Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

A MOLECULAR GENETIC INVESTIGATION OF IRISH ORIGINS

A thesis submitted to the University of Dublin for the degree of Doctor of Philosophy

Emmeline Wynne Hill

Department of Genetics

University of Dublin

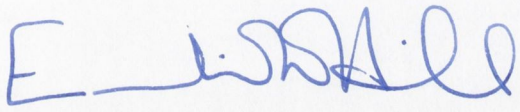
Trinity College

August 1999

Declaration

I hereby certify that this thesis, submitted to the University of Dublin for the degree of Doctor of Philosophy, has not been submitted as an exercise for a degree at any other university. I also certify that the work described here is entirely my own, unless stated as otherwise.

This thesis may be made available for consultation within the university library and may be photocopied or loaned to other libraries for the purposes of consultation.



Emmeline Wynne Hill

August 1999

Acknowledgements

My parents, without question, the best parents I ever had, are to whom I am mostly indebted: for your lifelong support and untiring encouragement, for the school fees to learn me and the phone bills to placate me, for the opportunities to find my dreams and to fulfil my goals and most importantly for your enthusiasm in all that I pursue, I thank you.

My utmost admiration for and thanks to the best supervisor and most inspiring teacher, Dan Bradley: for your patience and time dedicated to real understanding, your commitment to ensuring a thorough, well rounded education, and especially for your focus, keeping me on the right track and not letting me stray, thank you. The extended Bovine family, in particular my contemporary herd, Ciaran Meghen, Chris Troy and Stephen Park, I thank for their friendship and camaraderie over the years: Ciaran for his support since those early days as a bovine girl, a guide and mentor to whom I could always turn for advice; Chris for teaching me that apples + oranges = "big hands" and for being a constant source of amusement and distraction and ensuring that no work was ever done if he was bored; and Stephen for having answers to things that I thought didn't have answers and for teaching me that you can't make a mountain out of a mole, Hill. Also big thanks to Dave MAC Hugh and Ronan Loftus for being constant sources of endless information and for their worldly wisdom, and to Ashie for finally bringing some female support to the male dominance of the lab. I would also like to thank the calves of the lab, now too numerous to mention individually, for their support and friendship. There are also some other members of the department I would like to thank: in particular, David McConnell and Paddy Cunningham for their interest and enthusiasm in the work that I was doing, and for their wise words of direction. Also thanks to the Dave, Louis and Paul for their patience and help in solving the daily ordeals that were, I must say, frequent.

I am most grateful to Mark Jobling for his help when starting the Y-project, for pointing me in the right direction and for providing the link to the world outside Bovine genetics. I would also like to thank Matt Hurles for his help and for providing me with a means to date all the men in Ireland. To those who helped in the sampling effort I would like to thank Dr. Balabadi Dao, Dr. Okan Ertugrul, Dr. Bo Simonson and Mr. Kostas Papadopoulos. I would especially like to thank all the volunteers who agreed to take part in this work, who provided me with access to their X and Y chromosomes; this work would not have been possible without your help.

This would not be complete without thanking, from the bottom of my heart, my friends who have been by my side, always: Kate ("Do ya have a goal, 'cos ya gotta have a goal?"), Kathy and Joyce, the best friends anyone could ever wish for; Ben and Tom, Lu, Ants and Paddy, Brooks, O'Mahony, Hamilton and Harrison, Izzy, Amber, Tracy, Paula, Janice, Loo, all the cousins and family; and my absentee friends who have kept in touch and sent words of friendship and encouragement from around the world over the years. Without all of you the world would be a sombre (sober) place and I don't know where I'd be, thank you for believing in me.

This thesis is dedicated to my own ancestors, in particular to Grandpa Stamp and to the memory of Granny Stamp, Puffet and Dr. E. P. Hill.

Abbreviations

AD	after Christ
BC	before Christ
bp	base pairs
CI	confidence interval
D-loop	displacement loop
hg	haplogroup
ht	haplotype
MDS	multidimensional scaling
mtDNA	mitochondrial DNA
OTU	operational taxonomic unit
RFLP	restriction fragment length polymorphism
STR	simple tandem repeat
SNP	single nucleotide polymorphism
TMRCA	time to most recent common ancestor
UEP	unique event polymorphism
YAP	Y Alu polymorphism
YBP	years before present

Abstract

Three hundred and ninety-four male DNA samples, from Ireland, Turkey and Togo, were genotyped for 5 X chromosome simple tandem repeat (STR) loci. Six Y chromosome STR loci and 8 Y chromosome unique event polymorphisms were genotyped in 308 Irish, Togolese and Turkish male DNA samples. Additional samples representing other European populations were genotyped for greater European context.

The genetic diversity at X STR loci was highest in the Togolese population, intermediate in the Turkish population and lowest in the Irish population. The deepest evolutionary split between all populations was that between the Togolese and European populations. The estimated time to the most recent common ancestor between the Togolese population and the European populations was 168,000 *YBP* which concords with an Out of Africa model for human evolution.

The deepest evolutionary split determined by employing Y chromosome genetic markers was also that between the Togolese and European populations. The Turkish population was again intermediate to the Togolese and Irish populations.

In the Irish population surnames were used to subdivide the sample into components of known historic and prehistoric (Gaelic) origin, to which 1,000 year old geographical information may also be assigned. Unique event polymorphisms revealed 9 discrete haplogroup lineages unevenly distributed in the populations. Strikingly, western Gaelic samples are almost fixed (98%) for the putatively ancestral European haplogroup 1, as defined by biallelic markers of low mutation rate. Analysis of STR variation within this predominant Irish haplogroup suggests a late Neolithic / Early Bronze Age origin (4,200 *YBP*) possibly in a population expansion facilitated by the introduction of farming. The presence of other haplogroups (27%) in eastern Gaelic samples yields the first genetic evidence for a substantial secondary, prehistoric contribution to the gene pool of the island.

Table of contents

3 Y chromosome unique event polymorphisms

4 Genetic diversity on the Y chromosome

Figures

Tables

Chapter 2

Chapter 3

Chapter 4

Chapter 1:

General Introduction

1. General Introduction

1.1 Early hominid evolution

Genetic evidence suggests that the divergence of the hominid lineage from the anthropoid apes occurred little more than 5 million years ago *(Wilson and Sarich 1969; Sibley and Ahlquist 1984).* The earliest evidence for the hominid lineage in the fossil record is *Ardipithecus ramidus*, found in Ethiopia, Africa, dating to over 4.5 million *YBP (Wood 1996).* All of the earliest hominid remains are localised to Africa suggesting that the speciation event leading to the hominid lineage occurred on the African continent. The oldest direct ancestor for modern humans was *Australopithecus afarensis* dating to 3.5 million *YBP.* The many species of *australopithecines*, including *A. africanus*, *A. aethiopicus*, *A. robustus, A. garhi* and *A. boisei*, were confined to Africa and persisted until about one million *YBP (Johanson 1989, Asfaw et al 1999, Wood 1996).*

The genus *Homo* evolved from the small, sexually dimorphic, *A. afarensis* between 2.5 - 1.5 million *YBP.* Remains of the first members of the genus *Homo* are found at Olduvai Gorge and Koobi Fora and are found exclusively in Africa *(Leakey and Walker 1976, Leakey and Leakey 1978).* The first *Homo* species to emerge were *Homo habilis* and *Homo rudolfensis. H. habilis* was restricted to the African continent, had a larger brain than its *australopithecine* predecessors and was one of the first hominids capable of making simple stone tools *(Klein 1989).*

The first species to share a number of substantial anatomical features with later Homo species was *Homo ergaster*, or 'early *Homo erectus'*, which appears in the fossil record approximately 2 million *YBP (Wood 1996). Homo erectus* was the first hominid found outside Africa and remains dating to between 1.5 - 0.3 million *YBP* have been found in diverse geographical regions including Africa, Southeast Asia and Europe. The earliest unambiguous date for African *Homo erectus* comes from Olduvai Gorge and dates to more than one million *YBP (Swisher et al 1994).* In the east, the youngest mainland Asian *H. erectus* was found at Zhoukovdian, China, and dates to between 420,000 - 290,000 *YBP (Swisher 1994).* In Europe and the Middle East *H. erectus* remains have been found dating to over 500,000 *YBP (Cavalli-Sforza et al. 1994).* The widespread global distribution of

H. erectus suggests that the first global hominid population expansion from Africa occurred less than one million *YBP*.

Figure 1.1: 400,000 – 500,000 year old Acheulian hand axes and stone tools associated with *Homo erectus* populations, excavated by Louis and Mary Leakey at Olorgasaillie, Kenya.

1.1.2 The emergence of modern humans

The earliest, unambiguous, modern human remains were found at the Klaisies River Mouth site, Africa, and date to 116,000 - 127,000 *YBP (Swisher et al 1994).* The transition to modern *H. sapiens* is generally regarded to have occurred somewhere in Africa between 100,000 - 200,000 *YBP* although this has been questioned considerably. Recently an early *Homo* cranium was found in a succession of early - middle Pleistocene deposits in Eritrea, Africa, dating to approximately one million *YBP*. The cranium has a mixture of *Homo erectus* and *Homo sapiens* characters suggesting that early humans may have evolved much earlier than previously thought *(Abbate et al. 1998).*

Three hypotheses for the emergence of modern humans exist, for each of which there are a number of alternatives.

- The multiregional hypothesis holds that *H. sapiens* evolved from local archaic stocks in a number of geographically diverse regions of the world and that extensive global gene flow has prevented speciation and maintained the homogeneity of human populations *(Wolpoff 1989).*

- An alternative theory, advocating the admixture between ancient populations, suggests that *H. sapiens* evolved in Africa and spread to other regions of the world interbreeding with local *H. erectus* populations *(Brauer 1992).*

- However, it is widely held, that *H. sapiens* evolved in Africa and spread to other regions of the world either partially or completely replacing the indigenous *H. erectus* populations. There are a number of variants of the replacement hypothesis, the most well known being the Out of Africa hypothesis which suggests the emergence of modern humans in Africa within the past 100,000 - 200,000 *YBP* and the migration of anatomically modern humans throughout the globe with total replacement *(Cann et al. 1987; Stringer and Andrews 1988).*

1.1.3 Polycentric evolution of modern humans

The complete replacement of regional *H. erectus* and archaic *H. sapiens* populations has been contested by advocators of a polycentric origin for modern humans. A multiregional evolution of modern humans was first suggested in 1939 *(Weidenreich 1947)* and is still upheld by some scientists. The multiregional theory suggests that modern humans evolved in different regions of the world from local archaic hominid stocks. For example, it is proposed that modern Asians evolved from *H. erectus* populations represented by the Choukontien cave remains in China, commonly known as Peking Man, and that modern Southeast Asians and Australasians evolved from local *H. erectus* populations represented by the Indonesian Ngandong fossil, Java Man. If this were the case, based on fossil evidence, the differentiation of modern humans would have

begun over 1.8 million *YBP (Swisher et al 1994, Huang et al 1995, Gabunia and Veuka 1995).*

The basis for the multiregional hypothesis is gained from the continuity of physical features in the fossil record, which, it is claimed, demonstrates no evidence for a displacement of the local population by humans with modern anatomical features. Regional characteristics such as the prominent brow ridge in Australian aborigines *(Nei 1995)* are expected, therefore, to have been maintained in regional populations, without change, for over one million years. Stringer *(1995)* argues that modern human crania are distinct from archaic forms and do not suggest trends which might indicate regional continuity. If, as is often suggested, there is regional continuity between the Southeast Asian Ngandong fossil and Australians *(Wolpoff et al 1984),* then Australians should share a number of cranial traits with Southeast Asian *H. erectus.* However, the few characteristics that Australians do share with the Ngandong fossil are also found in early modern human remains from East Africa and the Near East *(Lahr 1996),* and most paleontological evidence now suggests that Australians resemble modern African remains more so than archaic Southeast Asian remains *(Foley 1998).* Recently Anton and Weinstein *(1999)* suggested that any archaic features demonstrated in fossil Australasians may be the result of artificial vault deformation, and further suggest that inferring relatedness between fossil Australians and Indonesian *Homo erectus* may be unwise.

The emergence of a single species demonstrating common physical features, *albeit* with specific regional or racial differences, requires a parallel evolution for global human populations over large geographic regions. In order to prevent speciation in disparate regions, a large amount of gene flow is required between local groups which, in turn, requires a large effective population size of the ancestral global population *(Wolpoff 1989).* The requirement of gene flow is somewhat contradictory to the regional continuity claim, as low levels of gene flow would be expected were regional physical characters to be maintained in local populations over a period of one million years.

Most genetic evidence suggests a small ancestral long-term effective population size which would not have facilitated the geographically widespread gene flow required for the parallel evolution of modern humans in disparate regions of the world. The rate at which lineages coalesce is proportional to the effective size of the population and the extent of gene flow between the coalescing lineages. Estimates of global coalescence and gene flow suggest the long-term effective size of the modern human population as ~10,000

breeding individuals *(Harding et al 1997; Reich and Goldstein 1998; Harpending et al 1998; Ziekowitch 1998; Nei and Graur 1984).*

If modern populations arose independently from ancient local hominids, then large levels of variation are expected between global populations *(Lewontin 1972; Nei and Roychudhury 1972).* Most genetic evidence, however, suggests very little between population variation, the majority of the variation partitioned in global populations being that between individuals within geographic population groups. For example, Lewontin *(1972)* showed that the majority of diversity within human populations is attributed to variation within populations (85.4%), intermediate diversity is found between populations within population groups or races (8.3%) and the least variation is attributed to variation between population groups (6.3%). More recently, less than 6% of the total global human genetic variation has been shown to occur among individuals of the same continent, the majority of diversity being found between individuals within discrete populations *(Barbujani et al 1997).* This suggests a low level of geographic differentiation among human populations and suggests that a polycentric evolution of modern humans, even with widespread gene flow, is unlikely.

1.1.4 A recent African origin with replacement

A variety of single origin models exist, alternatively coined the Garden of Eden, Noah's Ark and Out of Africa hypotheses. All of these models suggest an African origin for modern humans although the dates of divergence from a common ancestor differ, and the extent and effect of population bottlenecks and the extent of admixture with archaic humans throughout the evolution of modern *H. sapiens* also differ. The Out of Africa theory for modern humans requires an estimation for dates of divergence between global human populations after 200,000 *YBP.* Estimates for the coalescence time of β-globin genes, ~800,000 *YBP,* have been used to argue against a recent African origin for modern humans *(Fullerton et al. 1997).*

The divergence between African and all other global populations is expected to be the most ancient. Furthermore, a greater diversity within African populations is expected given a greater time for the accumulation of mutations in an older population. Most genetic analyses suggest a greater diversity within African populations compared to other global populations. Higher heterozygosities in sub-Saharan African populations than non-

African populations have been detected in a variety of genetic systems *(Bowcock et al. 1994; Relethford and Harpending 1994; Deka et al 1995; Jorde et al 1995, 1997; Tishkoff et al 1996).*

It has been suggested, however, that this greater diversity may be due to a larger ancestral population size of the African population. For example, Relethford and Harpending *(1994)* estimated a population size for the ancestral African population three times greater than other populations. More recently, Relethford and Jorde *(1999)* detected a significantly greater heterozygosity in sub-Saharan Africans than non-Africans which, they attributed to a larger ancestral population size in Africa compared to other global populations. A larger African population size does not discredit the single origin hypothesis, however, but rather suggests an important role for Africa in the evolution of modern human populations.

The Out of Africa hypothesis was proposed by Cann *et al. (1987)* following the analysis of mitochondrial DNA restriction fragment length polymorphisms in global human populations. A phylogenetic analysis demonstrated the initial separation of Africans from non-Africans at the root of a phylogenetic tree. The African clade contained African sequences only, whereas the non-African clade contained all non-African sequences and some diverse African sequences. This pattern suggested Africa as the ancestral source population for all modern humans. The divergence between the African cluster and the non-African cluster was estimated to have occurred ~200,000 *YBP* (140,000 - 290,000 *YBP*). Additionally, Africans were found to be 50% more different from each other than individuals within other non-African populations. Despite the wealth of criticism pertaining to this analysis *(Maddison 1991; Vigilant et al. 1991),* it remains the foundation genetic evidence for the Out of Africa hypothesis for modern human populations.

1.1.5 The extinction of the Neanderthals

The emergence of modern *H. sapiens* remains controversial, confused to a large part by the presence, in Europe, of the archaic human subspecies *H. s. neanderthalensis*. Neanderthal fossil remains suggest a robust hominid, retaining many primitive characteristics, with a comparable brain capacity to modern humans. Many Neanderthal archaeological sites are identified by the Mousterian material culture associated with the

Neanderthal lineage. Neanderthals appear in the European fossil record between 150,000 - 200,000 *YBP*, and in the Middle East ~300,000 *YBP (Stringer 1985, 1986; Vandermeersch 1985)* and persist until 30,000 - 40,000 *YBP*. The earliest anatomically modern humans in Europe were found at Cro Magnon, France, and date to 33,000 *YBP (Cavalli-Sforza et al 1994)* suggesting a coexistence of modern humans with archaics.

Three propositions have been made for the contribution of Neanderthal lineages to the modern human genepool. One claim posits that modern Europeans are direct descendants of Neanderthals. Alternatively Neanderthals may have contributed in some part to the modern European genepool by a process of admixture and extensive gene flow, or Neanderthals may have contributed nothing to the modern genepool and may have been completely replaced by modern humans in Europe and the Middle East.

That Neanderthals are direct modern human ancestors is discredited by the presence of both modern humans and archaic humans in comparable geological strata. Furthermore, mtDNA sequence analysis challenges the claim that Neanderthals are direct ancestors of modern humans suggesting no contribution of Neanderthal genes to the modern genepool. DNA extracted from a humerus bone of a Neanderthal fossil, dating to less than 100,000 *YBP*, was sequenced and compared to DNA sequences of contemporary modern humans representative of a number of global populations. The sequence comparisons did not detect a closer relationship between the Neanderthal sequence and European sequences which might be expected if Neanderthals were direct ancestors given that Neanderthals were confined to Europe and the Middle East. Furthermore, the estimated time for the divergence of the Neanderthal lineage and the modern lineage was 550,000 - 690,000 *YBP* which predates the first fossil evidence for Neanderthals. Additionally, pairwise sequence differences detected Neanderthal variation outside the variation for modern humans suggesting that no contribution of Neanderthal genes was made to the modern human lineage *(Krings et al 1997)*.

A common ancestor for modern humans and Neanderthals has been estimated from paleontological evidence as between 250,000 - 350,000 *YBP* in a population of African origin *(Lahr and Foley 1998)*. Most paleontological, archeological and genetic evidence suggests a complete replacement of archaic *H. sapiens* between 30,000 - 40,000 *YBP* with no genetic contribution to the modern genepool *(Cann et al. 1987; Vigilant et al 1991; Hammer 1995; Armour et al 1996; Krings et al. 1997; Torroni et al 1998)*. The disappearance of Neanderthal remains in the fossil record coincides with the first evidence

for modern humans in Europe and suggests an expansion of modern humans out of Africa into the Middle East and Europe and later into Asia replacing existing local hominids.

1.1.6 A global Paleolithic population expansion

Most genetic evidence suggests the recent evolution of human populations in Africa after 200,000 *YBP* and a subsequent migration from that continent to the rest of the world rapidly replacing indigenous archaic hominids. Recent genetic analyses have suggested a coincidence of a global population expansion at this time supporting a single origin for modern humans.

If a population is large now but was small at some time in the past, there will be few coalescence events since the expansion of the population. In this way, the size of the ancestral population is inversely proportional to the rate at which lineages coalesce. Phylogenetic analysis can investigate the presence of expansions in a population by analysis of the coalescent lineages. For example, gene trees in an expanding population are expected to have a star-like pattern, representative of a rapid recent expansion from an ancestral gene type *(Harpending et al. 1998; Castelloe and Templeton 1994).* Star-like genealogies are also produced following selective sweeps which mimic patterns that infer population growth. In both an expansion event and a selective sweep the present population will have a smaller than expected population size, with only a few alleles that coalesce in the distant past.

Characteristic signatures of past population dynamics may also be observed in distributions of mutational mismatches. Mismatch analyses have been successfully used to model ancient population expansions. In the case of a population expansion a smooth curve results because most of the coalescence has been contributed from few lineages, the remainder from closely related lineages. In contrast, a constant sized population at equilibrium will have a multimodal or ragged distribution *(Slatkin and Hudson 1991; Rogers and Harpending 1992).* It has been suggested that even if there has been continued exponential growth, a signature of the original expansion event will be observed in the data *(Rogers 1995).* Most distributions of global mitochondrial sequence mismatches share a common smooth appearance with the majority of the sequences coalescing some time in the Paleolithic *(Rogers 1995, Rogers et al 1996).* Dating the expansion of human mitochondrial lineages is based on a mutation rate for mtDNA sequence divergence of 7.5

x 10^{-4} and 1.5 x 10^{-3} *(Rogers and Harpending 1992).* For example, Rogers *(1995)* calculates the expansion of mtDNA lineages between 33,000 - 75,000 *YBP* and 66,000 - 150,000 *YBP* depending on the mutation rate used. This suggests an expansion of the population during the separation of African from non-African lineages. Generally a Pleistocene population expansion of 60,000 *YBP* is agreed *(Rogers and Harpending 1992; Sherry et al 1994; Rogers 1995).* However, it has also been suggested that this signature of expansion has been generated since the Neolithic within the past 10,000 years *(von Haeseler et al 1996).*

More recently the variance at microsatellite loci has been analysed for the detection of population expansions. In a population at equilibrium the variation at a large number of different loci will differ, whereas following a period of growth the variation at different loci is expected to be similar. Reich and Goldstein *(1998)* have estimated a population expansion between 49,000 - 640,000 *YBP* suggesting that a global population expansion occurred long before the Neolithic period and possibly even before the divergence of African populations from non-Africans. From the same data they estimate the divergence of Africans from non-Africans between 75,000 - 287,000 *YBP*. This suggests that the expansion occurred in Africa and not elsewhere. Similarly, Kimmel *et al (1998*) utilised the variance at microsatellite loci to determine the presence of population expansions in global populations. They found a departure from equilibrium which was strongest in Asian and weakest in African populations. They argue that this is consistent with an ancient bottleneck followed by a period of growth either just prior to or during the separation of African populations.

The multiple dispersal model of Lahr and Foley *(1998)* suggests a succession of bottlenecks in the ancestral human population coupled with a number of expansion events. Coinciding with changing climatic conditions, the first bottleneck might have occurred within Africa during the penultimate glacial period between 130,000 - 190,000 *YBP*. They suggest that a possible initial migration event from Africa ~100,000 *YBP* did not result in the replacement of archaic hominids in the world, but that a second expansion event ~45,000 *YBP* resulted in the permanent replacement and ultimate extinction of the archaic lineages. Following the initial replacement of European and Near Eastern archaic populations, further expansions and migrations both within the Near East and Europe and to more distant global regions occurred.

1.1.7 The Neolithic in Europe

Possibly the most important human cultural innovation after the manufacture of tools, the development of fire and the acquisition of language is the domestication of wild plant and animal species. The implications of domestication for human populations cannot be over emphasised as it is this innovation which led human populations to the transition from a nomadic hunter-gatherer lifestyle to a sedentary agriculturally based existence, ultimately enabling the highly complex technological civilisation of the world today. The transition from traditional nomadic hunting and gathering societies to sedentary agricultural societies is believed to have been a gradual process, and cannot be explained by a simple one-step model as is often implied by reference to an *"agricultural revolution"*.

Although a number of global centers of domestication have been identified, the best known region is the Fertile Crescent, running in a crescent shape from Israel to Syria, northern Iraq and western Iran, and into the southern regions of Anatolia. Other regions identified as major centers of domestication include Southeast Asia, Central America and some northern regions of South America *(Wenke 1990)*. The period in which the initial domestication of local plant and animal species occurred is known as the Neolithic period beginning over 12,000 *YBP*. Climatic changes at this time, associated with the end of the most recent Ice Age, may have been the most influential factor in the implementation of domestication. Recent evidence ascertained from the Greenland Ice-core project suggests that the past 12,000 years have seen the most stable climatic conditions in the history of human occupation of the globe *(Greenland Ice-core Project, 1993)*. It may only have been possible, therefore, in such favourable conditions, to harness and exploit natural resources in such a manner, although the potential for domestication may have long been previously present.

It has been suggested that the shift to a predominantly agricultural society during the Neolithic period was accompanied by rapid human population growth *(Cohen 1977)*. Hunter-gatherer societies may have only had population densities in the range 0.01 - 9.5 person / sq. km, although the upper limit would include only highly successful communities in areas with abundant natural resources *(Hassan 1975)*. For example, for the Basque population, Calderon *et al (1998)* suggest that the population density during the Paleolithic was 0.1 person/ sq. km, i.e. ~2,000 people. Slatkin and Ranala *(1997)*

estimated the rate of population growth in Western European populations as r = 0.005 per year (0.01 per generation, given a 20 year generation time). The population of the Basque country is now 3 million people *(Calderon et al 1998).*

It is not clear whether a population increase necessitated the adoption of farming to sustain the growing populous or whether the adoption of farming facilitated a growth in the population by increasing the carrying capacity of the land. Perhaps it was a combination of both. Contemporary hunter-gatherer societies have fewer children than agriculturally based societies *(Cavalli-Sforza 1986),* and the birth rate closely mirrors the death rate. Population numbers are therefore static without growth or demise of the population except under extreme conditions. In an environment capable of sustaining more people, as in an agriculturally based society, population numbers would have increased dramatically, probably over a short period of time. Marginal areas of the more lucrative lands would have been occupied and population pressures most certainly encouraged migration.

1.1.8 The demic diffusion hypothesis

The nature and extent of migration following the advent of agricultural technologies and the extent of the demographic influence of the Neolithics in Europe has been widely debated. That agricultural technologies have been adopted in all European societies over the course of the past 10,000 years is not disputed. However, whether the incorporation of agriculture resulted from a spread of the agriculturalists themselves, in a process of demic diffusion, or whether the diffusion involved the spread of the technology alone, in a process of cultural diffusion, is the central debate.

The formation of the demic diffusion hypothesis, otherwise known as the *"wave of advance"* model, was first suggested by Ammerman and Cavalli-Sforza *(1973).* They estimated the rate of advance of a population, growing and spreading at a constant rate in all directions away from the center of diffusion, as 1.1 km per year. This requires only local movement with marginal advancement into, and occupation of, peripheral lands.

The spread of agriculture, within Europe, has been identified from characteristic archaeological remains and relics of pottery cultures associated with ancient agricultural societies. However, the greatest support for a demic diffusion of agriculturalists from the Near East into Europe following the Neolithic period comes from the interpretation of

synthetic maps of gene frequencies first created by Menozzi *et al. (1978)*. Synthetic maps are created from isopleths of equal gene frequencies calculated by principal components, uncorrelated linear functions of the original gene frequencies, subsequently superimposed onto geographic maps. In the case of demic diffusion, circular gradients of gene frequencies are expected to emanate from a center of origin. Genetic drift is a major contributor to genetic differentiation in populations assuming low population density. For example, during the Paleolithic when population densities were low, drift would have played a major role in the differentiation of human populations. Demographic factors such as population expansions followed by migration and admixture would then have created clines of gene frequencies between previously differentiated populations.

Cavalli-Sforza *et al. (1994)* performed principal component analysis on 95 nuclear classical genetic polymorphisms in a pan-European data set. The map obtained from the first principal component, explained 28.1% of the total genetic variation and suggested a center of origin in the southeastern region of Europe, depicting a gradient of gene frequencies in a southeast to northwest cline. Significant correlations with dates for the spread of agriculture with this cline led them to suggest that the map of first principal components reflects the spread of agriculturalists into Europe and further suggests a replacement of the indigenous Mesolithic hunter-gatherer communities occupying the European continent at that time. In brief, the synthetic maps of Cavalli-Sforza *et al. (1994)* suggest a center of origin in the Near East and the migration of farmers into the northwest of the continent from ~ 10,000 *YBP* at a rate of 1 km per year.

This gradient of gene frequencies across Europe suggests a largely Neolithic European ancestry. A predominantly Neolithic ancestry for European populations has also been suggested by clines of nuclear DNA polymorphisms *(Chikhi et al 1998)*. Drawing on data from seven nuclear markers, including two minisatellites, four microsatellites and one alpha chain, northern and western populations demonstrate the greatest divergence from Near Eastern populations. Additionally, estimates for the divergences between most European populations postdate 10,000 *YBP*, the exceptions including the Basque, Finnish and Sardinian populations, all known to be genetic outliers within Europe *(Cavalli-Sforza et al 1994)*. This cline suggests, therefore, that the main demographic influences affecting the European genepool occurred with the Neolithic expansion of agriculturalists from the Near East.

1.1.9 Cultural diffusion and pioneer colonisation

A cultural diffusion of agricultural technologies has been suggested which opposes the demic diffusion hypothesis. A cultural diffusion suggests a main diffusion of ideas rather than movement of people from the Near East. Although the cultural diffusion model is plausible it seems unlikely that little or no element of demic migration occurred throughout Europe over the past 10,000 *YBP*. The pioneer colonisation hypothesis takes an intermediate stand and suggests some migrations from the Near East into Europe, although these migrations may have involved only small groups of colonists *(Zvelebil 1986)*. Richards *et al. (1996)* suggest that the pattern of gene frequencies in the synthetic maps of Cavalli-Sforza *et al. (1994)* represent more ancient migrations during the Paleolithic and question the Neolithic demic diffusion hypothesis. European mtDNA lineages coalesce in the Upper Paleolithic, suggesting that the major demographic influences in Europe were incorporated during the Upper Paleolithic and therefore the trends in the geographic distributions of gene frequencies may reflect earlier demographic population movements. They further suggest that the Neolithic expansion contributed little to the genetic composition of Europe. Additionally, Sokal *et al. (1999)* have warned that synthetic maps, based on principal component analysis, may be prone to large statistical errors and that the detected geographic trends can also be found in spatially randomised data.

1.1.10 The spread of Indo-European languages

The spread of the Indo-European languages has often been associated with the spread of agriculture. However, the debate over whether the Near East was the homeland of proto-Indo-European languages remains. Renfrew *(1989)* is the main advocate of an Anatolian origin for Indo-European languages, but others have suggested a center of origin for Indo-European languages further east and north of Anatolia *(Gamkrelidze and Ivanov 1990, Gimbutas 1970)*.

Most scholars accept, however, an expansion of Indo-European languages associated with the expansion of the Kurgan culture in the fourth millennium BC, suggesting the homeland in the steppes of the southern Ukraine and south Russia between the Black and Caspian Seas *(Gimbutas 1970)*. Linguistically, the dispersal of the Ugro-

Finnic and Kartvelian languages can be explained by an Indo-European homeland in the Russian steppes, as can the spread of Indo-Iranian languages to Iran and India. The spread of Indo-European languages from the steppes might be reflected in the third principal component synthetic map of Europe *(Piazza et al 1995).* They suggest, however, that an Anatolian origin is not completely incompatible with a Kurgan expansion of languages which may have been a secondary wave of diffusion of Indo-European languages.

1.1.11 European genetic isolates

Genetic evidence suggests that within Europe there are a number of populations genetically isolated from continental European populations who may not have their origins in a Neolithic expansion of agriculturalists and, furthermore, may have avoided recent homogenisation with European populations due to their geographically proximal locations. These populations include the Finns, Sardinians and Basques *(Cavalli-Sforza et al 1994).*

Basques

Those Mesolithics isolated in peripheral regions of the continent may have avoided large scale population replacement and influence from Neolithic agriculturalists. For example, the Basque population, confined to the southwest of France and the northwest of Spain, who speak the only surviving non-Indo-European language in Western Europe, Basque or Euskara *(Ruhlen 1987),* possess a genetic heritage unlike any other European population. It has been suggested that the Basques are relics of Mesolithic populations and may be direct descendants of cromagnoid peoples *(Cavalli-Sforza et al 1994; Bertranpetit and Cavalli-Sforza 1991).*

This concept is backed by strong genetic evidence. For example, the highest global frequency of Rh- has been detected in the Basque population *(Bertranpetit and Cavalli-Sforza 1991).* Additionally, differences in frequencies of polymorphisms at other classical genetic loci such as the ABO, HLA and Bf loci, and some protein and enzyme markers, have been found between the Basques and neighbouring populations. The highest frequencies of blood group O and Rhesus *cde*, complement components C4F and C3F and the properdin factor gene *BfF 1,* and the lowest frequency of blood group B in Europe are found in Basque peoples *(Lucotte and Hazout 1996 and references therein).* A principal

component analysis of classical genetic polymorphisms in the Iberian Peninsula demonstrates the main difference between the Basque population and the other Iberian populations, explaining 27.1% of the total genetic variation within Iberian populations *(Bertranpetit and Cavalli-Sforza 1991).* A recent principal component analysis of the HLA class I and II loci also suggests the distinct genetic nature of the Basque population within Europe *(Comas et al. 1998).* Generally, isolation by distance is assumed to have shaped the main differences between the Basque and neighbouring populations suggesting little or no admixture, and a distinct isolation of this population from immigrant Neolithics.

In a phylogenetic analysis of the mtDNA hypervariable region a differentiation of the Basques compared to other European populations was not detected *(Bertranpetit et al 1995).* This was explained by the high mutation rate of mtDNA leading to possible homogenisation of genetic distances between populations. By contrast, another phylogenetic study of mtDNA sequences did detect the Basque population as being different to other European populations *(Richards et al. 1996).* However, because these authors have suggested that the majority of the European gene pool was established during the Upper Paleolithic, the differentiation of the Basque population from other European populations was explained not simply by isolation from the Neolithics but by a long period of isolation by distance leading to substantial genetic drift in that population.

Until recently, the highest frequency of the Y chromosome haplotype XV, at the p49 (DYS1) locus, in Europe was observed in the Basque population, estimated as 72.2%, while the mean frequency in other European populations was 41% *(Lucotte and Hazout 1996).* The high frequency of haplotype XV and the low frequencies of the southern European and Near Eastern haplotypes at this locus suggests, assuming the Neolithic demic diffusion hypothesis, that substantial incursion by Neolithic agriculturalists did not occur in the Basque population. Analysis of haplotype frequencies and haplotype diversity values at the Y chromosome microsatellite locus, DYS413, also demonstrates the distinctive nature of Basque genetics compared to other European populations. The highest frequency of haplotype 23/23 is found in the Basque population (74.5%) which is much higher than the average frequency of this haplotype in six other southern European populations (25.8%). Additionally the low level of haplotype diversity in the Basque population (0.431) compared to other southern European populations (0.820) may suggest the increased effect of genetic drift in the Basques *(Scozzari et al. 1997)* or alternatively a lack of admixture with other European populations. The effect of drift in the Basque

population is also suggested by an analysis of Fst values at Y STR loci between the Basque population and the neighbouring Catalan population. By comparison to the overall differentiation between European populations *(Cavalli-Sforza et al. 1994),* the Fst value between Basques and Catalans amounts to almost two-thirds of the total European differentiation *(Pérez-Lezaun et al 1997).*

Phylogenetic analysis of global populations at the highly polymorphic minisatellite locus, MS205, illustrates the Basque population clustering with other European populations. However, in a neighbour-joining tree, using a genetic distance based on the average dissimilarity of alleles within and between populations, the long branch length leading to the Basque node suggests an ancient divergence of the Basque population from other populations within Europe. Bootstrap analysis determines the Basque population intermediate between European and African populations. Alonso and Armour *(1998)* suggest that although the Basques are a divergent European population, the intermediate phylogenetic position may reflect the retention of ancestral alleles in the Basque population. Estimations of ages for some of the Basque specific alleles at the MS205 locus all date to pre-Neolithic times and range from 15,079 *YBP* to 21,442 *YBP* (confidence limits ranging from 3,720 *YBP* to 33,580 *YBP*).

1.2. Mesolithic Ireland

Evidence for human occupation, from as early as 500,000 years ago, has been found in south-eastern England, but no such evidence has yet been found in Ireland *(Woodman 1986).* The first evidence for seasonal occupation of Ireland is found in the northern regions of the country at a time of approximately 9,000 *YBP (Woodman 1985*). These hunter-gatherer communities would probably have crossed into Ireland from Scotland via a landbridge that formerly connected the two landmasses.

The Midlandian glaciation, spanning over 100,000 years, was the last glaciation in Ireland. Substantial global temperature increases around 12,000 *YBP* marked the end of the Ice Age in Ireland *(GRIP 1993).* The melting of the glacial icesheets resulted in eustatic rises in sea levels which caused the ultimate separation of Ireland from Britain about 8,000 *YBP (Waddel 1998).*

Evidence from a number of archaeological sites suggests that the first large scale settlements of hunter-gatherer peoples was perhaps not until later in the Mesolithic period about 7,500 *YBP.* However, Ireland may still have remained largely isolated from the rest of the continent and population densities may have remained relatively low. It has been estimated that only several thousand people lived in Ireland at the end of the Mesolithic *(cited in Waddel 1998).*

There is no evidence in the archaeological record for a complete and rapid transition from hunting, foraging and fishing to farming, and therefore reliable dating of the introduction of agriculture, from archaeological records, is tenuous. Indications of agricultural communities are, however, unambiguously evident from shortly after 6,000 *YBP.* This is later than evidence for other continental adoptions of agriculture but a delay in the adoption of farming technologies in Atlantic Europe has been suggested by Zvelebil and Rowley-Conwy *(1986)* who suggest that in sedentary and socio-economically complex hunter-gatherer societies the adoption of agriculture might have been slowed.

1.2.2 The Bronze Age in Ireland

Evidence for an intensification of agriculture in Ireland between 4,000 - 6,000 *YBP* stems mainly from the excavation of domestic animal remains and the changes in flora and soil content associated with forest clearance. Additionally the appearance of megalithic

tombs and a pottery culture at this time indicate a major cultural transition which may have facilitated a rapid growth in the indigenous population. For example, in the west of Ireland a decrease in woodland has been radio-carbon dated to 4,900 - 5,600 *YBP* by the association of the increase in grasses, heather, plantain and bracken. A decrease in soil cover, also associated with woodland depletion has been uranium-thorium dated in the west of Ireland to 3,000 - 4,000 *YBP* by means of assessment of spleotherm growth. The growth of subterranean calcite deposits is proportional to the acidity of the soil. When soil is depleted, spleotherms cease to grow and this has be dated and correlated with a decrease in woodland cover *(Jones 1997).* Soil depletion may be associated with decreased woodland cover, or alternatively with overgrazing as a result of agricultural intensification. For example, there is evidence for substantial overgrazing at the Céide Fields in Co. Mayo, a large enclosed site centered on pastoral farming and dating to 3,700 - 3,200 BC *(Waddel 1998).* The overgrazing of this land would have resulted in nutrient depletion of the soil and the subsequent increase in the formation of bogland, a signature today of highly intensive farming activities at this time.

During this period there is evidence for the appearance of megalithic tombs in Ireland dating to between 4,000 - 2,000 BC *(Waddel 1998).* Megalithic tombs are unique to Europe and are found in northern Scandanavian countries as well as more southern Mediterranean countries, Britain and France. The appearance of Megalithic tombs in Scandinavia and in northern Europe are also associated with large scale forest destruction, possibly by burning, and the adoption of novel farming technologies *(Fagan 1980).* The scale and magnitude of such monuments as the formidable passage grave at Newgrange, Co. Meath, dating to 4,425 *YBP*, and their emergence at a time coinciding with the adoption of farming suggests a population of considerable size and density with a highly complex social structure.

1.2.3 The Celts

Celtic languages are spoken today only on the periphery of Europe although they once dominated much of Central and Western Europe. Traditionally Celtic languages are grouped according to a continental or insular center. Continental languages include Hispano-Celtic, Gaulish, Lepontic and Eastern Celtic, none of which is spoken today. The insular Celtic languages include Breton, Scots Gaelic, British, Welsh, Manx, Pictish and

Irish Gaelic all of which are confined to the British Isles and closely neighbouring regions, some of which are still spoken today *(Mallory 1991).*

In Ireland, evidence for a Celtic language affinity emerges in the sixth century BC with the labelling of Ireland as *Ierne*, although it was not until the first millennium AD that Ireland possessed its own script, known as *ogham*. Prior to this there is no recorded evidence for a language in Ireland and it is therefore unknown what language was spoken on the island prior to Celtic Gaelic.

The introduction of novel languages has often been associated with the introduction of material cultures. The La Tène culture is assumed to be the archaeological presence associated with Celtic languages just as the Jastorf culture has been associated with Germanic languages and some African pottery cultures have been associated with Bantu speakers *(Mallory 1989).* The La Tène culture, developed in the Rhine and Danube valleys and named after a famous discovery of typical artifacts near Lake Neuchatel, Switzerland, was an adaptation of ironwork technologies brought about with the demise of bronze metallurgy in the third millennium BC.

The question of the origin of the Gaelic language in Ireland has caused some debate. A simple explanation for the introduction of the Celtic language suggests an intrusion of Indo-European Celtic speakers from the continent in the first millennium BC bringing with them the La Tène material culture. The scale and success of such an invasion is questionable, however, with little archaeological evidence to suggest a large scale demic incursion by exogenous peoples (*Waddel 1998).* By comparison, there is abundant evidence for Celtic intrusions in the archeological record of Hungary, Romania and Yugoslavia *(Mallory 1989).* It is arguable that the Celtic language was introduced through the increasing trade and communication of the indigenous Irish peoples with Britain and continental Europe. Renfrew *(1987)* links the origin of the Celtic language in Ireland, not with the invasion of a dominant people but rather through a mechanism of élite dominance, the invasion by a small number of powerful peoples who spread the language throughout the island with little demographic influence. In fact, Renfrew rejects suggestions of any Bronze or Iron Age migration into Ireland. With little archaeological evidence to suggest a demographic invasion to the country and no records of spoken language, it might also be conceivable that Gaelic was spoken in Ireland before the spread of the Celts.

1.2.4 The Vikings in Ireland

The first Viking invasions in Ireland are documented as occurring in 794 AD in the Dublin Bay area. Viking settlements became common on coastal regions particularly in Dublin, Waterford, Wexford and Limerick. Internal waterways such as the Shannon river also facilitated the invasion of midland regions of Ireland. By 832 AD permanent Viking settlements had been established in the country, mainly comprised of Vikings of Norwegian origin. Later invasions saw the incursion of Danish Vikings who again initially attacked the eastern shores of the island but soon established settlements both on the coast and inland. Constant attacks on the island persisted until about 1200 AD, by which time there was a large population of exogenous origin on the island. Frequent contact with the indigenous insular population has been suggested by many historians, some even suggesting a great deal of intermarriage between the two groups *(Waddel 1998, Relethford and Crawford 1995 and references therein).*

1.2.5 Anglo-Norman and Tudor England influences

The Anglo-Norman invasions of the twelfth century and the later plantation settlers from both England and Wales in the early seventeenth century had most of their influences in the north, east and southeast regions of the country.

1.2.6 The population genetic structure of the Irish population

There has been little investigation of the population genetic structure of the Irish population. Recently the analysis of quantitative traits such as skin colour and anthropometrics has been less popular than the use of genetic markers for studies of population demography. However, most population data for Ireland comes from analyses of anthropometric variation and also the distribution of ABO blood group genes. During the 1930s, Dupertuis and Dawson collected a large set of anthropometric data for 10,000 males and almost 2,000 females *(Hooton 1955).* Analysis of this data was, however, limited due to constraints on analytical techniques with large amounts of data. Relethford *et al. (1980)* reanalysed the data from 347 males and 261 females from 12 towns in three counties, Mayo, Galway and Clare, of western Ireland. Sixteen anthropometric variables

were chosen for their analysis. To assess population differentiation, an analysis of variance to estimate the proportion of among group variance was performed. Highly significant variation ($p<0.05$) was found among the 12 towns. It was shown that migration reduced the degree of among group variation and furthermore that relationships between geographic distance and anthropometric differentiation are highly correlated.

It has also been shown that English admixture has had a significant effect ($r = 0.829$, $P = 0.047$) on anthropometric variation within Irish populations *(Relethford 1988).* The degree of population differentiation within western Irish populations was found to be due to the effects of recent English admixture in these populations, and isolation by distance. Using data from the same large anthropometric data set Relethford and Crawford *(1995)* estimated genetic distances between 31 of the 32 counties in Ireland. A principal coordinates analysis identified the initial separation of the midland counties form all others, representing 35% of the total genetic variation within the country. Other population analyses of Ireland had shown the distinctiveness of the midland counties from the coastal counties. Relethford and Crawford *(1995)* suggest that the distinctiveness of the midland populations may be due to extensive contact with Viking populations in the ninth to thirteenth centuries. This is, however, questionable as most of the Nordic influences were to be found along the eastern and southern seaboards. They further suggest that recent population history, particularly that of historical settlements, has been the major influence on the population genetic structure of Ireland.

The second principal coordinate of Relethford and Crawford's anthropometric data accounted for 18% of the total genetic variation and demonstrated a longitudinal gradient across the country, clearly separating eastern counties from western counties. An east-west longitudinal gradient across the country has also been established in the analysis of ABO blood group frequencies in different populations in Ireland. In the west of Ireland the frequency of blood group O reaches its highest frequency in Europe *(Hackett and Dawson 1958).* Blood group O frequencies were found to decline steadily across Ireland towards eastern regions with a corresponding increase in the frequency of blood group A. They found that, in general, frequency gradients were found across the country with no sharp changes in frequencies between any regions. This pattern is also found in frequencies of blood group distributions in England but is unlike the distributions for Wales where sharp local differences have been detected.

Hackett and Dawson *(1958)* suggested that the high frequency of O genes in the west of the island might reflect the antiquity of the populations of the west relative to the east. Group A genes are common in the south of England and they have suggested that it is the English influence in the east of the country that introduced the A group genes. In an analysis of blood groups in Ulster, Hart *(1944, cited in Hackett and Dawson 1958)* showed that in a comparison of people with English, Scottish and Irish surnames, those with English surnames had the highest frequency of group A genes, similar to the frequencies found in the south of England. Additionally, those regions of Ireland that have had the most Anglo-Norman influence had greater frequencies of group A genes compared to other regions. In a comprehensive survey from a sample of 1 in 18 of the population Dawson *(1964)* showed that the patterns demonstrated in the earlier study were well founded and that the frequencies of group O were highest in the west and the frequencies of group A were highest in the east.

In a similar survey in the 1970's, Tills *et al. (1977)* further demonstrated conformity with the original study of 1958. However, with the exception of group O the distribution and frequencies of the other blood groups were similar to those in the rest of Europe, and did not show any significant difference between the provinces within Ireland. In a study of gene frequencies of a number of red cell and serum protein polymorphisms, Tills *(1977)* showed that, in general, frequencies within Ireland were not significantly different from frequencies in other western European countries, although AK and PGD frequencies were slightly lower and ADA frequencies were slightly higher.

Although these studies were significant in their time and instigated a suspicion that western Irish populations were perhaps significantly different from other European populations, their low level of resolution has not been able to further suggest population demographics either within Ireland or in a continental context. More recently molecular genetic techniques have been used to make more precise predictions about the population history of Ireland.

For example, Zschocke *et al. (1997)* analysed mutations at the common autosomal recessive disease gene locus, phenylketonuria (PKU), to infer patterns of ancient demography in Ireland. PKU is very heterogeneous but some populations have mutations that account for the majority of PKU alleles. In this way, population specific alleles can be identified and their distribution in disparate regions may indicate past patterns of migration between different populations. The mutation I65T is common in Ireland, Britain and

Spain, but rare elsewhere in Europe, and is an ancient mutation found on at least five different haplotypes. It has been suggested, therefore, that this mutation is a pre-Neolithic marker representing Mesolithic European populations. The most common mutation in Ireland is R408W, which may have been prevalent in Neolithic populations, and it is suggested that it may have been introduced with the introduction of farming in Ireland. Other PKU mutations in Ireland have been linked to later migrations such as the Viking invasions, but no genetic evidence was found to exist for a significant genetic contribution by the Celts in Ireland.

1.3 Genetic variation and population genetic studies

Genetic variation at the biochemical level was first demonstrated at the beginning of the century by showing that different humans displayed heritable variation at the ABO blood group system *(Nuttal 1904, cited Avise 1994).* This concept was first applied in a study of genetic variation of blood groups in human populations *(Hirszfield and Hirszfield 1919, cited Cavalli-Sforza 1994).* In 1949, Pauling *et al. (1949)* demonstrated the molecular nature of the mutation leading to sickle-cell anaemia, but it was not until the 1960s that the study of evolutionary processes using molecular techniques began in earnest. The first demonstrations of protein polymorphism were demonstrated by the separation of non-denatured proteins using simple electrophoresis by separation of the proteins according to their net charge *(Harris 1966; Lewontin and Hubby 1966).* Initial studies of simple protein polymorphisms revealed an abundance of genetic variation within animal genomes leading to the formulation of the *"neutral theory of molecular evolution" (Kimura 1968).*

The derivation of the concept of the *"molecular clock"* by Zuckerkandl and Pauling *(1962, cited Nei 1987)* was a breakthrough in the study of molecular evolution. They proposed that various proteins and DNA might evolve stochastically at steady rates over time, thus providing built in clocks for dating past evolutionary events and correlating genetic dissimilarity with time.

1.3.2. Mitochondrial DNA studies

Mitochondrial DNA is useful for population genetic studies as the transmission of mitochondria occurs exclusively through the maternal line with no recombination with male derived mitochondria and therefore mtDNA sequences are effectively haploid. Additionally, the mutation rate of mitochondrial sequences is 5 - 10 times greater than nuclear sequences *(Brown 1985).* In particular, the non-coding control region has the fastest rate of mutation with an estimated sequence divergence of between 7.5×10^{-4} and 1.5×10^{-3} per base per generation *(Rogers and Harpending 1992).* The variation at the control region has been widely used in human population genetic studies to infer evolutionary relationships *(Cann et al. 1987; Horai et al 1990; Vigilant et al. 1991, Torroni et al 1994, Richards et al. 1996; Richards et al 1998, Macaulay et al 1999, Sykes*

1999;). As already discussed, mtDNA studies first determined that the initial split of human populations was that between African and non-African populations < 200,000 *YBP*, thus suggesting an Out of Africa hypothesis for human evolution. More recently, mtDNA studies of populations within Europe have challenged the demic diffusion hypothesis originally proposed by Ammerman and Cavalli-Sforza *(1973)* by detecting the majority of mtDNA lineages coalescing in the Upper Paleolithic, with only a small proportion being attributed to the influence of Neolithics from the Near East.

A wealth of data exists for mtDNA lineage relationships within Europe. Richards et *al. (1996)* analysed 821 individuals of European and Middle Eastern origin for diversity at the mtDNA control region. Five major lineages were identified, some of which were widespread in Europe, others being confined to a more specific ancestry. More recently data for over 900 individuals has been published *(Richards et al 1998)*. Haplogroup U is thought to be the most ancient cluster in Europe, originating ~36,500 *YBP,* and was found to be widespread. The diversity in the sub-cluster U5 was estimated to have originated over 50,000 *YBP*. Cluster U5 was the only ancient cluster specific to Europe and was detected in 7% of the European sample, suggesting that the original Paleolithic settlers of Europe had haplogroup U5 mtDNA types.

They have suggested that other haplogroups including clusters I, X, H, V and K were also probably present in Europe prior to the Neolithic period. Haplogroup I was found in individuals of both British and Middle Eastern origin and was the oldest Caucasian lineage, estimated to have originated 50,500 *YBP*. Haplogroup H was also found to be widespread in Europe and was found at higher frequencies than haplogroup U. Haplogroup H had its highest frequency in the Basque population. The origin of haplogroup H was estimated at 21,000 *YBP (Richards et al 1998)*. Haplogroup V, a sister haplogroup to haplogroup H was estimated to be 12,500 years old. Haplogroup K also has a similar distribution and age to haplogroups H and V. When corrected for total European diversity, most of the divergence dates for the European lineages are estimated between 11,000 - 18,000 *YBP*, but may have been introduced earlier in the Paleolithic. Sykes *(1999)* argues that the dates for these haplogroups correspond with expansions associated with the end of the last glacial period in Europe when populations recolonised northern European regions following a period of refuge in southern more temperate conditions.

Haplogroup J probably originated in the Middle East. The overall age of haplogroup J in Europe has been estimated as 28,000 *YBP*. But it has been suggested that

haplogroup J lineages were introduced into Europe with Neolithic farmers ~8,000 *YBP (Richards et al 1998, Sykes 1999)* and that the distribution of haplogroup J lineages in Europe follows the trend of the demic diffusion of farmers from the Near East. Two major sublineages of haplogroup J have been detected, one follows a central and northern European direction, probably associated with the LBK culture, another, slightly older than the central and northern cluster, follows a southern Mediterranean path to Britain and may have been associated with the Impressed Ware or Cardial Ware cultures.

Torroni *et al. (1998)* also suggest a major Paleolithic population expansion from the Atlantic zone between 10,000 - 15,000 *YBP*. This may be associated with the northern advance of peoples following the end of the last Ice Age in Europe. Haplogroup V is suggested as the autochthonous European haplotype, found at a frequency of 20% in the Basque population, 10.4% in the Sardinian population, 5.4% and 4.1% in Swedes and Finns respectively and 40.9% in Saami population. Haplogroup V is absent in southeast Europe and the Near East except for being found in one individual of Turkish origin. Haplogroup H was found at a frequency of 50% in the Basque population, 45.8% in Sardinians, 40.5% and 40,8% in Swedes and Finns respectively, 21.6% in Italians and 9.8% in the Yakut population. Haplogroup H was found to have a higher diversity in the Near East than Europe. In Western Europe its origin is estimated as 14,100 - 15,100 *YBP* whereas in the Near East it is estimated as 27,700 - 29,600 *YBP*. These dates are older than those estimated by Richards *et al (1996,1998)* who used different mutation rate estimates.

Most of the mtDNA analysis of European populations therefore supports a major demographic movement of peoples into the continent some time in the Upper Paleolithic. More recent migrations have had some influence on the present population, but the Neolithic movement of peoples was not as pronounced as suggested by the demic diffusion hypothesis. Recently Sykes *(1999)* has claimed that the proportion of Neolithic haplogroup J lineages in Europe (~20%) corresponds to the percentage of the variation explained by the first principal component of the classical genetic dataset (~28%) as proposed by Cavali-Sforza *et al. (1994)*. This suggests that the two systems do not show anomalous results but rather that the interpretation of the genetic trends differs. The synthetic maps may indeed represent the Neolithic diffusion of agriculturalists, but the distribution may not be explained by a complete absorption of autochthonous Mesolithics.

The trend in gene frequencies may suggest a dilution of both Mesolithic and Neolithic genes as more northern and western regions of the continent are reached.

1.3.3. Simple tandem repeat loci

Southern blot analysis first revealed hypervariable regions of DNA in the human genome that became known as minisatellites *(Jeffreys et al 1985).* The DNA probes originally employed by Jeffreys *et al. (1985)* hybridised to conserved sequences of ~10 - 15 bp in length, that were found to be scattered in numerous arrays about the human genome as part of a system of dispersed tandem repeats, in what are now referred to as variable number of tandem repeat (VNTR) loci. Each repeat unit was found to be ~16 - 64 bp in length.

Another class of genetic marker, more commonly used in population genetic studies, are microsatellite loci *(Litt and Luty 1989, Tautz 1989, Weber and May 1989).* Microsatellites, or simple tandem repeat (STR) loci, are ubiquitously interspersed repetitive sequences found abundantly in eukaryotic genomes and in the chloroplastic genome of plants. An estimated 35,000 STR sequences are found in the haploid human genome occurring randomly every approximately 100 kb of sequence *(Weber 1990).* The repetitive sequences are composed of simple sequence motifs of 2 - 6 nucleotides organised in tandem array *(Hamada et al 1982).* They are codominant and are inherited in a Mendelian fashion. The most commonly studied microsatellites are dinucleotide $(dC\text{-}dA)_n \bullet (dG\text{-}dT)_n$ repeats which are the most common in mammalian genomes. At each locus the number of dinucleotide repeats is usually < 30.

The existence of simple repetitive sequences was first documented over 20 years ago although the ubiquitous distribution of these sequences throughout eukaryotic genomes was first shown by Hamada *et al. (1982)* who detected hundreds of copies of poly-$(dC\text{-}dA)_n$ motifs in yeast and tens of thousands in vertebrates. The systematic hybridisation of different simple sequences to genomic DNA from a range of organisms confirmed the abundance of these sequences in eukaryotes, and many more types of tandemly arrayed simple sequences were found *(Tautz and Renz 1984).* Tautz *et al. (1989)* later showed that simple sequence motifs were five to ten-fold more frequent than equivalent random motifs.

Microsatellite loci are highly polymorphic. Variation at these loci is reflected in the length of the repetitive sequence, ascertained following visualisation of polymerase chain reaction (PCR) amplification products using standard electorphoretic techniques. Savatier *et al. (1985)* first demonstrated the hypervariable nature of STRs by cloning a 5.5 kb fragment of the β-globin gene in chimpanzee DNA and comparing it with a corresponding human sequence. Di-, tri and pentanucleotide repeat arrays were found to vary in length between the two species displaying either amplification or contraction of the number of repeat elements.

Few functions for these sequences have yet been determined and they are therefore generally considered as neutral markers. The amplification of dinucleotide repeats using human primer sequences suggests that microsatellites are highly conserved in terms of chromosomal location and also that the origin of these loci predates the divergence between humans and chimpanzees *(Deka et al 1994).* The conservation of genomic location might therefore suggest functional conservation. Despite the general assumption of neutrality it has been suggested that microsatellites may play a role in recombination *(Pardue et al 1987),* however were this the case then haplotypic information among closely linked loci would be disrupted and this has been shown not to be the case *(Morral et al 1993; Pena et al 1994).* Additionally, Hamada *et al. (1984)* have shown that dinucleotide repeats can affect the regulation of transcription of at least one gene in mammalian cells.

Trinucleotide repeats have been found to be associated with some human neurodegenerative diseases such as Huntington's disease, Fragile X syndrome, Myotonic dystrophy and spinobulbar muscular atrophy *(Sutherland and Richards 1995).* The number of trinucleotide repeats in an array have been shown to be associated with increases in symptom severity through time in genealogical pedigrees. This phenomenon is known as anticipation, the complexities of which are only beginning to become known. In disease patients the number of trinucleotide repeats was found to have expanded far beyond the length range found in the normal population, thus disrupting the structure of the encoded protein *(Willems 1994).* It has been shown that the expansion of the number of repeats is not a single step process but that increasingly large, and hence unstable, arrays are increasingly predisposed to further mutation in successive generations. Length thresholds exist for all of these diseases, which once passed, result in the manifestation of

disease symptoms, increasing in severity as the array length increases even further *(Richards and Sutherlands 1994).*

However, no general function for microsatellites has been shown. Probably one of the more important demonstrations of function neutrality for these loci was shown by Tautz and Renz *(1984).* They detected simple sequences in the metabolically inactive micronucleus of the protozoan *Stylonychia*, but not in the metabolically active macronucleus, which is derived from the micronucleus by chromosome diminution.

The processes of mutation of simple repetitive sequences has caused much debate in recent years *(Levinson and Gutman 1987, Schlotterer and Tautz 1992, Richards and Sutherland 1994)* although it is now generally agreed that allelic states at STR loci are generated by a mechanism of slipped-strand mispairing during DNA replication *(Levinson and Gutman 1987; Weber and Wong 1993).* Direct evidence for slippage comes from the distribution of mutations in a large study *(Weber and Wong 1993).* Precise details of the slippage mechanism are still poorly understood although it is known that slippage occurs during strand replacement by an out of frame pairing of repeated sequences *(Levinson and Gutman 1987).* As repeats gain more units a greater substrate for slippage is provided and therefore the expansion or contraction of repeat numbers increases. *In vitro* experiments using synthetic oligonucleotides and a variety of polymerases have indicated that the rate of slippage is dependent on the size of the repeat unit and on the sequence of that unit. For example, dinucleotides have been found to have the greatest rate of mutation, those with $(dG\text{-}dC)_n$ rich regions have been found to have the slowest *(Schlotterer and Tautz 1992).* It has also been suggested that the rate of mutation is influenced by the copy number of the STR prior to replication *(Stephan and Kim 1998),* although this has been contested *(Weber 1990, Shriver et al 1993).*

The usefulness of microsatellites for inferring population relationships depends largely on the estimation of a correct rate of mutation for these loci. A great deal of variation in the estimates of mutation rates has been published. For example, the average per locus mutation rate for dinucleotides was estimated from 15 autosomal microsatellite repeats as 5.6×10^{-4} *(Goldstein et al 1995).* Average mutation rates of di-, tri- and tetranucleotide repeats at autosomal loci have been estimated as 1.5×10^{-4} *(Underhill et al 1996)* and 2.1×10^{-3} per locus per generation *(Weber and Wong 1993)* which is very close to the rate estimated for Y STR loci (0.2%) *(Heyer et al 1997).* Slower mutation rates at Y STR loci have been determined, ranging from 1.2×10^{-4} - 4.8×10^{-4} *(Cooper et al 1996).*

Estimates of mutation rates at X chromosome loci also vary, ranging from 1.0×10^{-2} *(Mahtani and Willard 1993)* to 1.5×10^{-3} *(Weissenbach et al 1992).* Although these rates are estimated from averages across loci, Brinkman *et al. (1998)* have shown that the mutation rates at different loci differ by several orders of magnitude and that at one locus different alleles displayed different mutation rates. Chakraborty *et al. (1997)* have shown that di- tri- and tetranucleotide repeats have mutation rates inversely related to their motif size. For example, dinucleotides were found to have mutation rates 1.5 - 2 times higher than tetranucleotides, non-disease causing trinucleotides having rates intermediate to di- and tetranucleotides.

The complexity of microsatellite mutation is still not fully known. However, it is generally agreed that STR loci mutate in a single-step fashion, either by the addition or deletion of one repeat unit *(Shriver et al 1993, Valdes et al 1993).* The model proposed for the mutational behaviour of microsatellites is commonly known as the stepwise mutation model (SMM). The SMM was originally conceived to explain the regularities in the distribution of allele frequencies that could be determined by protein electrophoresis *(Ohta and Kimura 1973).* More recently it has been revised in attempts to account for the patterns of length variation at microsatellite loci. The SMM differs from the infinite alleles model of neutral mutation in that it assumes that there are only two states to which an allele can mutate. Evolutionary models are therefore based on the assumption that the divergence of alleles is proportional to the number of mutational steps between them.

Generally, allelic distributions at these loci conform to this model, although larger mutations have been observed. For example, in a pedigree analysis only 78% of mutations were due to the gain or loss of a single repeat, the remaining 22% due to larger non-stepwise events *(Weber and Wong 1993).* Others have also observed larger multi-step mutations in studies of copy number changes in extensive pedigrees *(Di Rienzo et al. 1994; Primmer et al 1996).* It has been suggested that the influence of point mutations within a repeat array on slippage may cause multi-step mutations by causing imperfect repeats *(Weber 1990)* and that the introduction of a point mutation may also split an array thereby reducing the size of an array in one large step *(Bell and Jurka 1997).* Recently Sajantila *et al. (1999)* identified most mutational events resulting in small gains or losses of repeats of between one and three repeats, observing few large multistep gains or losses of repeats.

It has also been suggested that the mutation process may be asymmetric, preferably adding repeats rather than deleting them. Were this the case then the mean of allelic distributions might tend towards infinity. However, constraints acting on allele size at microsatellite loci have been shown to limit the range and variance of microsatellite alleles *(Garza et al 1995).* Recently, Palsboll *et al. (1999)* have suggested that single strand slippage of partial repeats may provide a mechanism to counteract the infinite expansion of microsatellite arrays. But such constraints have been shown not to retard accurate estimations of genetic distances between two populations using dinucleotides. However, constraints may interfere when tetranucleotides are used *(Reich and Goldstein 1998).* Garza *et al. (1995)* accounted for this constraint by predicting that smaller alleles tend to gain repeats whereas larger alleles tend to lose repeats. This model, however, may be difficult to integrate into models for evolutionary analysis.

Other constraints may also be imposed on microsatellites. For example, Stephan and Kim *(1998)* assessed the persistence time of STR arrays in a population given a number of parameters including changes in array size and changes in population size. They showed that the persistence time of a microsatellite array is highly dependent on the population size and that persistence time increases sublinearly with an increase in population size. Because small alleles are more common in a population, the loss of larger alleles in a population will be affected by processes of drift. In this way the persistence of finite arrays in a population are governed inversely by the size of the population.

Evolutionary studies using microsatellites have been performed in a large number of different species *(MacHugh et al 1994; Buchanan et al 1994, Morin et al 1994).* Perhaps the widest use of microsatellites in evolutionary studies has been the modelling of human population histories *(Bowcock et al 1994; Deka et al 1995, Goldstein et al 1995).* The most recent use has been in Y chromosome studies.

1.4 Aims and objectives

The primary goal of the work described in this thesis is to investigate the genetic origin of the people of Ireland. In particular, the genetic relationship between the population of the island and continental European populations is considered. For this reason some European populations have been included in the study to provide a greater European context for the Irish population. Additionally the inclusion of an African population provides the scope to investigate the European populations at a broader level.

The following three chapters describe the assessment of genetic diversity and the genetic relationships within and between these populations using different genetic systems. The first approach analyses the variation at microsatellite haplotypes on the X chromosome. The second approach investigates population relationships using Y chromosome specific genetic markers. Initially unique event polymorphisms subdivide the populations into discrete Y chromosomal lineages. Following subdivision of the samples the genetic diversity within those lineages is assessed using Y chromosome microsatellite haplotype data.

Chapter 2:

Genetic diversity on the X chromosome

2.1 Introduction

2.1.1. Disease genes and polymorphic loci

Disease genes located on the X chromosome display pathology predominantly in males. Recessive mutations are expressed in the hemizygous state in males whereas females may be carriers of the diseases. The most common X-linked disease in humans (1 in 2,000) is Fragile X syndrome *(Sherman et al. 1985)*. Other disease genes located on the X chromosome include X linked spinal and bulbar muscular atrophy *(Spiegel et al. 1996)*, Barth syndrome *(Orstavik et al. 1998)*, X linked lymphoproliferative disease *(Sayos et al. 1998)*, Hunter disease *(Gort et al. 1998)*, Duchenne muscular dystrophy *(Worton et al. 1988)*, and X-linked arthrogryposis *(Zori et al 1998)*.

The presence of a large number of disease genes on the X chromosome has led to the extensive mapping of the chromosome *(Drayna et al. 1984, Starr and Wood 1987, Aldridge et al. 1984)* yielding numerous informative polymorphic markers. Two major polymorphic types have been utilised in linkage analysis - restriction fragment length polymorphisms (RFLPs) and simple tandem repeat (STR) polymorphisms. Population studies of variation at these and other loci have detected significant differences in allele frequencies between diverse populations. Frequencies of the B2 allele at an RFLP site on the X chromosome, RC8, differ significantly between Germans and Turks *(Schurmann et al. 1987)* but not between English and French *(Rahuel et al. 1984)*. Probe LI.28 detects significant differences in allele frequencies between English and Indian populations but not between English and Nigerian or Indian and Nigerian populations *(Papiha et al. 1988)*. Additionally probes DXS9 and DXS7 have detected significant allele frequency differences between five diverse ethnic groups *(Wadhwa et al. 1989)*.

Evolutionarily diverse population groups have been shown to exhibit differential allele frequencies at variable microsatellite loci. Occasionally private alleles are detected at loci that may serve as diagnostic markers for a sub-species or race within a species *(Neel 1973; MacHugh et al. 1997)*. Additionally, some RFLP loci on the X chromosome detect population specific alleles. For example, at DXS9 allele B*3 is found mainly in European populations, and has been detected very rarely in non-European populations *(Wadhwa et al. 1989)*.

Differential allele frequencies and the detection of population specific alleles on the X chromosome have been used to infer human population history. Most of these studies conform to patterns of global human diversity detected at other polymorphic loci *(Summers 1987; al-Maghtheh et al. 1993; Cruciani et al 1994; Papiha et al. 1991).* Additionally, clines of allele frequencies have been detected including a north-south cline of allele *2 detected with the LI.28 probe in European, African and Asian populations *(Papiha 1991)* and an east-west cline of allele *2 detected with the probe L754 in European and Asian populations *(al-Maghtheh et al. 1993).*

The variation at RFLP loci on the X chromosome has not been reported as being less than variation at similar loci on autosomes. Recent studies suggest, however, that microsatellites on the X chromosome are up to 30% less variable and less densely distributed than microsatellites on autosomes *(Stephan and Kim 1998).* This may be explained either by the smaller effective population size of X chromosomes compared to autosomes or perhaps by lower mutation rates on the X chromosome. A comparative study of 33 X linked genes and 238 autosomal genes in rodents has detected a lower rate of synonymous substitution in X linked genes than in autosomes. The authors have suggested that this cannot be explained by the higher mutation rate in the male lineage, nor by strong purifying selection, and that a lower mutation rate on the X chromosome is the most consistent explanation *(McVean and Hurst 1997).*

2.1.2. The genetics of the X chromosome

It is expected that the mode of inheritance of X chromosomes conforms to Mendelian segregation and inheritance in the same manner as autosomes. There is evidence, however, to suggest that X chromosome inactivation or imprinting may affect Mendelian transmission patterns from mother to son. Females exhibiting skewed X inactivation patterns have been found to be three times more likely to transmit alleles at loci located on the inactive X chromosome than on the active X chromosome *(Naumova et al. 1995).* Additionally, a strong bias in favour of the inheritance of alleles at loci on the X chromosome arising from the maternal grandfather has been detected *(Naumova et al. 1998).*

2.1.3. Isolation of DNA haplotypes

A haplotype may be defined as a DNA segment with multiple informative polymorphic sites which segregate as a non-recombining unit. The hemizygous male state of the X chromosome makes it an ideal system for the determination of haplotypes from an array of linked loci.

In diploid organisms the only other convenient approach to the isolation of DNA haplotypes requires restricting analysis to non-recombining regions of the genome *(Harding et al. 1997)* and involves the sequencing of maternally inherited mitochondrial DNA *(Avise 1994)* and analysis of paternally inherited Y chromosomes *(Jobling and Tyler-Smith 1995).* Other than mtDNA and Y chromosome haplotypes, the isolation of haplotypes from diploid systems can be technically difficult and often ambiguous.

For autosomal loci, information on the genotypes of both parents in addition to the individual sample is required to directly infer the gametic phase of the individual *(Excoffier and Slatkin 1998).* This enables the determination of haplotypes by detecting those alleles that have arisen from each parental chromosome by resolving the complication of independently segregating alleles.

Alternatively, in the absence of parental data, an expectation-maximisation (EM) algorithm may be employed in the calculation of maximum-likelihood estimates of haplotype frequencies under the assumption of Hardy-Weinberg equilibrium *(Excoffier and Slatkin 1995; Slatkin and Excoffier 1996).* For example, maximum-likelihood estimates of haplotype frequencies from diploid population data can be calculated using the program HAPLO *(Hawley and Kidd 1995).* However, there are limitations associated with this, most notably increased computation time when a larger number of samples and a larger number of polymorphic loci are used *(Excoffier and Slatkin 1995; Slatkin and Excoffier 1996).* Also, simulation studies have shown that this approach to the assessment of linkage disequilibrium is not as powerful as direct analysis from known haplotype frequencies *(Excoffier and Slatkin 1998).* Direct haplotyping of closely linked polymorphic markers from individual sperm has also provided another alternative for obtaining molecular haplotypes for genetic analysis *(Crouau-Roy and Clayton 1995).*

2.1.4. Haldane's Rule

It has been demonstrated that deleterious mutations on the X chromosome do not persist within particular haplotypes over a large number of generations. The population dynamics of the X chromosome predicts that recessive mutations will survive for only 3 generations as reproductive fitness, associated with X linked disease, is severely retarded. It is predicted that the loss of these mutations through selection will be balanced by *de novo* mutations at disease loci. Conversely, autosomal recessive disease loci may associate with particular haplotypes for a time determined by the mutation rate of the disease gene *(Haldane 1935).* This rule is, of course, dependent on both the severity of the trait and the age of onset of disease.

2.1.5. Haplotypes of disease genes on the X chromosome

A number of population studies of haplotypes associated with disease genes on both the X chromosome and autosomes have made inferences about the population dynamics of disease loci.

On the X chromosome the Fragile X gene has been found to be highly associated with two polymorphic markers flanking the disease locus. Analysis of haplotype frequencies in Fragile X patients and in normal controls detects distinct haplotypes associated at high frequency with disease chromosomes and at low frequency with normal chromosomes *(Richards et al 1992; Oudet et al. 1993).* This suggests that the Fragile X genotype originated within a small number of founding haplotypes which have persisted in the population. This contradiction of Haldane's prediction of the loss of haplotype associations after a small number of generations may be explained by the phenomenon of anticipation in expanding triplet repeat genes. Mutation rates at these types of disease loci are determined by the copy number of the product of the previous mutation and therefore do not have fixed mutation rates *(Weber 1990).* Premutations consisting of an intermediate number of repeat units are asymptomatic. Premutations therefore avoid selection and permit the survivorship of Fragile X lineages in the population. However, by contrast, in a study of the Jewish Ashkenazi population no founding Fragile X chromosomes were identified *(Pesso et al. 1997).*

2.1.6. Haplotypes to determine population history

Not only can haplotype analysis resolve past histories of disease chromosome lineages but it can also assist in the inference of past population histories. For example, at the myotonic dystrophy locus diversity is greatest in African populations with non-African populations displaying a subset of this diversity. Strong linkage disequilibrium between myotonic dystrophy and flanking markers has confirmed the presence of founding haplotypes in African populations *(Harley 1991)* suggesting that the mutation responsible for the expression of myotonic dystrophy arose in an African population prior to the split between Africans and non-Africans. This concords with an African origin for modern humans *(Goldman et al. 1995; Tishkoff et al 1998).*

The geographic distribution of haplotypes associated with autosomal disease loci together with patterns of diversity may aid in the inference of population histories. For example, the diversity at CD4-receptor gene haplotypes indicates a greater variation in African populations than non-African populations *(Tishkoff et al 1996).* A cline across Europe, often associated with the movement of Neolithic peoples is found in the distribution of the ΔF508 cystic fibrosis mutation *(Morral et al. 1994).* The CCR5-Δ32 AIDS resistance allele at the CCR5 chemokine locus exhibits a north-south cline within Europe *(Stephens et al 1998).* The distribution of the apolipoprotein B-100 gene suggests an Asian origin for R3500W mutations *(Tai et al. 1998).* Haplotype analysis of oculopharyngeal muscular dystrophy detects a founder effect in an Hispanic American population *(Grewal et al 1998).* Two common mutations causing factor XI deficiency in Jews have been found to have arisen in two distinct populations, one of ancient Middle Eastern origin and another of more recent European origin *(Peretz et al. 1997).* Haplotype analysis at the phenylketonuria locus indicates a Nordic ancestry for Icelandic populations with no evidence for an Irish influence *(Guldberg et al. 1997).*

2.1.7. Population genetics of the X chromosome

Few studies have been conducted using selectively neutral polymorphisms, unassociated with disease genes, on the X chromosome despite the large number of recorded STRs identified on that chromosome. One recent study of linkage disequilibrium at loci on the X chromosome in a number of Scandinavian populations has introduced the

use of X chromosomal microsatellites in population genetic studies. Laan and Paabo *(1997)* assessed the extent of linkage disequilibrium between a set of microsatellite loci on the X chromosome in a number of geographically closely related populations. A greater proportion of linkage disequilibrium was detected in the Saami population compared to the Finnish population. This was interpreted as evidence that the Finnish population is an expanded population whereas the Saami are at equilibrium and have been of constant size for many generations.

A study of neutral DNA single nucleotide polymorphisms from exon 44 of the X chromosome dystrophin gene in non-disease subjects has estimated an origin for all human populations ~100,000 - 200,000 *YBP (Zietkiewicz et al. 1998)* which is in agreement with many other molecular genetic studies of human population history. Ancestral states for 36 novel alleles detected at polymorphic sites in exon 44 were determined from a comparison with non-human ape sequences. Seventeen of the new alleles were present in all global populations suggesting that they had arisen prior to the separation of African populations and non-African populations. An estimated age of 200,000 years for these alleles was described. Any polymorphisms that were found to be younger were found only in either African or non-African populations but not both. This asserts that the split between African and non-African populations occurred prior to 200,000 years ago. Additionally the highest levels of variation at these loci were detected in African samples indicating that African populations are older than other worldwide populations *(Zietkiewicz et al. 1998)*.

A non-coding 5 Mb segment in the Xq22 region of the X chromosome was recently examined in 23 male subjects. A total of 102 segregating sites were detected in 58 1.5 kb segments. Seven of the segregating sites were shared between African Pygmies, Ashkenazi Jews and Europeans. Two sites were shared between Pygmies and Europeans, two were shared between Pygmies and Ashkenazim and seventeen sites were shared between Europeans and Ashkenazim. Europeans were found to display the least amount of diversity, calculated as the average pairwise difference (15.0), and Pygmies displayed the greatest diversity (23.6). The Ashkenazim population had an intermediate diversity to the two (16.8) *(Anagnostopoulos et al. 1999)*.

In 4,200 bp of sequence at the pyruvate dehydrogenase E1 α subunit (PDHA1) locus, 35 males divided into groups of African and non-African origin were found to have significantly different sequence haplotypes. Diverse sequences were found associated uniquely either with the African or non-African populations. Divergence dates between

the two groups were estimated at about 200,000 *YBP* and suggests that the emergence of modern humans occurred in an ancient subdivided population with localised geneflow and selection allowing for parallel evolution. Unlike mtDNA studies, the PDHA1 locus does not show any evidence for population expansions in global populations *(Harris and Hey 1999).*

Kaessmann *et al. (1999)* recently sequenced a ~10 kb non-coding region of low recombination (one eighth of average X chromosome recombination) at Xq13.3. Comparisons of 69 male X chromosome sequences representing all major global linguistic population groups revealed 33 polymorphic sites. A greater diversity in African populations was demonstrated. Deep divergences were also found in Asian populations. By sequencing an orthologous sequence in a chimpanzee and a gorilla and assuming a divergence between chimpanzees and humans of 5 million years, the most recent common ancestor for all X chromosomes was estimated as 535,000 (+/- 119,000) *YBP*.

2.2 Materials and methods

2.2.1. Sampling strategies

Several strategies were employed for the sampling of subjects for human genetic diversity study. Most importantly only male samples were collected. Generally, in order to ascertain the geographic origin of a sample, population genetic surveys require that the birthplaces of all 4 grandparents are known *(see Passarino et al 1998).* Here, however, the novel coupling of male samples and the analysis of the X chromosome requires that only the birthplace of the mother be known. By tracing the maternal lineage a sample can be assigned to a population group one generation back. If the birthplaces of both maternal grandparents are known, and these are the same, then the origin of the X chromosome can be identified two generations back.

Recent population migration, particularly rural-urban transition, resulting in the genetic admixture of often distinct, previously isolated populations highlights the importance of knowledge of geographic ancestry of chromosomes. Additionally, some societies are patrilocal, suggesting that females move from their birthplace to the birthplace of their spouse. Therefore knowledge of X chromosomal maternal origin is highly beneficial.

Over 500 X chromosomes in Ireland were collected and about 30 from each of Denmark, Turkey, Togo, Greece and Malta. Most population genetic surveys to date have collected samples in the country of origin without knowledge of individual family history. For example, in a study of European Y chromosome diversity Cooper *et al. (1996)* collected blood samples from a hospital in East Anglia, UK, without any information relating to each sample. The selection of samples with non-English sounding surnames was, however, avoided. Although all individuals were classified as English neither the geographic nor the ethnic origin of the samples were known *(David Rubenzstein pers. comm.).*

For this study it was decided that the use of a non-invasive sampling procedure would allow for the sampling of male subjects in universities and local educational institutions and detailed individual knowledge of family history could be provided by voluntary questionnaire.

2.2.2. Ethical considerations for human sampling

The collection of human genetic material has raised strong ethical questions in recent years, particularly with commercially valuable medical genetic research in which often illiterate, underprivileged and uneducated societies may be exploited. The patenting of human DNA sequences for the development of biotechnical products has been opposed and questions have been raised regarding the safeguarding of indigenous human population rights[1].

For legal and ethical reasons it was necessary that informed consent should be gained from all sample providers. The American Council of International Organisations for Medical Sciences defines informed consent as follows: *"Consent is informed when it is given by a person who understands the purpose and nature of the study, what participation in the study requires a person to do and to risk, and what benefits are intended to result from the study."*

In each location, before any sampling was performed a short informative presentation was made outlining the nature, goals and methods of the study. A detailed introduction to human genetic diversity was given where appropriate. Any questions raised by the participants were answered accordingly.

All participants under the age of 18 years required a signed parental consent form. Parental consent forms were only required for the sampling of students (believed to be over 16 years old) in St. Muiredach's College, Co. Mayo. All other individuals sampled were understood to be over 18 years old.

Each participant completed a consent form that requested information about family geographic history and required the participant's signature. All samples and consent forms were numbered and anonymised thereafter.

---

[1] World Medical Association "Declaration on the Human Genome Project", Marbella, September 1992; Comite Permanente Dos Medicos Da Comunidade Europeia "Resolution regarding the proposed directive on the protection of biotechnological inventions", Estoril, 13/14.11.1992
See http://www.psych.bangor.ac.uk/deptpsych/Ethics/HumanResearch.html and http://pw2.netcom.com/~alalli/pageone.html for discussions on ethical issues in medicine and science.

2.2.3. Populations sampled

Samples were collected from five ethnically and geographically diverse populations. Most of the analyses concentrate on the Irish, Turkish and Togolese populations.

Ireland

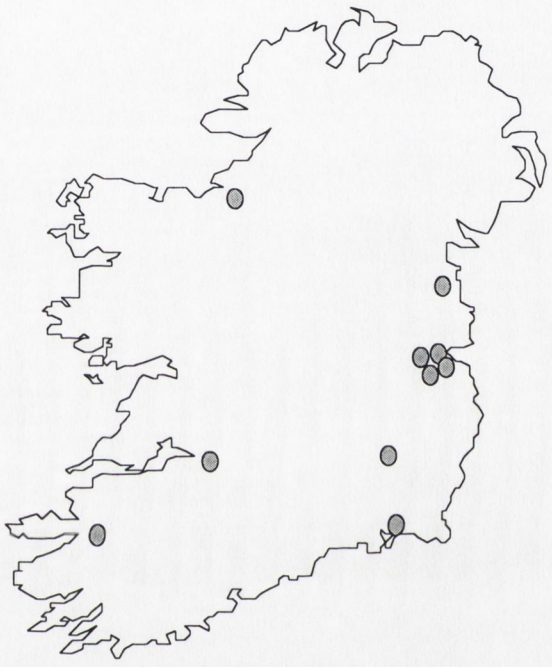


Figure 2.1: Approximate sampling locations in Ireland

The sampling of Irish male individuals was carried out over a period of 18 months between October 1995 and April 1997. Initially sampling was carried out in Natural Sciences, Medicine, Dentistry and Pharmacy practical classes in the University of Dublin, Trinity College. It became clear however, that there was a strong bias towards samples originating in Dublin when, in fact, a widespread representative national sample was sought. A countrywide sampling effort was required. This involved the collection of samples from student volunteers in the following educational institutions around the country:

- University College Dublin
- University of Limerick
- R.T.C. Carlow
- R.T.C. Tallaght
- R.T.C. Waterford
- R.T.C. Dundalk
- R.T.C. Tralee
- St. Muiredach's College, Co. Mayo
- The Kings Hospital School, Co. Dublin

Most of the sampling groups ranged in size from 30 - 100 individuals. In most cases less than half of these were males. In total more than 1,400 Irish male samples, believed to be unrelated, were collected in the 18 month sampling period.

In addition to the male samples ~95 female samples were collected for use as diploid allele size standards in the genotyping of the male samples. No information regarding family geographic history was recorded.

Greece

Eighteen unrelated male samples were collected from the University of Thesaloniki, Greece, by Mr. Kostas Papadopoulos. Consent forms were signed but no information (other than being of Greek descent) concerning the geographic history of the samples was obtained.

Denmark

Thirty-three unrelated male samples were collected from the University of Copenhagen, Denmark by Dr. Bo Simonsen. Consent forms were signed but no detailed information (other than being of Danish descent) concerning the geographic history of the samples was obtained.

Turkey

The present landhold of Turkey includes the eastern European region known as eastern Thrace, bounded on the north by Bulgaria and the Black Sea and on the west by the Aegean Sea and Greece, and the western Asian region of Anatolia, bounded on the east by Armenia, Georgia and Iran, on the south by Iraq, Syria and the Mediterranean Sea and on the north by the Black Sea. The current population of Turkey is estimated as 63,400,000 people of which 55% live in rural areas by means of subsistence farming. Although widely known as Turks, the Turkish people are made up of a number of ethnic groups of diverse origin.

Fifty-five unrelated male samples were collected from 3 rural regions (near Ankara, Erzurum and Izmir) of Turkey by Dr. Ronan Loftus and Dr. Okan Ertugrul. In some cases family histories were recorded along with signed consent forms. The samples were not divided into subpopulations, however, as sufficient sample sizes were not obtained for each region. Therefore, despite the complex history and heterogeneity of the Turkish peoples *(Comas et al. 1997),* all of the Turkish samples were pooled as one population.

Togo, West Africa

The West African state of Togo encompasses 56,000 sq. km and is bounded on the north by Bukino Faso, on the west by Ghana and on the east by Benin. The southern border constitutes 50 km of coastline with the Gulf of Guinea. The present population of Togo is estimated as 4,900,000 people (July 1998 est.). The population is predominantly rural with over 70% living in rural areas, although the estimated annual growth of the urban population in the capital, Lomé, is 6.1%.

The population of Togo is made up of over 37 ethnic African tribes of which the largest and most important are the Ewes and the Mina in the south and the Kabye in the north. The Ewes, who were hunters and farmers, moved to the region that is now Togo from the Niger river valley between the years AD 1100 and 1300. Other important ethnic groups include the Akposso, Basser, Kotcoli, Tehambe, Konkomba, Lamba, Hansa, Peuhl, Mossi, Tamberma, Guni, Losso, Ontachi and Gemuy tribes.

Forty-one African samples were collected by Dr. Dan Bradley, with the help of Dr. Balabadi Dao in Togo, West Africa. All samples were collected in the region of Sokodé, Togo.

2.2.4. Population assignment based on maternal origin of the X chromosome

Individuals sampled in Ireland were requested to record details of their birthplace and the birthplaces of their mother and maternal grandparents. Information recorded by the sample providers was entered into a File Maker Pro™ database designed for the purpose of this survey by Aoife MacLysaght *(Department of Genetics, Trinity College, Dublin).*

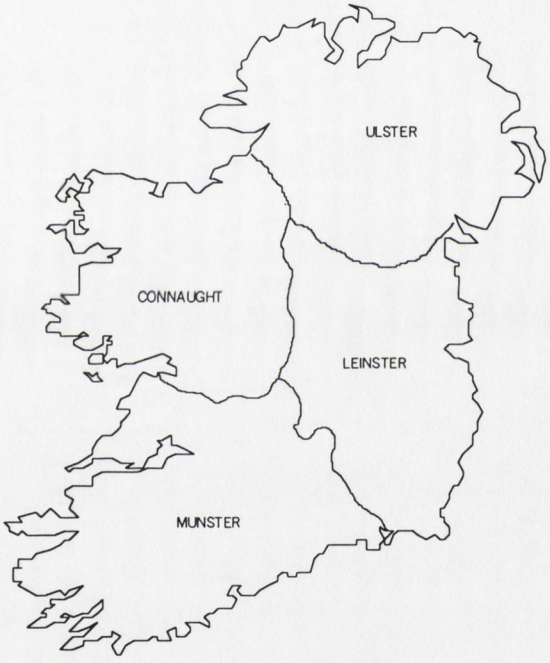


Figure 2.2: The four ancient provinces of Ireland, delimited by the county borders in the lists below.

On the basis of this information, samples were grouped into populations corresponding to the four ancient provinces of Ireland. When the birthplaces of the maternal grandparents were the same then that region was preferred to assign the sample to a population. When the mother's birthplace was known then this was used. It was rarely necessary to use the birthplace of the sample provider as the population delimiter. However, it became clear over the course of 18 months that knowledge of family history

was often poor, particularly when communicating information regarding grandparental geographic history. Provincial subdivisions were made according to **table 2.1.**

Ulster	Munster	Leinster	Connaught
Antrim	Clare	Carlow	Galway
Armagh	Cork	Dublin	Leitrim
Cavan	Kerry	Kilkenny	Mayo
Derry	Limerick	Laois	Roscommon
Donegal	Tipperary	Longford	Sligo
Down	Waterford	Louth	
Fermanagh		Meath	
Monaghan		Offaly	
Tyrone		Westmeath	
		Wexford	
		Wicklow	

Table 2.1: Provincial subdivisions into which each sample was classified.

A number of samples had their origin outside Ireland. These samples were grouped according to the country of origin (e.g. England).

2.2.5. Sampling procedure

Following a demonstration of the sampling procedure, cheek cell swabs were collected by the participants themselves. Buccal squamous epithelial cells were collected on sterile nylon bristle *CytoSoft*™ cytology brushes by gently scraping the inside of both cheeks, avoiding contact with the teeth and gums. Following the collection procedure the brushes were carefully replaced in their packages, numbered and sealed for transportation back to the lab. This self-administered procedure confirmed the personal consent of each participant.

2.2.6. DNA extraction from buccal cell swabs

Although sample collection was done in the field, DNA extraction was carried out in the laboratory. In all cases DNA was extracted within 14 days of collection. Controlled experiments demonstrated that DNA yields did not dissipate up to 14 days after collection provided the brushes were well sealed and stored in a cool, dry place. Once the brushes were returned to the laboratory they were removed from the packages and the bristle ends

were cut, using ethanol/heat sterilised scissors, into labelled 1.5 ml Eppendorf tubes and stored at 4 °C until extraction. For the first ~100 extractions, extraction controls were carried out to ensure that the sample material was free of contamination.

Whole genomic DNA was extracted from the buccal cell samples using the Biorad Instagene Matrix™. The Eppendorf tube was opened and 500 µl of 50 mM NaOH was added. The cap was replaced and the tube was vortexed on high for 60 seconds. The tube was then placed in a heatblock at 95 °C for 10 minutes before the addition of 50 µl of 1 M Tris pH 8.0. The tube was vortexed for 30 seconds. The Biorad Instagene Matrix™ was shaken well before use. Using a 1 ml pipette tip 200 µl Biorad Instagene Matrix™ was added to the tube. The tube was then incubated in a waterbath at 56 °C. After 30 minutes the tube was removed and vortexed on high for 10 seconds. A final incubation in a heatblock at 100 °C for 8 minutes was then carried out before a further vortex for 10 seconds. The brush was then removed from the tube and discarded using ethanol/heat sterilised forceps. The tube was then centrifuged at 12,000 rpm for 3 minutes. The supernatant, containing whole genomic DNA, was removed and transferred to another labelled Eppendorf tube *(Gill et al. 1997)*. All stock samples were stored at -20 °C.

2.2.7. Storage of sample aliquots in 96-well microtitre plates

Samples were aliquoted in 200 µl batches into 96-well microtitre plates and mixed with 20 µl of a dilute solution of cresol red dye. The samples were plated in a particular sequence for use with a Hamilton multi-channel loading syringe leaving empty wells at appropriate intervals for size standard samples. Aliquot plates were stored at 4 °C.

2.2.8. Assembly of a panel of linked polymorphic X chromosome microsatellites

A number of inherited disease genes map to the Xq22 region in the proximal half of the long arm of the X chromosome *(Davies et al. 1991; Vetrie et al. 1994)* including Pelizaeus-Merzbacher disease *(Willard and Riordan 1985)*, X-linked agammaglobulinemia *(Vetrie et al. 1993)*, Fabry disease *(Bernstein et al. 1989)* and Alport syndrome *(Barker et al. 1990)*. The region at Xq22 between markers DXS366 and DXS87

provided a well mapped region of the X chromosome to search for polymorphic repeat sequences.

Highly complex genomic DNA libraries have been constructed *(Brownstein et al. 1989, Larin et al. 1991)* from the yeast artificial cloning system *(Burke et al. 1987)* which has enabled the isolation of overlapping clones from particular chromosomal regions *(Schlessinger et al. 1991)* Between markers DXS366 and DXS87, 22 cosmids were detected. Two overlapping cosmids, hsu177e8 and hsu693a8, which spanned ~100 kb were detected. The cosmid clone hsu693a8 contained few repeat regions. The cosmid clone hsu177e8 (GenBank accession # Z68694) was selected on the basis of the large number of repeat sequences located within its 40,822 bp. No gene sequences are believed to be present on hsu177e8 and therefore selection should not act in this cosmid region .

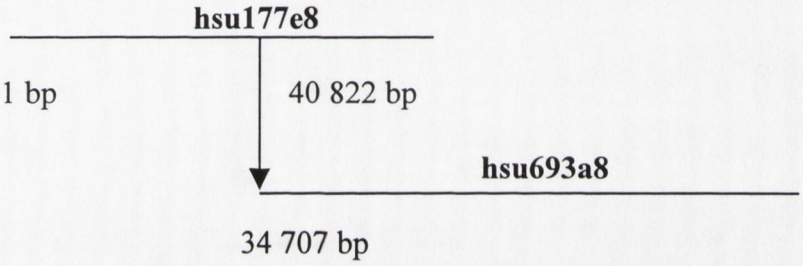


Figure 2.3: Overlapping cosmid clones on chromosome X located between microsatellite markers DXS366 and DXS87. The small region of cosmid hsu177e8 (~40kb) suggests that little recombination will occur between loci on this cosmid and that they will behave, in males, in a haploid manner.

2.2.9. Selection of suitable marker loci on hsu177e8

Cosmid hsu177e8 was found, on inspection by eye, to contain 61 repeat regions, 9 of which were microsatellite repeat regions. These were selected on the criterion of a high number of repeat units, deemed necessary for the detection of polymorphic differences between alleles. Each possessed 12 or more repeat units. Eight were comprised of dinucleotide repeats of the $(CA)_n$ type. $(CA)_n$ repeats are the most common dinucleotide repeat type in vertebrate genomes *(Hamada et al. 1982).* One trinucleotide repeat was found, comprised of a $(CAT)_n$ repeat motif.

MICROSATELLITE LOCUS	NUMBER OF REPEAT COPIES	REPEAT TYPE	POSITION ON COSMID	% SEQUENCE CONSERVATION	POLYMORPHIC / MONOMORPHIC
hsu177e8 - F3	17	CA	88 - 121	100%	polymorphic
hsu177e8 - F7	24	CA	1367 - 1414	100%	polymorphic
hsu177e8 - F8	15	CA	6535 - 6564	87%	monomorphic
hsu177e8 - F9	12	CA	12800 - 12847	81%	monomorphic
hsu177e8 - F13	19	CA	21120 - 21157	100%	polymorphic
hsu177e8 - F14	20	CAT	28904 - 28963	85%	polymorphic
hsu177e8 - F18	16	CA	29721 - 29752	94%	monomorphic
hsu177e8 - F23	20	CA	35011 - 35050	100%	polymorphic
hsu177e8 - F29	20	CA	39038 - 39077	100%	polymorphic

Table 2.2: Locus descriptors for microsatellite loci on cosmid hsu177e8. The number of repeat copies were >12 and <24, $(CA)_n$ repeats were abundant, and a high sequence conservation was favourable. Six of the originally selected loci were found to be polymorphic.

2.2.10. Design of primer sets from flanking sequences of repeat regions

Sets of primer sequences were designed from flanking sequences (approximately 200 bp of sequence both 5' and 3') surrounding the repeat regions. The following considerations were made when designing the primers to ensure complete specificity of the amplification reaction:

- Primers should be 18 - 25 bp in length
- GC content should be >40% - <60%
- Repetitive stretches of bases should be avoided
- No 3' complementarity between primers
- Where possible, two cytosine or two guanine residues should be present at the 3' end of each primer
- Opportunity for internal secondary structure should be minimal

Furthermore primers were designed with the aim to limit the size of the PCR product resulting from each primer pair. PCR products for all primer sets fell in the range ~100 bp - 250 bp.

Primer sequences and locus details are shown in **table 2.3.**

MICROSATELLITE LOCUS	PRIMER SEQUENCES 5' - 3'	APPROXIMATE ALLELE SIZE	$MgCl_2$ mM	Anneal °C
hsu177e8 - F3	CAG TGA ACT ATA AAT GCA CC TAA AGA GAG CTC ATC TAT TCC	164 bp	2.0	55
hsu177e8 - F7	AGC ACT CAG TAA TCG TTT GG CTT TTG CCA TAT ACA CAA GG	172 bp	2.0	55
hsu177e8 - F8	TTG AAA ATG CAG GAT CCT GG ATG GCA AGG AAA ACT GAC GG	224 bp	1.5	55
hsu177e8 - F9	GGC AGA GAT TGT CAG ACT GG ATA CCA TGC TTT ACC TCT AG	250 bp	2.0	55
hsu177e8 - F13	ACA TCT GAC CAT CTT CCT GG CGT GTA AAC CCT TAA CAA AGG	115 bp	1.5	55
hsu177e8 - F14	AGG TCA AAC AGA AAA TGT CC CCT GTC CCC AAC CAG AAT CC	133 bp	1.5	55
hsu177e8 - F18	AAA ATT CCC TCC ACA TTT CC ACA GAG GAT GAA GAT GAA AC	254 bp	1.5	57
hsu177e8 - F23	CCA GCA TGG GTG ATA GAG GG CTG CAT TTT GGC CTG TTT GG	194 bp	2.0	57
hsu177e8 - F29	AGC CTT ATT TGC CTG TTT CC ATG AGT GCC TAC CAA TTA CC	154 bp	2.0	55

Table 2.3: Forward and reverse primer sequences (5' – 3') and approximate allele sizes for cosmid hsu177e8 microsatellite loci.

2.2.11. Microsatellite Genotyping

The complete microsatellite genotyping of all samples was performed following the optimisation of the PCR reaction for each microsatellite.

2.2.12. Optimisation of PCR for each microsatellite

For each primer pair optimal $MgCl_2$ concentrations were determined by titrating $MgCl_2$ in a 10X PCR reaction buffer[2] with a range between 5 mM and 50 mM in increments of 5 mM (details for each primer pair are shown in **table 2.4**). PCR reactions were carried out in either a Hybaid OmniGene thermal cycler or a PTC-200 thermal cycler. Annealing temperatures for the PCR amplification were initially set to 55 °C.

Component	Volume	Final composition
10X PCR reaction buffer	2.0 µl	1X
Primer B (2 µM)	1.5 µl	0.3 µM
Primer A (2 µM)	1.5 µl	0.3 µM
dNTPs (3.32 mM each dNTP)	1.2 µl	200 µM
Taq polymerase	0.1 µl	1.0 unit
dH_2O	11.7 µl	...
DNA template	2.0 µl	~50 ng
Mineral oil	10.0 µl	...

Table 2.4: The composition of the PCR amplification reactions (20 µl final volume).

6X electrophoresis tracking buffer[3] was added to each PCR product. 10 µl of PCR product was then electrophoresed on a 1.5 % agarose gel with 0.5 µg/ml ethidium bromide in 1X TBE buffer[4]. The optimum $MgCl_2$ concentration was determined from visual analysis of electrophoresed products. It was often necessary to further optimise the specificity of the reaction by either increasing or decreasing the annealing temperature. Also it was often necessary to increase the number of cycles in each reaction to increase the quantitative yield of the PCR product.

[2] 500 mM KCl; 100 mM TrisCl (pH 9.0); 5 - 50 mM $MgCl_2$; 0.1% gelatin; 1.0% Triton X - 100
[3] 0.25% bromophenol blue; 0.25% xylene cyanol FF; 30% glycerol
[4] 1.0 M Tris-borate (ph 8.3); 0.02 M EDTA (pH 8.0)

PCR reactions were carried out with the following cycling parameters. A two minute denaturation step was followed by 30 - 40 cycles of amplification:

Denaturation	45 seconds	93 °C
Annealing	45 seconds	52 - 60 °C
Extension	45 seconds	72 °C

This was followed by a final extension step for 3 minutes at 72 °C.

2.2.13. PCR amplification of microsatellites

For the generation of microsatellite genotypes the PCR reaction components were modified to include α - ^{32}P dCTP [3,000 Ci/mM].

Component	Volume	Final composition
10X PCR reaction buffer	1.10 µl	1X
Primer mix (2 µM)	1.65 µl	0.3 µM
dNTP : dCTP (200 : 6.25 µl)	0.66 µl	200 µM : 20 µM
Taq polymerase	0.05 µl	0.5 unit
α - ^{32}P dCTP [3,000 Ci/mM]	0.05 µl	0.5 µCi
dH_2O	6.38 µl	...
DNA template	1.0 µl	~50 ng
Mineral oil	10.0 µl	...

Table 2.5: The composition of radiolabelled PCR amplification reactions (11 µl final volume)

PCR amplification of the sample DNA and radioactive labeling by random incorporation of α - ^{32}P dCTP was performed in Hybaid™ 96-well polycarbonate microtitre plates following the transfer of 1 µl DNA from the storage aliquoted 96-well microtitre plates. Each sample was overlaid with 10 µl mineral oil to prevent evaporation during PCR.

A mastermix of all PCR components (except the DNA template) was made and 10 µl aliquots were added to each sample using an Eppendorf repeating multipipette. PCR amplification was carried out in either a Hybaid OmniGene thermal cycler or a PTC-200 thermal cycler using the optimal conditions previously determined for each microsatellite.

Amplification was improved by covering the samples with aluminium foil to encourage even conduction of heat during thermal cycling.

PCR products were electrophoresed on 6% acrylamide denaturing sequencing gels [final composition: 6% acrylamide ; bisacrylamide (19:1); 50% urea; 100 mM Tris-borate (pH 8.3); 2mM EDTA (pH 8.0)][5] with 1X TBE electrophoresis buffer. In preparation for electrophoresis 10 µl formamide denaturing buffer[6] was added to each sample following PCR amplification. The samples were heat denatured for 3-4 minutes at 95 °C and then snap-cooled on ice.

Approximately 1.5 µl PCR product was loaded onto a pre-warmed gel using a Hamilton multi-channel loading syringe. The gels were secured on Gibco BRL vertical sequencing rigs and electrophoresed at 65 - 70 W for 2 - 4 hours depending on the size of the PCR product. After this time the gel was removed and transposed onto Whatman chromatography filter paper, covered with cling film and dried on a Savant SGD 4050 vacuum slab gel drier at 80 °C for 1 - 2 hours. The dried gel was placed in a film cassette with Agfa autoradiographic X-ray film and exposed for 24 - 48 hours depending on the intensity of the radioactive signal. The autoradiographic films were developed in an Agfa Curix film developer.

The microsatellite allele sizes were read manually from the autoradiograms by referring to the size standards on each gel (described below). All gels were double checked. The allele sizes were recorded in an Excel™ database file on the Macintosh computer.

2.2.14. M13 sequencing and female size standards

A panel of >30 female samples were PCR amplified and electrophoresed according to the above method. A number of samples were selected for use as size standard markers on the basis of heterozygosity and a large repeat size difference between the two alleles. The female size standards were then electrophoresed alongside sequenced M13mp18

---

[5] Sequencing gels were made using the Sequagel™ sequencing system. The following proportions of buffer:concentrate:diluent were used 10:25:65. Gels were polymerised with 400 µl of 10% APS and 40 µl TEMED solutions.

[6] 98% deionised formamide; 10 mM EDTA (pH 8.0); 0.025% bromophenol blue; 0.025% xylene cyanol FF

plasmid DNA and their allele sizes were determined by comparison with the known M13 sequence.

M13 sequencing was performed using *Sequenase 2.0* DNA sequencing kits. Sequencing was performed according to a modified *Sequenase* protocol.

The annealing reaction was carried out in a 37 °C waterbath for 15 minutes and then cooled to room temperature. The following reagents were required for the annealing reaction.

Component	Volume
Forward M13 sequencing primer (3 ng / µl)	4 µl
5X *Sequenase* buffer	8 µl
M13 template DNA (0.2 ug / µl)	28 µl

note: Forward M13 sequencing primer[7]; 5X *Sequenase* buffer[8]

While the annealing reaction was being carried out 4 termination tubes were labelled (dATP, dCTP, dTTP, dGTP) and 10 µl of each termination solution[9] was placed in each tube. Meanwhile the following labeling reaction was prepared and placed on ice.

Component	Volume
TrisCl (5 mM, pH 7.5)	15 µl
Dithiothreitol (DTT) (100 MM)	4 µl
5X *Sequenase* labelling mix	1 µl
α - ^{35}S dATP [1,000 Ci/mM)	2 µl

[10]

After the addition of 1 µl *Sequenase* enzyme, 22 µl of the labeling reaction mix was added to the annealing reaction mix. After 4 minutes 14 µl of the labeling reaction was added to each of the 4 termination solutions which were then placed in a 37 °C waterbath for 5 minutes. The termination reactions were then stopped by the addition of 16 µl formamide denaturing buffer. Following heat denaturation 4 µl of each base was electrophoresed alongside the female size standards

[7] 5' GTT TTC CCA GTC ACG AC 3'
[8] 200 mM TrisCl (pH 7.5); 100 mM $MgCl_2$; 250 mM NaCl
[9] 80 µM of each of dATP, dCTP, dGTP, dTTP; 50 mM NaCl; 8 µM of the corresponding ddNTP
[10] 7.5 µM of each of dCTP, dGTP, dTTP

2.2.15. Computational Methodologies

Most of the following computations were preformed on either an Apple Macintosh Performa 630 compute or a Gateway 2000 EV700 PC, unless otherwise specified.

Genetic distance measures

Genetic distances (D_{dm}, D_{sw}, R_{st}) were calculated using MICROSAT *(Erich Minch).* The genetic distances D_{asm} and D_{swasm} were calculated in SWAMP, a program written by Garret Taylor *(Department of Genetics, Trinity College, Dublin),* run on a unix work station.

Phylogenetic tree construction

Phylogenetic trees were constructed in the PHYLIP package *(Felsenstein 1989)* using the neighbour-joining method of *Saitou and Nei (1987)* implemented in the NEIGHBOR facility. The neighbour-joining method groups operational taxonomic units (OTUs) as neighbours in such a way as to minimise the total branch length of the tree. The performance of the neighbour-joining method is largely dependent on the accuracy of the genetic distance measure used. Trees were drawn using the DRAWGRAM and DRAWTREE accessories and were manipulated to their final presentation in MacDraw Pro™ on the Macintosh computer.

Maximum parsimony networks

Maximum parsimony trees were created in the PHYLIP package *(Felsenstein 1989).* Allelic information was recoded to binary format using the program FACTOR in PHYLIP that factors multistate characters to binary characters. The binary recoded data set was then entered into PENNY in PHYLIP which attempts to find all the most parsimonious trees possible from the input data set. PENNY uses a "branch and bound" algorithm, which attempts to create trees by joining the first two OTUs and adding the third OTU to all possible places and then the next OTU again to all possible places. There are therefore a very large number of possible trees depending on the number of OTUs in the data set. The

input data set was randomised in order to avoid bias from the binding order of branches. Depending on the number of OTUs in the data set the number of groups of 100 most parsimonious trees searched for ranged from 1,000 - 10,000. Output files from PENNY were entered into CONSENSE which records information on the number of times that a monophyletic clade is separated from all other nodes in each group of 100 trees. This is similar to the bootstrapping method. The majority consensus tree is that which contains all clades which occur as monophyletic groups as often as possible in the data and at least more than 50% of the time.

Output files from CONSENSE were entered into MacClade version 3.05 *(Maddison and Maddison 1992).* Parsimony trees were drawn noting the tree length, branching patterns and branch lengths.

The final maximum parsimony networks were drawn in MacDraw Pro™ using the data from the MacClade output. Additionally the frequency of the occurrence of each haplotype was recorded and the nodes were depicted on the network accordingly - the most frequent haplotype having the largest area.

AMOVA

Analysis of Molecular Variance (AMOVA) was calculated using the program ARLEQUIN *(Schneider et al. 1997*). AMOVA is similar to an analysis of variance and includes molecular haplotype distance information. The frequency of each haplotype and the composition of the haplotype are used to estimate genetic structure *(Excoffier et al. 1992).* A distance matrix of squared Euclidean distance (the number of mutational steps between pairs of haplotypes) was calculated in Arlequin from haplotype frequency lists for each population. From this distance matrix variance components were calculated - the total molecular variance (σ^2) is the sum of the variances due to differences between haplotypes within a population (σ^2_c), the sum of the variances due to differences among haplotypes in different populations within a group (σ^2_b), and those due to differences among the populations (σ^2_a).

Two tests of genetic structure were performed with the following population groupings.

1. a Ulster, Munster, Leinster, Connaught
   b Africa
   c Turkey

2. a Ulster, Munster, Leinster, Connaught
   b Africa
   c Turkey
   d Denmark

Mismatch distributions

All pairwise comparisons between haplotypes for each population were computed for D_{asm} and D_{swasm} in the program SWAMP. All classes of the number of differences between haplotypes were counted from the matrices in an Excel™ spreadsheet. Histograms of the mismatches were created in Excel™ and manipulated in MacDraw Pro™ for final presentation.

Coalescence dating

The average squared distance (ASD) measure *(Goldstein et al. 1995)* was calculated using the program Microsat *(Erich Minch)*.

2.3 Results

2.3.1. Assessment of polymorphism and final assembly of microsatellite loci

All of the microsatellites detected on cosmid hsu177e8 were tested for polymorphism in a small subset (n = 90) of the Irish samples. Three loci were found to be monomorphic. The monomorphic loci had a lower percentage of sequence conservation than did the polymorphic loci. The sequence conservation for the monomorphic loci ranged from 81% - 94%, and for the polymorphic loci the range was between 85% - 100%. All of the polymorphic dinucleotide repeats had 100% sequence conservation. In total 6 microsatellites were found to be polymorphic in the initial screening process.

Locus F29 exhibited a large number of alleles ranging from 192 bp to 224 bp (17 alleles). Despite the large genetic variability at this locus the primer set was deemed unsuitable as allele calling proved to be unrepeatable and inconclusive. Therefore of the 9 microsatellite repeat loci on cosmid hsu177e8 five were selected for the attainment of X chromosomal haplotypes, four CA repeats (F3, F7, F13, and F29) and the single CAT repeat (F14).

2.3.2. Basic allelic data

The results described below refer, in the most part, to the Togolese, Turkish and Irish populations. For the purpose of some analyses the Irish population has been split into four subpopulations representing Ulster, Munster, Leinster and Connaught. For comparison, the Danish and Greek data has been added in some cases to provide a richer context for the major populations.

2.3.3. Allele frequencies and frequency histograms

Allele frequencies were determined at each locus by counting. The frequencies were entered into an Excel™ spreadsheet where frequency histograms for each population at each locus were produced. **Figures 2.4 a and 2.4 b** show these histograms.

Allele frequencies at each of the five loci in each of the population groups are shown in **table 2.6**. The total number of alleles detected at the five loci was 43, the mean number of alleles observed therefore being 8.6, with a range from 4 repeat variants at the triplet repeat locus F14 to 14 repeat variants locus F7. The most frequent allele detected at each locus was common between all populations except for the African Togolese population which demonstrated different MFAs at two of the loci.

With the exception of F7 and F13 each locus shows a unimodal distribution with one frequent allele and less frequent alleles differing by a series of single repeat units. At locus F7 a smaller allele of size 152 bp is observed in both Ireland (3.5%) and the Turkish sample (7%) but is absent in the Togolese sample. This allele is seven repeat units smaller than the next allele 166 bp, and no intermediate alleles are found in either the Irish or the Turkish samples. However in the Togolese population an intermediate allele of size 156 bp is observed (2.7%) which is absent in the European populations. A rare allele (105 bp) at locus F13 is observed in both the Irish (0.96%) and Turkish (2.3%) samples but not in Togo. Again, no intermediate alleles are observed between 105 bp and 111 bp.

Locus	Allele size (bp)	Togo n = 37	Denmark n = 14	Greece n = 7	Turkey n = 43	Leinster n = 122	Munster n = 99	Connaught n = 30	Ulster n = 63
	160	0.135	0	0	0	0	0	0	0
	162	0.378	0.071	0	0.047	0.131	0.061	0.067	0.063
F3	164	0.324	0.643	1.000	0.814	0.852	0.848	0.800	0.810
	166	0.135	0.286	0	0.116	0.016	0.091	0.133	0.127
	168	0.027	0	0	0.023	0	0	0	0
	152	0	0	0	0.070	0.025	0.020	0.100	0.048
	154	0	0	0	0	0	0	0	0
	156	0.027	0	0	0	0	0	0	0
	158	0	0	0	0	0	0	0	0
	160	0	0	0	0	0	0	0	0
	162	0.081	0	0	0	0	0	0	0
F7	164	0.216	0	0	0	0	0	0	0
	166	0.027	0	0	0.023	0.016	0.020	0	0
	168	0.135	0.071	0.286	0.047	0.033	0.081	0.133	0.032
	170	0.052	0.071	0.143	0.116	0.049	0.061	0.067	0.063
	172	0.324	0.571	0.429	0.535	0.705	0.667	0.600	0.540
	174	0.108	0.286	0.143	0.163	0.156	0.091	0.067	0.222
	176	0.027	0	0	0.023	0.008	0.061	0.033	0.079
	178	0	0	0	0.023	0.008	0	0	0.016
	105	0	0	0	0.023	0.008	0.010	0.033	0
	107	0	0	0	0	0	0	0	0
	109	0	0	0	0	0	0	0	0
	111	0.541	0	0	0.023	0.008	0.020	0	0
F13	113	0.081	0	0	0	0.016	0.030	0	0
	115	0.297	0.929	0.857	0.860	0.885	0.889	0.933	0.937
	117	0	0.071	0.143	0.070	0.066	0.030	0.033	0.048
	119	0.081	0	0	0.023	0.008	0.010	0	0.016
	121	0	0	0	0	0	0.010	0	0
	123	0	0	0	0	0.008	0	0	0
	128	0	0	0	0	0	0	0	0.032
F14	131	0.946	0.643	0.714	0.628	0.574	0.596	0.400	0.508
	134	0.054	0.357	0.286	0.349	0.410	0.404	0.600	0.460
	137	0	0	0	0.023	0.016	0	0	0
	148	0	0	0	0.023	0	0	0	0
	150	0.027	0	0.143	0.047	0	0.020	0.067	0.079
	152	0	0	0	0.070	0.066	0.040	0.067	0.048
	154	0.486	0.571	0.857	0.372	0.607	0.475	0.467	0.508
F29	156	0.189	0.286	0	0.349	0.27	0.394	0.367	0.302
	158	0	0.071	0	0.116	0.057	0.051	0	0.048
	160	0.081	0	0	0	0	0.010	0.033	0.016
	162	0.081	0.071	0	0.023	0	0	0	0
	164	0.108	0	0	0	0	0	0	0
	166	0.027	0	0	0	0	0	0	0

Table 2.6: Allele frequencies of X chromosome microsatellite loci in all populations. In most cases the most frequent allele was shared between all populations. The most frequent allele in the African population differed in some cases to the European populations. Unusually smaller allelic variants were detected at the F7 and F13 loci.

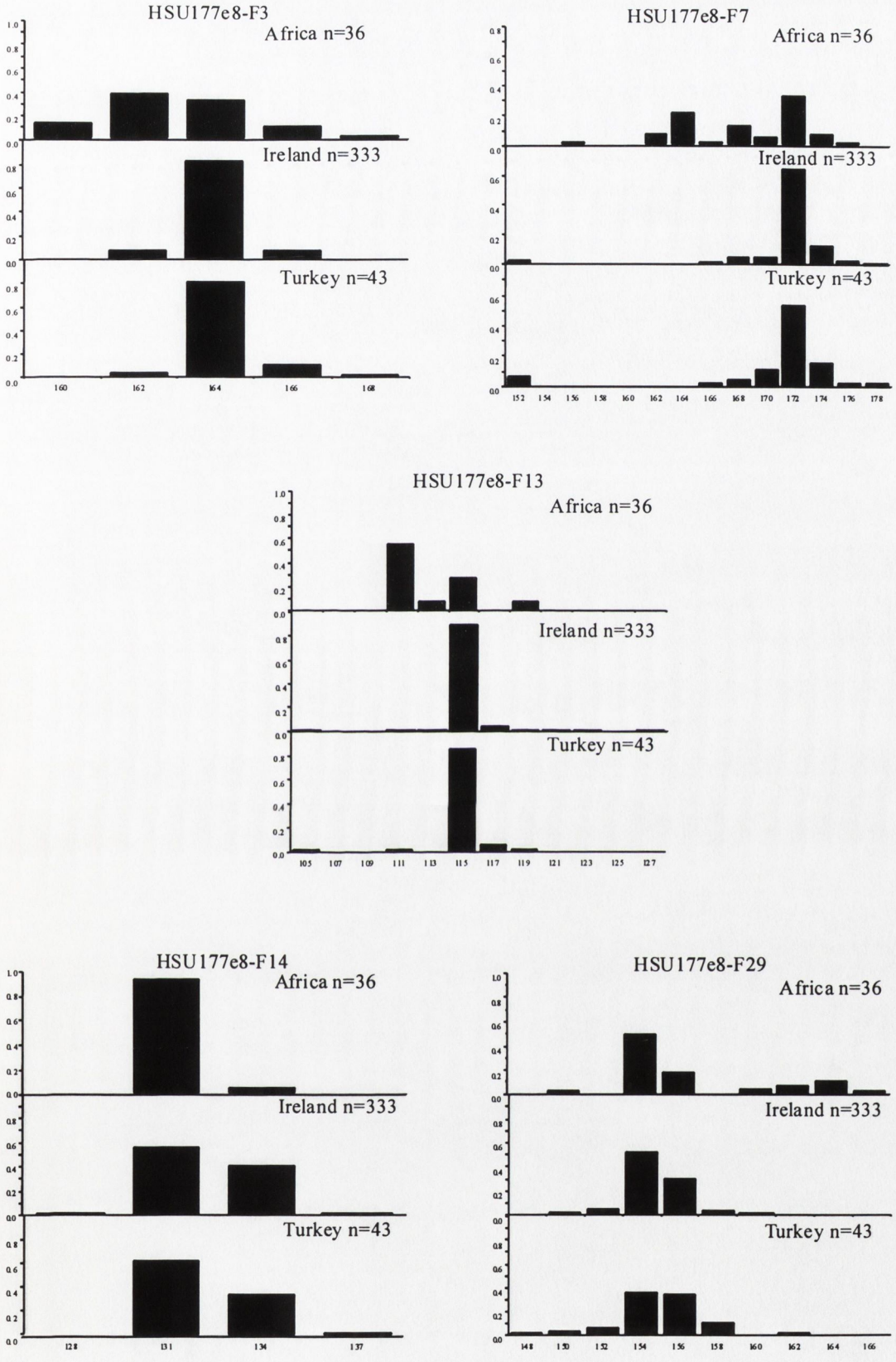


Figure 2.4 a: Allele frequency histograms for 5 loci in 3 populations. The most frequent alleles are shared between all populations except at the F3 and F13 loci.

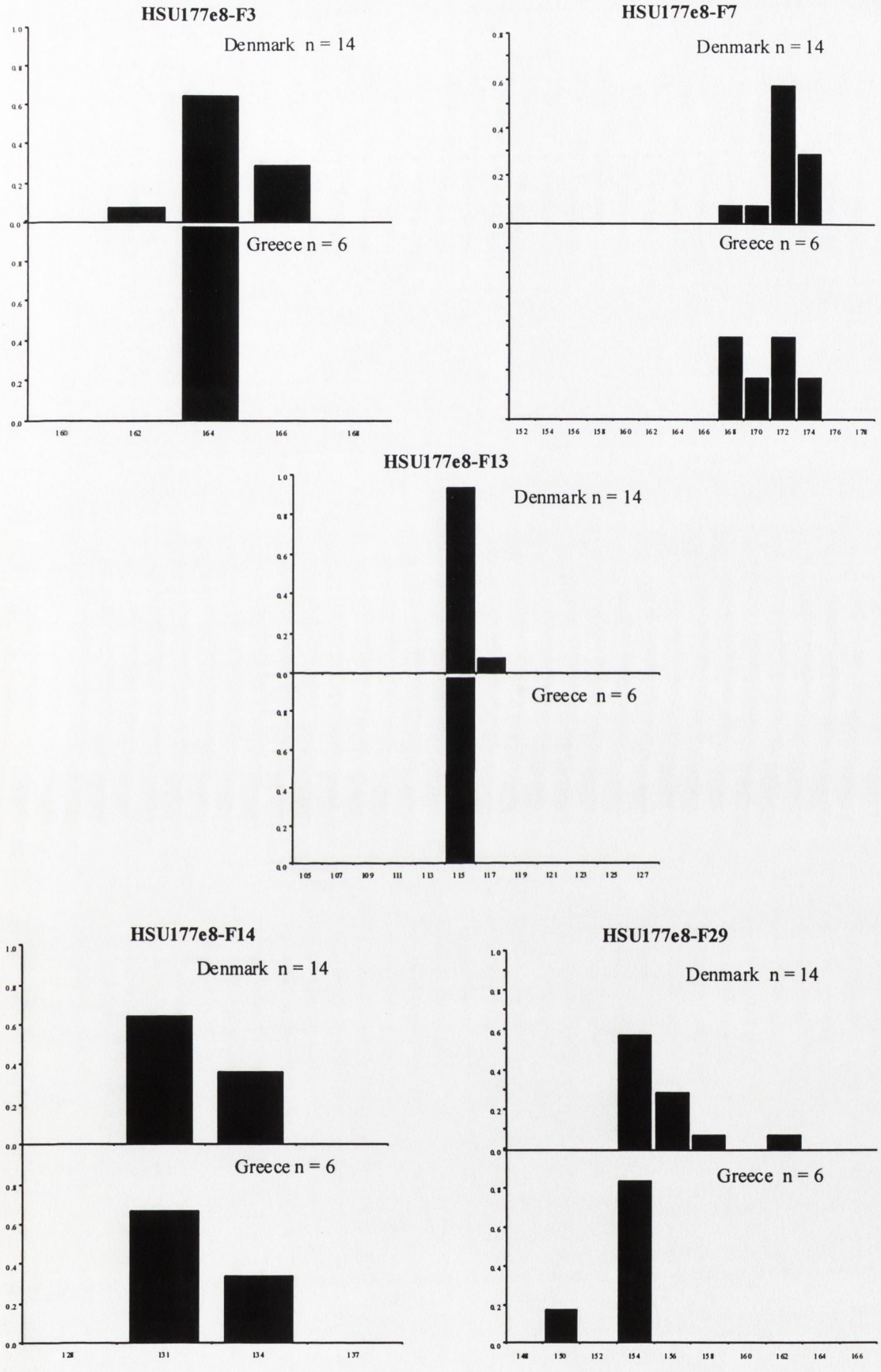


Figure 2.4 b: Allele frequency histograms for 5 X chromosome microsatellite loci in Denmark and Greece. The most frequent alleles are shared between both populations.

2.3.4. Microsatellite allele length variance

Allele length variance is an indication of genetic diversity and is proportional to coalescence time. The allele length variance at each locus in each sample is given in **table 2.7.** Not surprisingly, considering the large number of alleles at the locus, the variance at locus F7 is the greatest in each population. Although the Togolese population displays the greatest overall variance, at locus F7 the Turkish population has the highest variance which is presumably the result of the presence of the smaller 152 bp allele which is absent in the Togolese population. Surprisingly the variance at the triplet repeat locus, F14, is not substantially less than at the other loci despite it containing the least number of repeat variants. In the Irish population the variance at F14 is comparable to that at F13 and F29 and greater than F3. In the Turkish population the variance is not too far removed from the variance at F13 and F29 and is greater than F3. In the African population, however, the variance at F14 is substantially less than all other loci. The variance of the repeat units in the African population averaged across all loci is 2.433, in the Turkish population 2.106 and in Ireland 1.259.

Ireland						
	177e8-F3	**177e8-F7**	**177e8-F13**	**177e8-F14**	**177e8-F29**	**Mean variance**
Raw variance	0.6506	17.3091	2.4173	2.4207	3.7297	5.3055 nucleotides
Repeat variance	0.1627 dinucleotide	4.3273 dinucleotide	0.6043 dinucleotide	0.2690 trinucleotide	0.9324 dinucleotide	1.2591 repeat units
Turkey						
	177e8-F3	**177e8-F7**	**177e8-F13**	**177e8-F14**	**177e8-F29**	**Mean variance**
Raw variance	0.9922	30.5781	3.4197	2.6312	5.9623	8.7167 nucleotides
Repeat variance	0.2481 dinucleotide	7.6445 dinucleotide	0.8549 dinucleotide	0.2924 trinucleotide	1.4906 dinucleotide	2.1061 repeat units
Togo						
	177e8-F3	**177e8-F7**	**177e8-F13**	**177e8-F14**	**177e8-F29**	**Mean variance**
Raw variance	3.7714	21.4444	6.5111	0.4857	16.7206	9.7867 nucleotides
Repeat variance	0.9429 dinucleotide	5.3611 dinucleotide	1.6278 dinucleotide	0.054 trinucleotide	4.1802 dinucleotide	2.4332 repeat units

Table 2.7: Allele length variance statistics for 5 microsatellite loci in 3 populations. The raw variance is calculated as the total variance at each locus, the repeat variance is calculated the variance of the repeat lengths (i.e. for trinucleotides, repeat variance = raw variance/3^2 and for tetranucleotides, repeat variance = raw variance/4^2)

2.3.5. Linkage disequilibrium

Population	locus 1	locus 2	P value	S.E.
Togo	F7	F13	**0.0022**	-6.1013
Togo	F7	F29	0.0690	-2.6731
Togo	F13	F29	0.4491	-0.8006
Togo	F7	F14	0.4184	-0.8713
Togo	F13	F14	0.3733	-0.9853
Togo	F29	F14	0.7543	-0.2819
Togo	F7	F3	**0.0006**	-7.4186
Togo	F13	F3	**0.0003**	-8.1807
Togo	F29	F3	**0.0161**	-4.1265
Togo	F14	F3	0.2879	-1.2452

Population	locus 1	locus 2	P value	S.E.
Ireland	F7	F13	**0.0001**	-9.2103
Ireland	F7	F29	0.2414	-1.4213
Ireland	F13	F29	0.0649	-2.7355
Ireland	F7	F14	**0.0001**	-9.2103
Ireland	F13	F14	0.1861	-1.6814
Ireland	F29	F14	**0.0001**	-9.2103
Ireland	F7	F3	**0.0001**	-9.2103
Ireland	F13	F3	**0.0004**	-7.8753
Ireland	F29	F3	0.4416	-0.8174
Ireland	F14	F3	**0.0001**	-9.2103

Population	locus 1	locus 2	P value	S.E.
Turkey	F7	F13	0.1888	-1.6669
Turkey	F7	F29	**0.0015**	-6.5293
Turkey	F13	F29	1.0000	0.0000
Turkey	F7	F14	**0.0109**	-4.5227
Turkey	F13	F14	0.5019	-0.6894
Turkey	F29	F14	**0.0021**	-6.1658
Turkey	F7	F3	**0.0412**	-3.1883
Turkey	F13	F3	**0.0470**	-3.0576
Turkey	F29	F3	0.2124	-1.5495
Turkey	F14	F3	**0.0240**	-3.7280

Table 2.8: Linkage disequilibrium summarised across populations for all locus x locus comparisons using Fisher's combined probability test. Bold type indicates significant linkage disequilibrium ($p<0.05$) between two loci.

Locus x locus comparisons were calculated for all possible combinations of loci (10) in each of three populations (Togo, Turkey and Ireland). A summary of the results is given in **table 2.8.** In the three populations allelic associations ($P < 0.05$) were detected

for 53.3% of comparisons. Fisher's combined probability was <0.0001 for all locus x locus comparisons in the 3 populations. Each of the five loci found on the hsu177e8 cosmid on the X chromosome located within 40 kb displayed significant linkage disequilibrium with 2 or more of the other loci in at least one of the populations. Significant linkage was not detected between F13 and either of F14 or F29, nor between F14 and F13 or between F29 and F13.

It is possible that recombination between loci might contribute to the lack of detectable linkage disequilibrium between some pairs of loci in some of the populations. However, by considering the physical distance between the loci on the cosmid (see **table 2.2**), there is little evidence to suggest a large influence of recombination on this region of the chromosome. For example, loci F3 and F7 are separated by ~1300 bp and significant linkage is detected between these loci in all of the populations. The largest physical distance between any two adjacent loci is between F7 and F13 (~ 19,700 bp) and in both the Irish and Togolese populations significant linkage is detected between this pair. Furthermore, F29 is the most distal on the cosmid from F3 (~38,900 bp) and F7 (17,600 bp) and in both the Togolese and Turkish samples significant linkage disequilibrium is detected between F29 and F7.

2.3.6. Haplotypes: frequencies and measures of diversity

A total of 110 haplotypes were observed in the 3 populations with the 5 loci. Only 17.3% of these were shared between populations, but the majority (82.7%) were unique to each population. All detected haplotypes and their frequencies in each population are given in **table 2.10.**

2.3.7. Haplotype distributions between populations

In the Irish population 80 haplotypes were detected in 333 individuals. The most frequent haplotype in the Irish population was ht 57 (14.41%) which was also the most frequently sampled haplotype in the Turkish population (23.26%) but was found at a low frequency in the Togolese population (2.78%). Haplotype 63 was also a common haplotype in the Irish population (13.81%) and has diverged by two mutational steps from ht 57, differing by one repeat unit each at locus F14 and F29.

	Togo	Ireland
Ireland	4.0%	
Turkey	4.1%	17.8%

Table 2.9: The proportion of haplotypes shared between populations. More haplotypes were shared between the Irish and Turkish populations than either with the Togolese population.

In the Turkish population 26 haplotypes were detected in 36 individuals. The most frequent haplotype was ht 57 (23.26%), also found at high frequency in the Irish population. Similarly, the second most common Irish haplotype, ht 63, was common in the Turkish population (6.98%). Haplotype 26, was found as frequently in the Turkish population as ht 63, however, ht 26 was not common in the Irish population (1.2%). Haplotype 26 is associated with the rare 152 bp allele at locus F7, which is more frequent in Turkey than in Ireland. The Irish and Turkish populations shared 16 of 90 (17.8%) possible haplotypes.

Fewer haplotypes were shared between the Togolese population and the Irish and Turkish populations. In the Togolese population a total of 25 haplotypes were detected in 43 individuals. The most common Togolese haplotype, ht 9 (16.67%) was not sampled in either the Irish or the Turkish populations. Eight individual mutations occurring at loci F3, F7, F13 and F29, separate ht 9 and ht 57, the most common Irish and Turkish haplotype. In ht 9 and ht 57, however, allele 131 bp at F14 is common in all populations. Haplotype 70 was also detected at high frequency in the Togolese population (11.11%) but was not sampled in the Irish or the Turkish populations. Interestingly 11 mutational steps separate ht 70 from ht 9, but only 7 mutational steps separate ht 70 from ht 57. Two haplotypes of a possible 49 (4.1%) were shared between the Turkish and Togolese populations. Four haplotypes of a possible 101 (4.0%) were shared between the Irish and Togolese populations.

Ht	177E8-F3	177E8-F7	177E8-F13	177E8-F14	177E8-F29	Ireland (n=333)	Turkey (n=43)	Africa (n=36)
1	160	166	113	131	154			2.78
2	160	168	111	131	162			2.78
3	160	172	111	131	154			2.78
4	160	174	111	131	160			2.78
5	160	174	111	131	162			2.78
6	162	152	115	134	152	0.30		
7	162	156	115	131	150			2.78
8	162	162	111	131	154			2.78
9	162	164	111	131	154			**16.67**
10	162	164	111	131	156			2.78
11	162	164	111	134	154			2.78
12	162	166	105	131	156	0.60	2.33	
13	162	166	123	131	152	0.30		
14	162	166	123	137	156	0.30		
15	162	166	127	131	152	0.30		
16	162	168	111	131	154			5.56
17	162	170	115	134	148		2.33	
18	162	170	115	134	154	1.50		
19	162	172	115	131	150	0.30		
20	162	172	115	131	154	0.30		2.78
21	162	172	115	131	156			2.78
22	162	172	115	134	154	3.60		
23	162	172	115	134	156	0.60		
24	164	152	115	131	156	0.30		
25	164	152	115	134	150	0.30		
26	164	152	115	134	152	1.20	**6.98**	
27	164	152	115	134	154	0.60		
28	164	152	115	134	156	0.60		
29	164	152	117	131	154	0.30		
30	164	162	115	131	154			2.78
31	164	166	115	131	154	0.30		
32	164	168	115	131	154	1.20		
33	164	168	115	131	156	0.60		
34	164	168	115	131	158	0.30		
35	164	168	115	134	154	1.50	4.65	
36	164	168	115	134	156	0.90		
37	164	168	115	137	154	0.30		
38	164	170	111	131	156	0.30		
39	164	170	111	131	162	0.30		
40	164	170	115	131	154	0.60	2.33	
41	164	170	115	131	156	0.90		
42	164	170	115	131	158	0.30	4.65	
43	164	170	115	134	154	1.20		
44	164	170	115	134	156	0.30		
45	164	170	115	134	158	0.30		
46	164	170	119	131	156			2.78
47	164	170	121	134	154	0.30		
48	164	172	111	131	166	0.30		2.78
49	164	172	111	134	154	0.30		
50	164	172	113	131	152	0.30		
51	164	172	113	131	154	0.30		
52	164	172	113	131	156	0.30		
53	164	172	115	128	154	0.30		
54	164	172	115	128	156	0.30		
55	164	172	115	131	150	1.50	2.33	

Table 2.10: X STR haplotypes denoted by allele length and their frequencies in each population. Bold text denotes the more common haplotypes in each population.

Ht	177E8-F3	177E8-F7	177E8-F13	177E8-F14	177E8-F29	Ireland (n=333)	Turkey (n=43)	Africa (n=36)
56	164	172	115	131	154	**9.91**	2.33	5.56
57	164	172	115	131	156	**14.41**	**23.26**	2.78
58	164	172	115	131	158	3.60	4.65	
59	164	172	115	131	160	0.90		2.78
60	164	172	115	131	166	0.30		
61	164	172	115	134	150	0.30		
62	164	172	115	134	152	0.90		
63	164	172	115	134	154	**13.81**	**6.98**	
64	164	172	117	131	154	0.60		
65	164	172	117	131	156	2.10	2.33	
66	164	172	117	134	154	1.20		
67	164	172	119	131	154	0.30		
68	164	172	119	131	156	0.30		
69	164	172	119	134	156		2.33	
70	164	174	111	131	164			**11.11**
71	164	174	113	131	154	0.30		
72	164	174	115	131	150	0.60	2.33	
73	164	174	115	131	152	2.40		
74	164	174	115	131	154	5.41	2.33	
75	164	174	115	131	156	3.30	4.65	
76	164	174	115	131	158	0.30		
77	164	174	115	131	162		2.33	
78	164	174	115	134	154	2.10		
79	164	174	115	134	156	2.10		
80	164	174	115	134	158	0.30		
81	164	174	117	131	154		2.33	
82	164	174	117	134	154		2.33	
83	164	174	117	134	156	0.30		
84	164	176	115	131	154	0.30		
85	164	176	115	131	162			2.78
86	164	176	115	134	154	0.60		
87	164	178	115	131	154		2.33	
88	164	178	115	131	156	0.30		
89	164	178	115	134	154	0.30		
90	166	162	113	134	156			2.78
91	166	168	105	131	156	0.30		
92	166	168	115	134	154	0.30		
93	166	168	119	131	154			2.78
94	166	170	115	131	156			2.78
95	166	172	113	131	154			2.78
96	166	172	115	131	154	0.60	2.33	
97	166	172	115	131	156	0.30		
98	166	172	115	134	152	0.60		
99	166	172	115	134	154	1.80	4.65	
100	166	172	115	134	158		2.33	
101	166	172	117	131	156	0.30		
102	166	172	119	134	154	0.30		
103	166	174	115	131	160	0.30		
104	166	176	115	134	154	2.10		
105	166	176	115	134	156	0.60		
106	166	176	115	137	154		2.33	
107	166	178	115	134	154	0.30		
108	166	178	115	134	156	0.30		
109	168	168	119	131	156			2.78
110	168	170	111	134	154		2.33	

Table 2.10 *cont.*: X STR haplotypes, denoted by allele length and their frequencies in each population. Bold text denotes the more common haplotypes in each population.

2.3.8. Haplotypic diversity measures

Gene diversity is equivalent to the expected heterozygosity in diploid systems *(Nei 1987)* and is given in **table 2.11** for each population. It is defined as the probability that two randomly chosen haplotypes are different in a population and is estimated by the following equation:

$$H = n / n\text{-}1 \ (1 - \Sigma p^2)$$

Haplotypic diversity measures illustrate a greater variability in the Togolese population compared to the European populations.

	Average gene diversity (H)	Standard Error
Africa	0.602	0.357
Turkey	0.500	0.305
Ulster	0.459	0.282
Munster	0.425	0.264
Leinster	0.402	0.253
Connaught	0.452	0.284

Table 2.11: Average gene diversity *(Nei 1987)* for five-locus X STR haplotypes, calculated in the Togolese, Turkish and Irish populations

2.3.9. Genetic distance measures: populations as OTUs

Genetic distance, originally defined by Nei *(1972)* for protein-electrophoretic data, is a quantitative estimate of how divergent two individuals are genetically. Recently a number of genetic distance measures have been developed for use with microsatellite data which account for the stepwise mutational mechanism of these loci *(Goldstein et al. 1995a;1995b; Shriver at al 1995; Jin et al. 1996; Slatkin 1995).*

In order for a distance measure to be useful in the estimation of relative times of divergence a linear increase with respect to time is essential and a low variance is beneficial. The combination of these two parameters determines the performance of a genetic distance measure *(Pollock and Goldstein 1995).*

The D_{dm}, delta mu squared, genetic distance measures the squared mean difference in the number of repeats between alleles of two populations and its average has a linear

expectation over time *(Goldstein et al. 1995).* The average squared distance between two populations at a given locus is

$$D_{dm} = (\mu (A) - \mu (B))^2$$

where μ (A) is the mean allele size for population A and similarly μ (B) is the mean allele size for population B.

The D_{dm} genetic distance measure is linear with time and is also independent of population size allowing direct estimates of divergence times between populations without the addition of further parameter estimates to account for different sample sizes *(Takezaki and Nei 1996).*

2.3.10. Tree drawing

Phylogenies impart two important sources of information about individual and population relationships. First, the branching pattern or ordering of nodes suggests the nature of the relationships according to the extent of clustering of OTUs. Second, branch lengths are suggestive of the time passed since an evolutionary split either between species, population groups or individuals provided that reversions of mutations to previous character states are rare.

2.3.11. Phylogenetic trees

The resulting tree **(figure 2.5)** demonstrated two apparent features, the clear separation of Togolese and all non-African populations by the longest branch, and the clustering of all non-African populations with no obvious pattern. That the deepest split between all populations is that between Africans and non-Africans has been recorded in many previous investigations of human population relatedness *(Cavalli-Sforza 1996)*

2.3.12. Phylogenetic analysis: populations as OTUs

Ddm	TOGO	DENMARK	GREECE	TURKEY	LEINSTER	MUNSTER	CONNAUGH	ULSTER
TOGO		*0.519*	*0.445*	*0.166*	*0.353*	*0.317*	*0.186*	*0.361*
DENMARK	1.089		*0.171*	*0.099*	*0.025*	*0.018*	*0.268*	*0.038*
GREECE	1.143	0.288		*0.095*	*0.062*	*0.111*	*0.064*	*0.048*
TURKEY	0.669	0.139	0.118		*0.039*	*0.031*	*0.041*	*0.045*
LEINSTER	0.966	0.070	0.107	0.063		*0.005*	*0.168*	*0.006*
MUNSTER	0.845	0.042	0.149	0.039	0.013		*0.146*	*0.012*
CONNAUGHT	0.576	0.389	0.175	0.067	0.215	0.194		*0.183*
ULSTER	1.071	0.062	0.091	0.067	0.008	0.019	0.227	

Table 2.12: Ddm genetic distances (lower diagonal) and their standard errors (upper diagonal)

The greatest Ddm genetic distances were detected between the African population and all other populations (average = 0.908). The Connaught population was found to have the smallest genetic distance with the African population. Indeterminate distances were detected between the European populations.

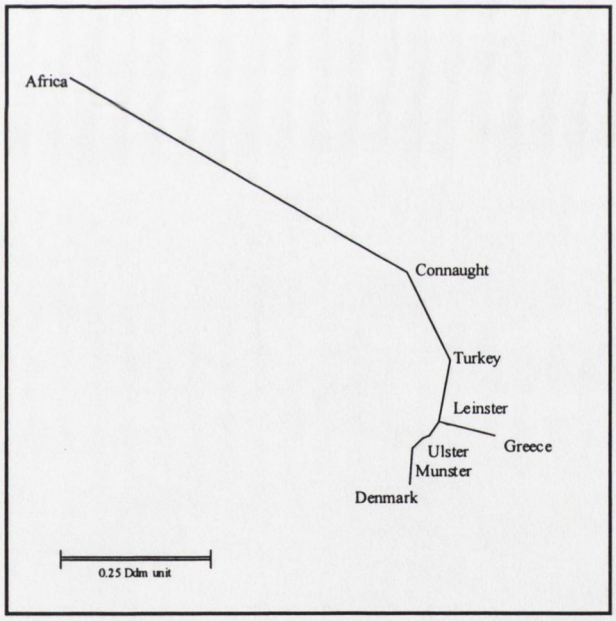


Figure 2.5: Neighbour-joining tree relating Ddm genetic distances. The greatest genetic distance is that between the African population and all other populations.

In order to confirm the deep rooted split between African and non-African populations and to further assess the extent of clustering of non-African populations an analysis of population structure was performed. One approach to the molecular analysis of

geographic population structure is the use of F-statistics *(Wright 1951).* F-statistics are generally used to estimate gene flow between diverse geographic populations assuming selective neutrality *(Weir and Cockerham 1984).* The relative magnitude of genetic differentiation between populations is measured by calculating the coefficient of gene differentiation from allele frequencies. Wright's fixation index Fst may be defined as

$$Fst = (\sigma^2_a + \sigma^2_b) / \sigma^2_T$$

and describes the ratio between the diversity found between subpopulations and the total genetic diversity in all populations. A high Fst index indicates population structure whereas a low Fst index indicates little differentiation.

Fst	Ulster	Munster	Leinster	Connaught	Denmark	Turkey	Togo	Asia	England
Munster	0.000								
Leinster	0.000	0.000							
Connaught	0.034	0.046	0.062						
Denmark	0.000	0.000	0.002	0.045					
Turkey	0.000	0.000	0.010	0.000	0.000				
Togo	0.274	0.295	0.373	0.083	0.231	0.123			
Asia	0.057	0.129	0.141	0.012	0.059	0.007	0.171		
England	0.068	0.159	0.171	0.000	0.184	0.000	0.075	0.000	
Greece	0.000	0.007	0.000	0.000	0.185	0.000	0.183	0.000	0.000

Table 2.13: Fst genetic distances calculated for 6 geographically diverse populations and four Irish subpopulations.

The Fst fixation index is a simple measure of the amount of genetic differentiation existing between populations relative to a hypothetical group of populations. The greatest Fst measure was that detected between the African and Leinster populations (0.373). This suggests a great differentiation between these two populations. Similarly, high Fst measures were detected between the African population and the Ulster and Munster populations, but, between the African population and Connaught the Fst measure was relatively low (0.083) indicating little differentiation. Within the Irish subpopulations no differentiation was detected except between Connaught and the Leinster population (0.062).

2.3.13. Population structure: Analysis of Molecular Variance

In an Analysis of MOlecular VAriance (AMOVA) a comparison of the three major populations ([Ulster, Munster, Leinster, Connaught] V [Africa] V [Turkey]) calculated that 87.09% of genetic diversity is distributed as differences between individuals within populations and only 12.3% of the total genetic variance accounts for the differences between populations. For four groups ([Ulster, Munster, Leinster, Connaught] V [Africa] V [Turkey] V [Denmark]) the within population variance was 88.49% and the fraction of variance detected among the groups was 10.86%. The proportion of variance attributed to variance among populations was highly significant with a p value of <0.0001. The fraction of variance detected among populations within groups was significant at the 10% level and the fraction of variance detected among groups was significant at the 25% level.

Population groups	Ulster; Munster; Leinster; Connaught Africa Denmark Turkey	
Source of variation	**% variation**	**p value =**
Among groups	10.86	0.234
Among populations withing groups	0.65	0.096
Within populations	88.49	0.000

Table 2.14: AMOVA test of population substructure between four groups ([Ulster, Munster, Leinster, Connaught]V [Africa] V [Turkey] V [Denmark]).

2.3.14. Population structure: Multidimensional Scaling

Graphical representations of the association among genotypes were created using multidimensional scaling (MDS) in the program SPSS version 6.1.1. Eigenvalues were obtained from a matrix of pairwise Fst distances **(table 2.13)** between populations. **Figure 2.6** shows the MDS plot for these distances.

In the MDS plot the separation of the African population from all other populations is most evident. Although there is little further structure between non-African populations it is clear that 3 of the 4 Irish subpopulations (Ulster, Munster and Leinster) cluster together in the opposite plot quadrant from the African population. It was clear that molecular analysis utilising conventional phylogenetics could not seperate the non-African populations at this level. MDS, however, tentatively suggests a greater differentiation between Togolese and Irish populations.

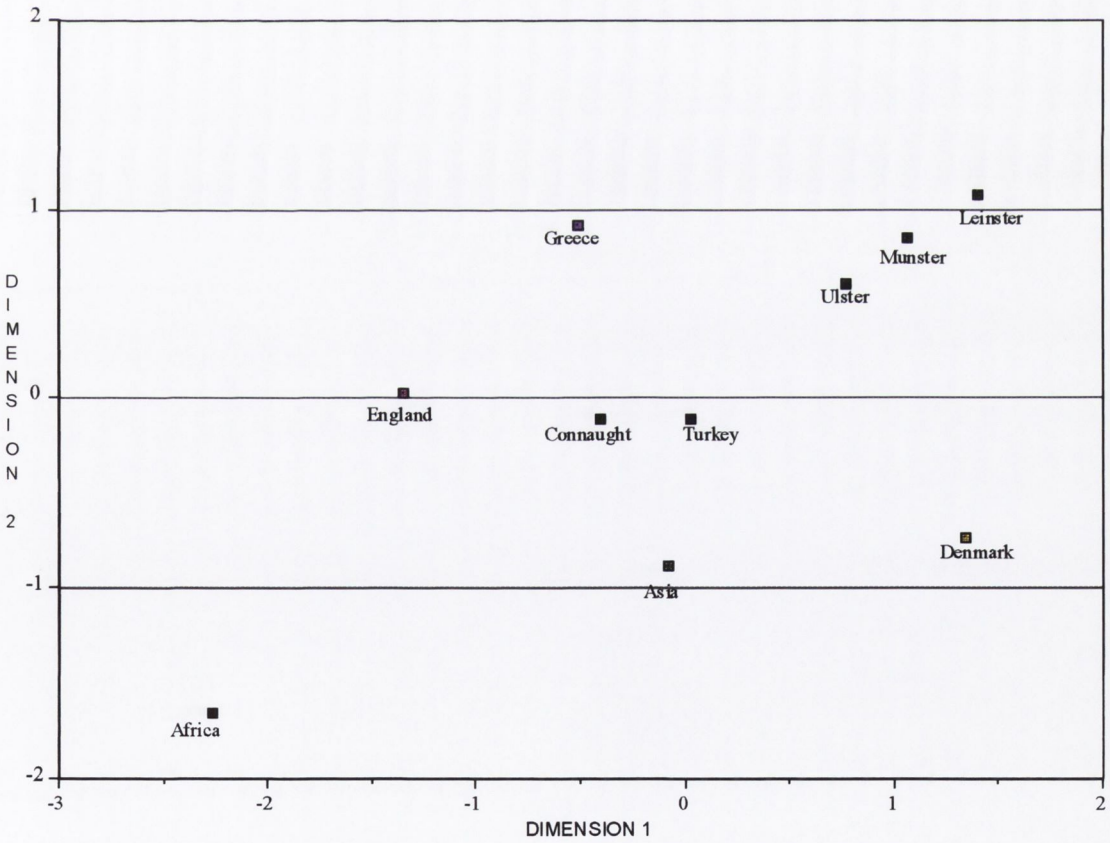


Figure 2.6: MDS plot from computed pairwise Fst values using X chromosome microsatellite haplotypes

2.3.15. Maximum Parsimony Phylogenies

A most parsimonious tree is one that requires the smallest number of evolutionary changes to explain the observed differences among OTUs. A number of computer algorithms exist which search alternative trees for those of minimum total branch length. Many trees of different topology are often found to be equally parsimonious.

Maximum parsimony networks relating haplotypes within populations were constructed for all 4 Irish subpopulations and also for the Togolese, Turkish and Danish populations.

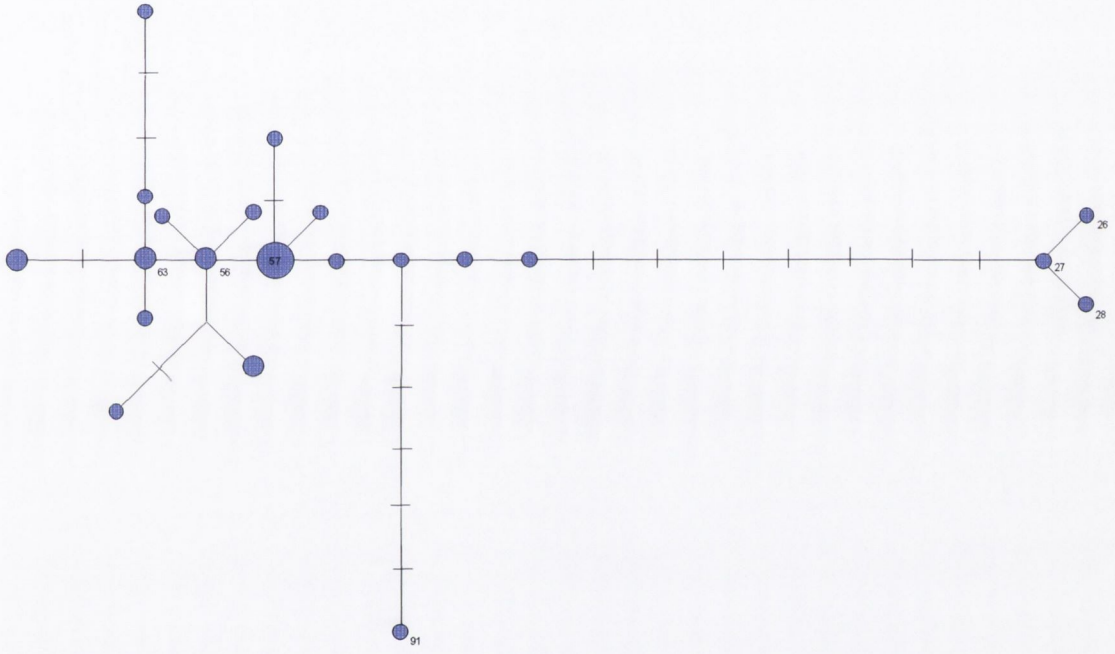


Figure 2.7: Maximum parsimony network relating haplotypes sampled in the Connaught population (n = 30). The area of the nodes is proportional to the number of haplotypes sampled. Each division represents the gain or loss of a single repeat unit (one mutational step). The most frequent haplotype sampled was ht 57, also common in the Ulster, Munster and Leinster populations. The three haplotypes associated with the 152 bp allele at the F7 locus are peripheral on the network and are found on long branches distal from the main core of common haplotypes. Similarly, ht 91, associated with the uncommon 105 bp allele at locus F13, is found on an unusually long branch.

Haplotype (ht) 57, ht 56 and ht 63 are the most common haplotypes in all four Irish provincial populations and expectedly are comprised of the most common alleles at each locus detected in the Irish population. Similarly for the Turkish and the Danish populations haplotype 57 is the most common. In the Togolese population the most common

haplotype is ht 9 and is comprised of the most common alleles at each locus found in the African sample.

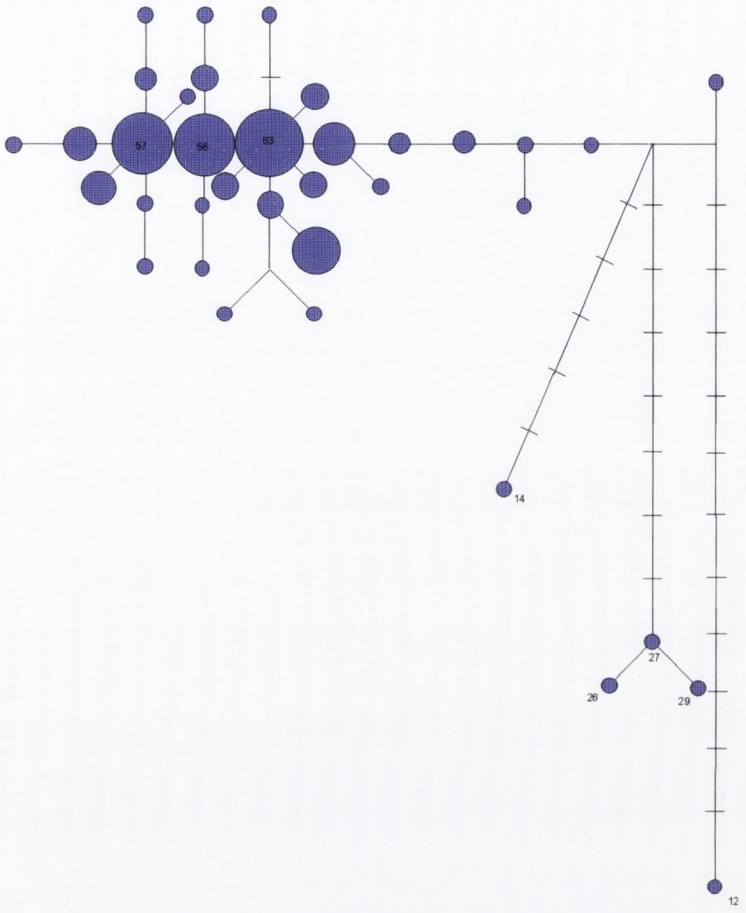


Figure 2.8: Maximum parsimony network relating haplotypes sampled in the Leinster population (n = 122). The most common haplotype was ht 63, common also in the Ulster and Munster populations and present in the Connaught population. Haplotypes associated with the rare alleles, 152 bp and 105 bp at the F7 and F13 loci, are found on distal branches from the core of the network. The common haplotypes cluster closely at the core of the network with numerous, closely related haplotypes radiating from the more common nodes.

In all four provincial Irish networks there is a main core of haplotypes, with the most common haplotype central in the core with less common haplotypes radiating from it. Additionally, a number of haplotypes are found at the end of long branches distal on the network from the core. These haplotypes are associated with alleles 152 bp and 105 bp at loci F7 and F13 and also with allele 123 bp at locus F13. For example, in Ulster, the three extreme nodes represent ht 6, ht 26 and ht 28, all of which contain the 152 bp allele at F7.

In Leinster the most distal haplotype, ht 12 is associated with the 105 bp allele at F13 and is 19 mutational steps from the most common ht 57. Also in Leinster three

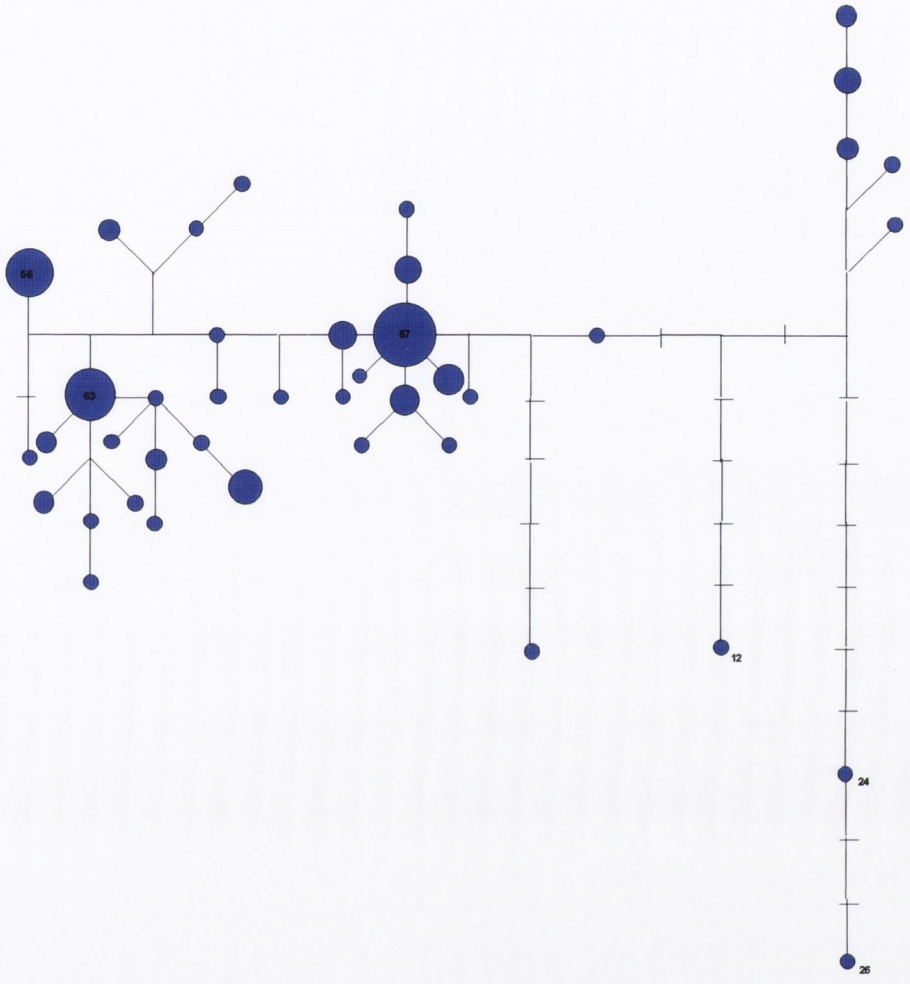


Figure 2.9: Maximum parsimony network relating the haplotypes sampled in the Munster population (n = 99). The common ht 57 lies central on the network surrounded by closely related haplotypes. Outliers, ht 26 and ht 24, are associated with the 152 bp allele at locus F7. Haplotype 12 is associated with the 105 bp allele at locus F13.

haplotypes, ht 26, ht 27 and ht 29, associated with the 152 bp allele at F7 are present on long distal branches, and ht 14, containing the 123 bp allele at F13 is also an obvious outlier node. In Munster and Connaught the most distal haplotypes are found 17 and 13 and mutational steps respectively from ht 57. In Munster five outlier haplotypes are found that are not associated with the less common alleles at F7 and F13 but are all closely related containing alleles 164 bp, 166 / 168 bp and 115 bp (at F3, F7 and F13 respectively). This might suggest that the most parsimonious network has not been found indicating the possible limitations of this approach. However, a number of outlier haplotypes containing

the 152 bp allele at F7 (ht 26, ht 24) and the 105 bp allele at F13 (ht 12) are also present. Similarly in Connaught the outlier haplotypes are associated with 152 bp at F7 (ht 26, ht 27, ht 28) and 105 bp at F13 (ht 91).

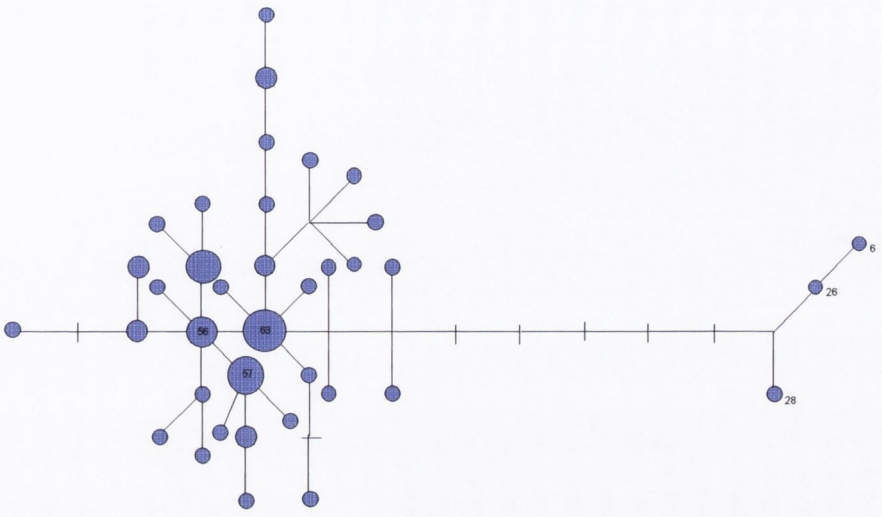


Figure 2.10: Maximum parsimony phylogeny relating haplotypes sampled in the Ulster population (n = 63). The most common haplotype, ht 63, is found centrally in the network. Prominent star-like patterns are evident from all common nodes. The three distal haplotypes are associated with the uncommon 152 bp allele at locus F7.

The Togolese and Turkish networks are more diverse. Haplotypes are spread across a large mutational distance and there is a lack of an obvious central core haplotype with a star-like radiation pattern.

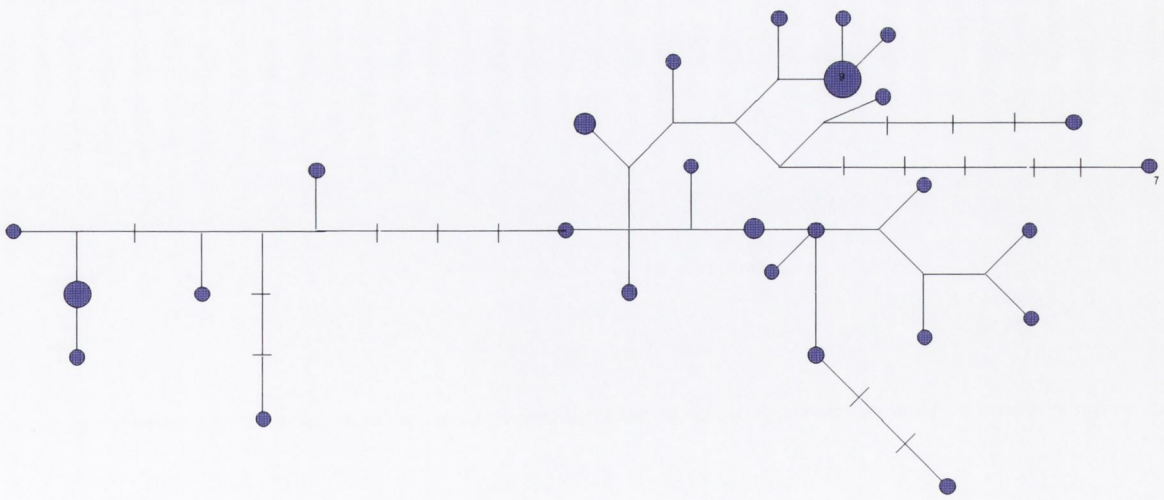


Figure 2.11: Maximum parsimony network relating haplotypes sampled in the Togolese population (n = 36). Unlike the Irish phylogenies the most common ht 9 is not centrally located and does not demonstrate a star-like pattern.

The most distal haplotypes from the most common haplotypes in the Togolese and Turkish networks are 16 and 15 mutational steps respectively. Two of the outlier haplotypes detected in the Irish networks (ht 26, ht 12) are also found distally in the Turkish network. Although in the Irish populations the outlier haplotypes were never sampled more than once, in the Turkish population ht 26 was sampled in 7% of all chromosomes.

The three most common Irish haplotypes are also found in the Turkish sample but they do not cluster in the Turkish network as they do within the Irish subpopulations. Haplotype 57, the most common in Turkey is found distal from the other common Irish haplotypes, ht 56 and ht 63.

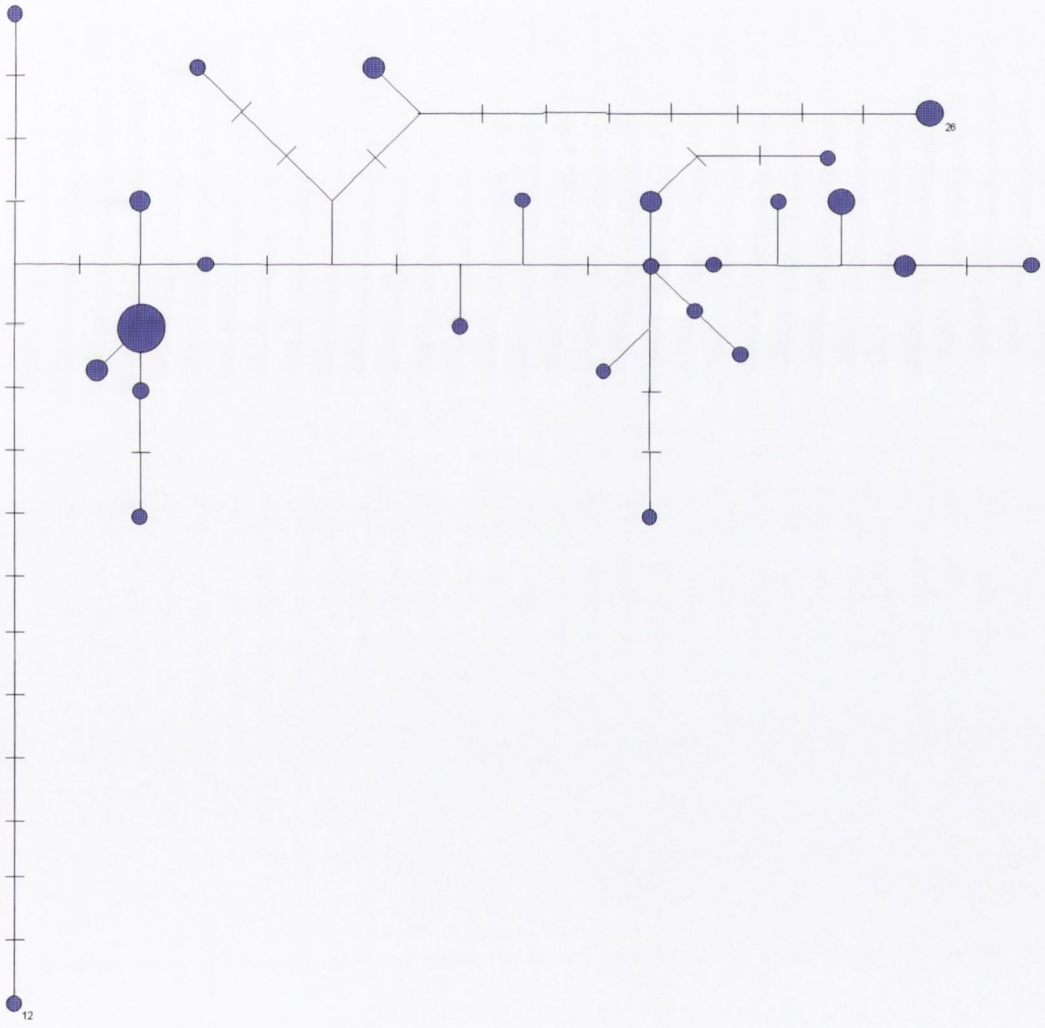


Figure 2.12: Maximum parsimony network relating haplotypes sampled in the Turkish population (n = 43). Two of the outlier haplotypes present in the Irish populations, ht 26 and ht 12, are also detected in the Turkish population and are indicated on the network.

2.3.16. Genetic distance measures and phylogenetic analysis: haplotypes as OTUs

The distance measure, D_{asm}, developed to compare haplotypes of individuals within and between populations was applied here among individuals among populations. This distance measure, devised by Bowcock *et al. (1994),* was originally developed to measure the genetic distance between individuals in terms of the proportion of alleles shared between them, 1 - Ps, where Ps is the number of shared alleles summed over loci / (2 x number of loci compared). This measure was used to examine variation among 30 microsatellites in 10 individuals in each of 14 distantly related human populations. Of the individuals examined 87.8% formed distinct clusters corresponding to the continent of origin of the sample. This strongly suggested that a simple allele sharing measure could reliably distinguish between individuals of diverse ethnic groups.

This allele sharing measure may be defined by the formula

$$D_{asm} = 1 - \{ \sum_{j=1}^{r} i_{AS} / 2r \}$$

where r is the number of loci and i_{AS} is the number of alleles two individuals have in common at the j-th locus, summed over all loci, and r is the number of loci surveyed. D_{asm} can have a value of either $1 \rightarrow n$ or 0 depending on similarity (0) or difference ($1 \rightarrow n$) between two individuals where n is the number of differing loci between two haplotypes. For example in a 5 locus haplotype n = 5 if all loci differ, n = 4 if alleles at 4 loci are different but similar at the fifth, and so on.

D_{asm} was calculated using individual haplotypes as OTUs. Because most of the genetic diversity is found between individuals within populations (88.5%) and few haplotypes are shared between individuals it was reasoned that for this comparison an approximately equal number of individuals from each population should be compared. A random subset (n = 50) of the Irish sample was taken and compared to the Togolese (n = 36) and Turkish (n = 44) populations. A total of 65 haplotypes were detected in the 130 individuals. In the Togolese population 64% of haplotypes were unique to individuals, in the Turkish population 39% were unique to individuals and in the Irish population 32% were unique to individuals. Only 8% of haplotypes were shared in more than one population. The haplotypes and their occurrence in each of the three populations are given in **table 2.15.**

Figure 2.13: Neighbour-joining tree of haplotypes related by the D_{asm} genetic distance. Only two haplotypes are shared between all three populations. Most evident is the African specific monophyletic clade (indicated by dark line) in which 59.3% of the African haplotypes are found. African haplotypes not found within the African specific clade are found on longer branches than non-African haplotypes.

The D_{asm} distance matrix with all pairwise comparisons of the 65 haplotypes is given in **table 2.16.** A neighbour-joining tree relating all haplotypes was drawn **(figure 2.13).** The neighbour-joining tree clearly shows three main branches extending from the root of the tree which define 3 major clades. Most evident is an African-specific clade incorporating 59.3% of all of the African haplotypes. In this major clade only 2 of the 18 haplotypes are non-African, one being present in the Turkish sample alone, and the other being shared between the Turkish population and Ireland. Of the African haplotypes which do not cluster in the African-specific clade most exhibit longer branch lengths than the Turkish and Irish haplotypes. Neither the Irish nor the Turkish haplotypes cluster separately to any significant extent. Only two (ht 37 and ht 43) of the sixty-five haplotypes are shared between all three population groups. These shared haplotypes tend to have shorter branches and central branching points. Haplotype 43 is the most common haplotype in both the Turkish and Irish sample, and ht 37 is the second most common haplotype in the Irish sample. The most common African haplotype (ht 7) is not detected in either the Turkish or the Irish sample.

Ht	177e8-F3	177e8-F7	177e8-F13	177e8-F14	177e8-F29	TOGO (n =36)	TURKEY (n = 44)	IRELAND (n=50)
1	164	152	115	134	152	...	3	...
2	164	152	115	134	154	...	...	1
3	162	156	115	131	150	1	...	...
4	162	162	111	131	154	1	...	...
5	166	162	113	134	156	1	...	...
6	164	162	115	131	154	1	...	...
7	162	164	111	131	154	6	...	...
8	162	164	111	134	154	1	...	...
9	162	164	111	131	156	1	...	...
10	162	166	105	131	156	...	1	...
11	160	166	113	131	154	1	...	...
12	162	168	111	131	154	2	...	...
13	160	168	111	131	162	1	...	...
14	164	168	115	131	154	...	...	2
15	164	168	115	134	154	...	3	...
16	166	168	119	131	154	1	...	...
17	168	168	119	131	156	1	...	...
18	168	170	111	134	154	...	1	...
19	162	170	115	134	148	...	1	...
20	162	170	115	134	154	...	...	2

Table 2.15: A total of 65 haplotypes were detected in the African population (n = 36), the Turkish population (n = 44) and a random subset (n = 50) of the Irish population. Haplotypes and their absolute frequencies in each of the three populations are given. Haplotype numbers do not correspond to those previously used. For reference, **appendix A** gives corresponding haplotype identities.

Ht	177e8-F3	177e8-F7	177e8-F13	177e8-F14	177e8-F29	TOGO (n =36)	TURKEY (n = 44)	IRELAND (n=50)
21	164	170	115	131	154	...	1	...
22	164	170	115	131	156	...	...	1
23	166	170	115	131	156	1	...	...
24	164	170	115	131	158	...	2	...
25	164	170	119	131	156	1	...	...
26	160	172	111	131	154	1	...	...
27	164	172	111	131	164	3	...	...
28	164	172	111	131	166	1	...	...
29	164	172	113	131	152	...	...	1
30	166	172	113	131	154	1	...	...
31	164	172	115	131	150	...	1	2
32	164	172	115	131	152	...	...	1
33	164	172	115	134	152	...	...	3
34	166	172	115	134	152	...	...	1
35	162	172	115	131	154	1	...	...
36	162	172	115	134	154	...	...	2
37	164	172	115	131	154	2	1	5
38	164	172	115	134	154	...	3	4
39	166	172	115	131	154	...	1	1
40	166	172	115	134	154	...	2	1
41	162	172	115	131	156	1	...	...
42	162	172	115	134	156	...	...	1
43	164	172	115	131	156	1	10	8
44	164	172	115	134	156	...	...	1
45	164	172	115	131	158	...	2	3
46	166	172	115	134	158	...	1	...
47	164	172	115	131	160	1	...	1
48	164	172	117	134	154	...	...	1
49	164	172	117	131	156	...	1	1
50	164	172	119	131	156	...	...	1
51	164	172	119	134	156	...	1	...
52	160	174	111	131	160	1	...	...
53	160	174	111	131	162	1	...	...
54	164	174	111	131	164	1	...	...
55	164	174	115	131	150	...	1	...
56	164	174	115	131	154	...	1	1
57	164	174	115	131	156	...	2	3
58	164	174	115	131	162	...	1	...
59	164	174	117	131	154	...	1	...
60	164	174	117	134	154	...	1	...
61	166	176	115	134	154	...	...	1
62	166	176	115	137	154	...	1	...
63	164	176	115	131	162	1	...	...
64	164	178	115	131	154	...	1	...
65	164	178	115	134	154	...	...	1

Table 2.15 *cont.*: A total of 65 haplotypes were detected in the African population (n = 36), the Turkish population (n = 44) and a random subset (n = 50) of the Irish population. See **appendix A** to cross reference previous haplotype nomenclature.

2.3.17. Multidimensional scaling: population substructure at the haplotypic level

The D_{asm} genetic distance matrix for 65 haplotypes was used to produce eigenvalues to assess the difference between haplotypes within populations using MDS. The plot values for each haplotype are given in **table 2.17.** When haplotypes are plotted using eigenvalues obtained from matrices of D_{asm} distances, the distribution is characterised by a lack of clustering of any population specific (including African) haplotypes **(figure 2.14).**

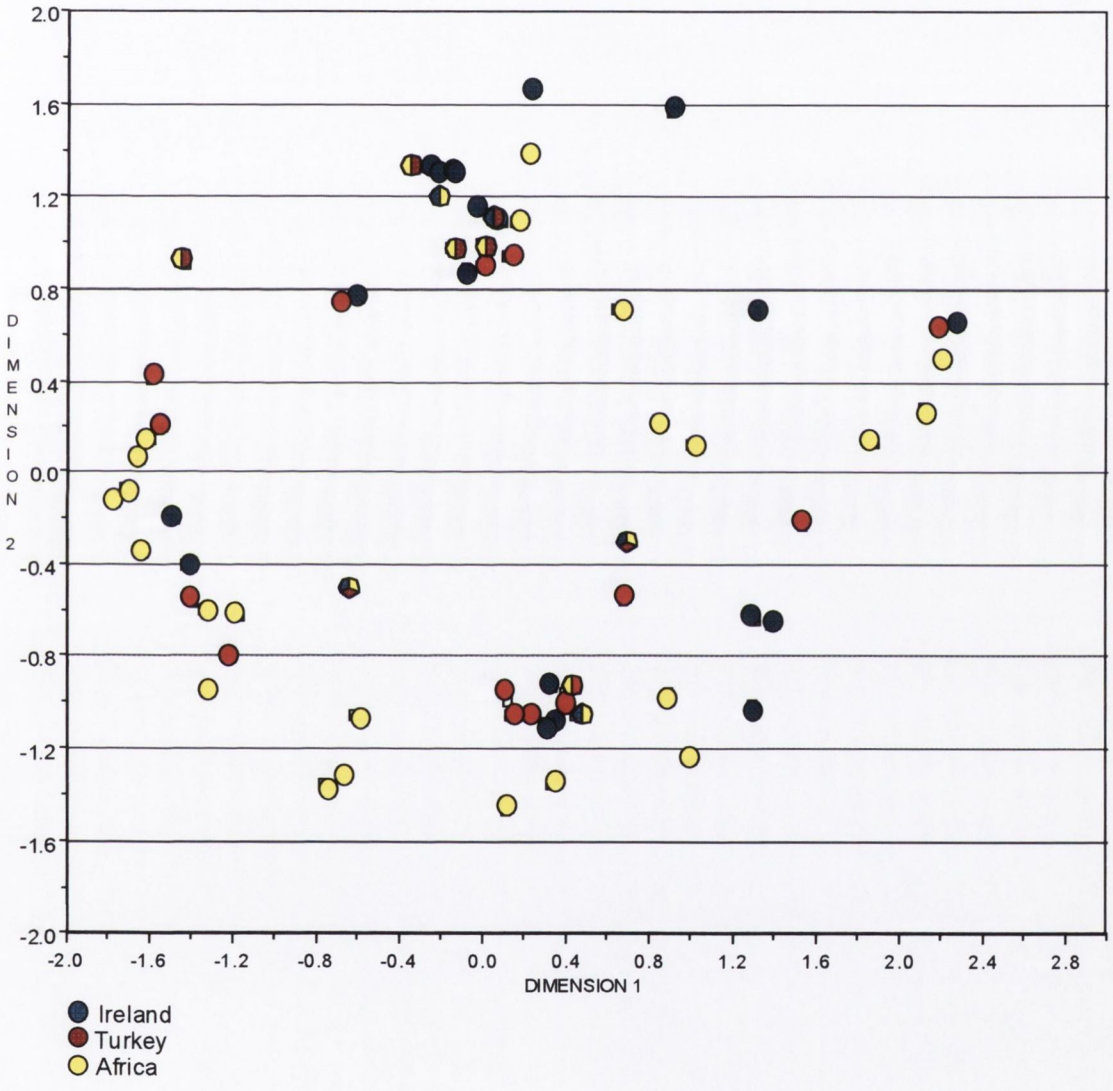
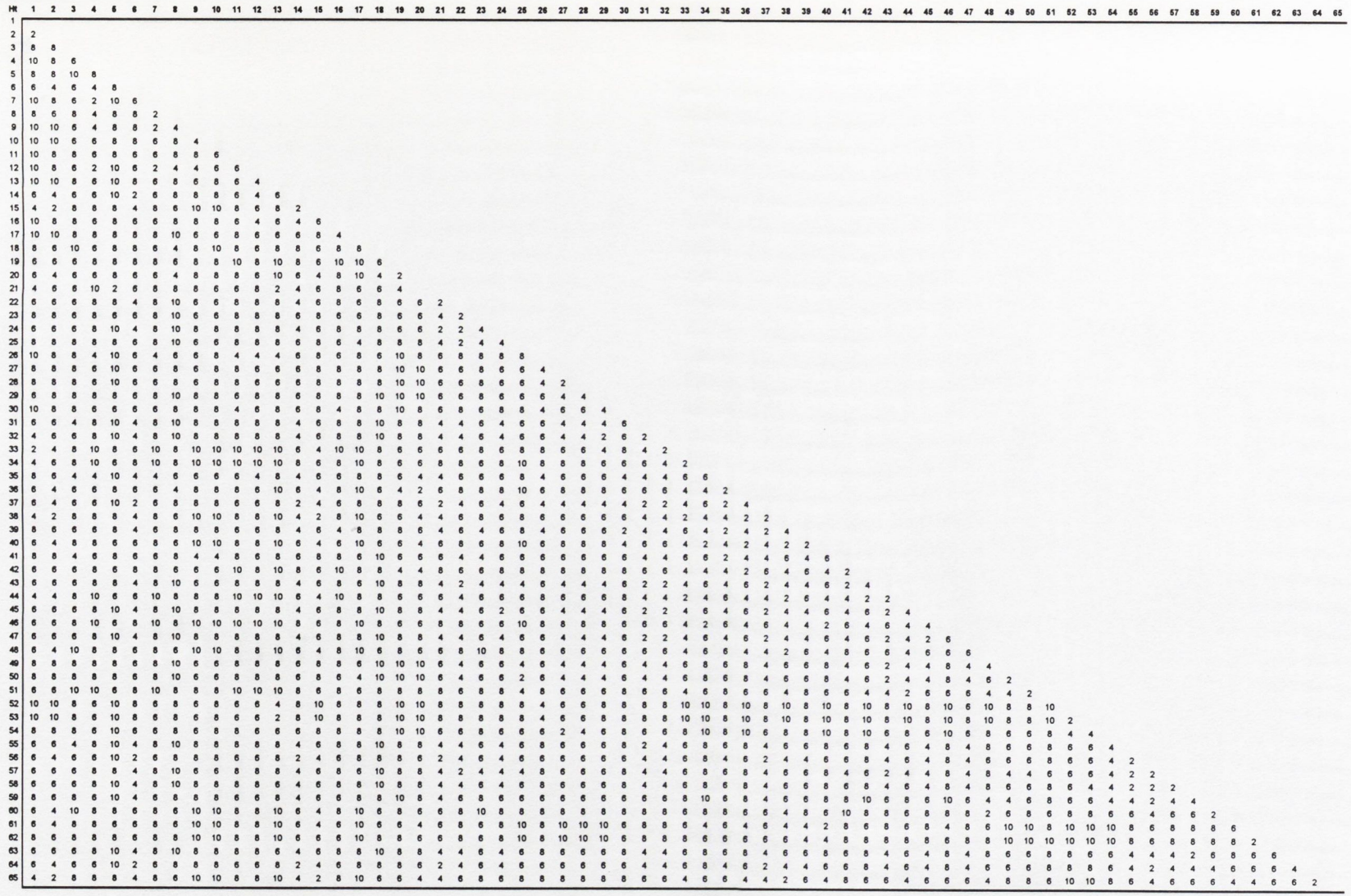


Figure 2.14: MDS plot of haplotypes related by D_{asm} genetic distance. No association among population specific haplotypes is evident.



Ht	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65
1
2	2
3	8	8
4	10	8	6
5	8	8	10	8
6	6	4	6	4	8
7	10	8	6	2	10	6
8	8	6	8	4	8	8	2
9	10	10	6	4	8	8	2	4
10	10	10	6	6	8	8	6	8	4
11	10	8	8	6	8	6	6	8	8	6
12	10	8	6	2	10	6	2	4	4	6	6
13	10	10	8	6	10	8	6	8	6	8	6	4
14	6	4	6	6	10	2	6	8	8	8	6	4	6
15	4	2	8	8	8	4	8	6	10	10	8	6	8	2
16	10	8	8	6	8	6	6	8	8	8	6	4	6	4	6
17	10	10	8	8	8	8	8	10	6	6	8	6	6	6	8	4
18	8	6	10	6	8	8	6	4	8	10	8	6	8	8	6	8	8
19	6	6	6	8	8	8	8	6	8	8	10	8	10	8	6	10	10	6
20	6	4	6	6	8	6	6	4	8	8	8	6	10	6	4	8	10	4	2
21	4	6	6	10	2	6	8	8	8	6	6	8	2	4	6	8	6	6	4
22	6	6	6	8	8	4	8	10	6	6	8	8	8	4	6	8	6	8	6	6	2
23	8	8	6	8	6	6	8	10	6	6	8	8	8	6	8	6	6	8	6	6	4	2
24	6	6	6	8	10	4	8	10	8	8	8	8	8	4	6	8	8	8	6	6	2	2	4
25	8	8	8	8	8	6	8	10	6	6	8	8	8	6	8	6	4	8	8	8	4	2	4	4
26	10	8	8	4	10	6	4	6	6	8	4	4	4	6	8	6	8	6	10	8	6	8	8	8	8
27	8	8	8	6	10	6	6	8	6	8	8	6	6	6	8	8	8	8	10	10	6	6	8	6	6	4
28	8	8	8	6	10	6	6	8	6	8	8	6	6	6	8	8	8	8	10	10	6	6	8	6	6	4	2
29	6	8	8	8	8	6	8	10	8	8	6	8	8	6	8	8	8	10	10	10	6	6	8	6	6	6	4	4
30	10	8	8	6	6	6	6	8	8	8	4	6	8	6	8	4	8	8	10	8	6	8	6	8	8	4	6	6	4
31	6	6	4	8	10	4	8	10	8	8	8	8	8	4	6	8	8	10	8	8	4	4	6	4	6	6	4	4	4	6
32	4	6	6	8	10	4	8	10	8	8	8	8	8	4	6	8	8	10	8	8	4	4	6	4	6	6	4	4	2	6	2
33	2	4	8	10	8	6	10	8	10	10	10	10	10	6	4	10	10	8	6	6	6	6	8	6	8	8	6	6	4	8	4	2
34	4	6	8	10	6	8	10	8	10	10	10	10	10	8	6	8	10	8	6	6	8	8	6	8	10	8	8	8	6	6	6	4	2
35	8	6	4	4	10	4	4	6	6	6	6	4	8	4	6	6	8	8	6	4	4	6	6	6	8	4	6	6	6	4	4	4	6	6
36	6	4	6	6	8	6	6	4	8	8	8	6	10	6	4	8	10	6	4	2	6	8	8	8	10	6	8	8	8	6	6	6	4	4	2
37	6	4	6	6	10	2	6	8	8	8	6	6	8	2	4	6	8	8	8	6	2	4	6	4	6	4	4	4	4	4	2	2	4	6	2	4
38	4	2	8	8	8	4	8	6	10	10	8	8	10	4	2	8	10	6	6	4	4	6	8	6	8	6	6	6	6	6	4	4	2	4	4	2	2
39	8	6	6	6	8	4	6	8	8	8	6	6	8	4	6	4	8	8	8	6	4	6	4	6	8	4	6	6	6	2	4	4	6	4	2	4	2	4
40	6	4	8	8	6	6	8	6	10	10	8	8	10	6	4	6	10	6	6	4	6	8	6	8	10	6	8	8	8	4	6	6	4	2	4	2	4	2	2
41	8	8	4	6	8	6	6	8	4	4	8	6	8	6	8	8	6	10	6	6	6	4	4	6	6	6	6	6	6	6	4	4	6	6	2	4	4	6	4	6
42	6	6	6	8	6	8	8	6	6	6	10	8	10	8	6	10	8	8	8	4	4	8	6	6	8	8	8	8	8	8	6	6	4	4	4	2	6	4	6	4	2
43	6	6	6	8	8	4	8	10	6	6	8	8	8	4	6	8	6	10	8	8	4	2	4	4	4	6	4	4	4	6	2	2	4	6	4	6	2	4	4	6	2	4
44	4	4	8	10	6	6	10	8	8	8	10	10	10	6	4	10	8	8	6	6	6	4	6	6	6	8	6	6	6	8	4	4	2	4	6	4	4	2	6	4	4	2	2
45	6	6	6	8	10	4	8	10	8	8	8	8	8	4	6	8	8	10	8	8	4	4	6	2	6	6	4	4	4	6	2	2	4	6	4	6	2	4	4	6	4	6	2	4
46	6	6	8	10	6	8	10	8	10	10	10	10	10	8	6	8	10	8	6	6	8	8	6	6	10	8	8	8	8	6	6	6	4	2	6	4	6	4	4	2	6	4	6	4	4
47	6	6	6	8	10	4	8	10	8	8	8	8	8	4	6	8	8	10	8	8	4	4	6	4	6	6	4	4	4	6	2	2	4	6	4	6	2	4	4	6	4	6	2	4	2	6
48	6	4	10	8	8	6	8	6	10	10	8	8	10	6	4	8	10	6	8	6	6	8	10	8	8	6	6	6	6	6	6	6	4	6	6	4	4	2	6	4	8	6	6	4	6	6	6
49	8	8	8	8	8	6	8	10	6	6	8	8	8	6	8	8	6	10	10	10	6	4	6	6	4	6	4	4	4	6	4	4	6	8	6	8	4	6	6	8	4	6	2	4	4	8	4	4
50	8	8	8	8	8	6	8	10	6	6	8	8	8	6	8	6	4	10	10	10	6	4	6	6	2	6	4	4	4	6	4	4	6	8	6	8	4	6	6	8	4	6	2	4	4	8	4	6	2
51	6	6	10	10	6	8	10	8	8	8	10	10	10	8	6	8	6	8	8	8	8	6	8	8	4	8	6	6	6	8	6	6	4	6	8	6	6	4	8	6	6	4	4	2	6	6	6	4	4	2
52	10	10	8	6	10	8	6	8	6	8	6	6	4	8	10	8	8	8	10	10	8	8	8	8	8	4	6	6	8	8	8	8	10	10	8	10	8	10	8	10	8	10	8	10	8	10	6	10	8	8	10
53	10	10	8	6	10	8	6	8	6	8	6	6	2	8	10	8	8	8	10	10	8	8	8	8	8	4	6	6	8	8	8	8	10	10	8	10	8	10	8	10	8	10	8	10	8	10	8	10	8	8	10	2
54	8	8	8	6	10	6	6	8	6	8	8	6	6	6	8	8	8	8	10	10	6	6	8	6	6	6	2	4	6	8	6	6	8	10	8	10	6	8	8	10	8	10	6	8	6	10	6	8	6	6	8	4	4
55	6	6	4	8	10	4	8	10	8	8	8	8	8	4	6	8	8	10	8	8	4	4	6	4	6	8	6	6	6	8	2	4	6	8	6	8	4	6	6	8	6	8	4	6	4	8	4	8	6	6	8	6	6	4
56	6	4	6	6	10	2	6	8	8	8	6	6	8	2	4	6	8	8	8	6	2	4	6	4	6	6	6	6	6	6	4	4	6	8	4	6	2	4	4	6	6	8	4	6	4	8	4	6	6	6	8	6	6	4	2
57	6	6	6	8	8	4	8	10	6	6	8	8	8	4	6	8	6	10	8	8	4	2	4	4	4	8	6	6	6	8	4	4	6	8	6	8	4	6	6	8	4	6	2	4	4	8	4	8	4	4	6	6	6	4	2	2
58	6	6	6	8	10	4	8	10	8	8	8	8	6	4	6	8	8	10	8	8	4	4	6	4	6	8	6	6	6	8	4	4	6	8	6	8	4	6	6	8	6	8	4	6	4	8	4	8	6	6	8	6	4	4	2	2	2
59	8	6	8	6	10	4	6	8	8	8	6	6	8	4	6	6	8	8	10	8	4	6	8	6	6	6	6	6	6	6	6	6	8	10	6	8	4	6	6	8	8	10	6	8	6	10	6	4	4	6	8	6	6	4	4	2	4	4
60	6	4	10	8	8	6	8	6	10	10	8	8	10	6	4	8	10	6	8	6	6	8	10	8	8	8	8	8	8	8	8	8	6	8	8	6	6	4	8	6	10	8	8	6	8	8	8	2	6	8	6	8	8	6	6	4	6	6	2
61	6	4	8	8	6	6	8	6	10	10	8	8	10	6	4	6	10	6	6	4	6	8	6	8	10	8	10	10	10	6	8	8	6	4	6	4	6	4	4	2	8	6	8	6	8	4	8	6	10	10	8	10	10	10	8	6	8	8	8	6
62	8	6	8	8	8	6	8	8	10	10	8	8	10	6	6	6	10	8	8	6	6	8	6	8	10	8	10	10	10	6	8	8	8	6	6	6	6	6	4	4	8	8	8	8	8	6	8	8	10	10	10	10	10	10	8	6	8	8	8	8	2
63	6	6	6	8	10	4	8	10	8	8	8	8	6	4	6	8	8	10	8	8	4	4	6	4	6	8	6	6	6	8	4	4	6	8	6	8	4	6	6	8	6	8	4	6	4	8	4	8	6	6	8	8	6	6	4	4	4	2	6	8	6	6
64	6	4	6	6	10	2	6	8	8	8	6	6	8	2	4	6	8	8	8	6	2	4	6	4	6	6	6	6	6	6	4	4	6	8	4	6	2	4	4	6	6	8	4	6	4	8	4	6	6	6	8	8	8	6	4	2	4	4	4	6	6	6	4
65	4	2	8	8	8	4	8	6	10	10	8	8	10	4	2	8	10	6	6	4	4	6	8	6	8	8	8	8	8	8	6	6	4	6	6	4	4	2	6	4	8	6	6	4	6	6	6	4	8	8	6	10	10	8	6	4	6	6	6	4	4	6	6	2

Table 2.16: Dasm distances (lower diagonal) calculated between 65 haplotypes in 3 populations.

Ht	177e8-F3	177e8-F7	177e8-F13	177e8-F14	177e8-F29	Dimension 1	Dimension 2
1	164	152	115	134	152	2.1798	0.6288
2	164	152	115	134	154	2.2726	0.6603
3	162	156	115	131	150	2.1933	0.4928
4	162	162	111	131	154	2.1118	0.2902
5	166	162	113	134	156	-1.1836	-0.6202
6	164	162	115	131	154	1.8703	0.1400
7	162	164	111	131	154	-0.6207	-1.0459
8	162	164	111	134	154	-1.6586	0.0772
9	162	164	111	131	156	-1.7719	-0.1031
10	162	166	105	131	156	-1.3968	-0.5545
11	160	166	113	131	154	-1.7316	-0.0624
12	162	168	111	131	154	-1.3521	-0.5919
13	160	168	111	131	162	-1.6387	-0.3274
14	164	168	115	131	154	-1.4420	-0.4006
15	164	168	115	134	154	-1.4434	0.9027
16	166	168	119	131	154	-0.6733	-1.3192
17	168	168	119	131	156	-1.3417	-0.9452
18	168	170	111	134	154	-1.2363	-0.8000
19	162	170	115	134	148	-1.5985	0.4069
20	162	170	115	134	154	-1.4852	-0.1937
21	164	170	115	131	154	-1.6037	0.1431
22	164	170	115	131	156	1.3768	-0.6613
23	166	170	115	131	156	-1.5599	0.1854
24	164	170	115	131	158	1.5421	-0.2227
25	164	170	119	131	156	0.1078	-1.4556
26	160	172	111	131	154	-0.7672	-1.3496
27	164	172	111	131	164	0.2082	1.3920
28	164	172	111	131	166	0.9814	-1.2410
29	164	172	113	131	152	1.2926	-1.0036
30	166	172	113	131	154	0.3260	-1.3569
31	164	172	115	131	150	1.3129	-0.6338
32	164	172	115	131	152	1.3065	0.7190
33	164	172	115	134	152	0.8933	1.5784
34	166	172	115	134	152	0.2365	1.6722
35	162	172	115	131	154	0.2895	-1.0847
36	162	172	115	134	154	0.0888	1.1020
37	164	172	115	131	154	-0.1634	0.9814
38	164	172	115	134	154	-0.2138	1.2999
39	166	172	115	131	154	-0.1467	1.3207
40	166	172	115	134	154	-0.3088	1.3244
41	162	172	115	131	156	0.1436	1.1000
42	162	172	115	134	156	0.3944	-0.9705
43	164	172	115	131	156	-0.0212	1.1574
44	164	172	115	134	156	-0.2560	1.3231
45	164	172	115	131	158	0.6879	-0.2963
46	166	172	115	134	158	0.1301	-1.0460
47	164	172	115	131	160	-0.2144	1.2061
48	164	172	117	134	154	0.3617	-1.0596
49	164	172	117	131	156	0.4162	-1.0619
50	164	172	119	131	156	0.3938	-1.0079
51	164	172	119	134	156	0.1023	-0.9856
52	160	174	111	131	160	0.8887	-0.9563
53	160	174	111	131	162	1.0022	0.1215
54	164	174	111	131	164	0.6328	0.7192
55	164	174	115	131	150	0.3366	-0.9242
56	164	174	115	131	154	0.2353	-1.0534
57	164	174	115	131	156	-0.6631	-0.5109
58	164	174	115	131	162	-0.0140	0.9934
59	164	174	117	131	154	0.1009	0.9490
60	164	174	117	134	154	-0.0169	0.8970
61	166	176	115	134	154	-0.6081	0.7524
62	166	176	115	137	154	-0.6913	0.7424
63	164	176	115	131	162	0.8283	0.2215
64	164	178	115	131	154	0.6735	-0.5439
65	164	178	115	134	154	-0.0957	0.8893

Table 2.17: Eigenvalues for D_{asm} MDS plot of 65 haplotypes.

2.3.18. Mismatch distributions

Mismatch distributions, histograms showing the frequency of the number of pairwise differences between haplotypes at linked loci, have been investigated in mtDNA systems in an attempt to assess evidence for population expansions in human populations *(Rogers and Harpending 1992; Rogers 1995; Rogers et al 1996; Harpending et al. 1998; Reich and Goldstein 1998).* In an expanding population, or a population that has been exposed to selective sweeps, haplotype lineages have a greater chance of survival and therefore many pairs of haplotypes will share the same number of differences between them. Mathematical models and simulations suggest that a smooth, unimodal distribution is a signature of either a population expansion or a selective sweep, whereas a multimodal distribution is typical of a population of long term fixed size.

In a comparison of the past dynamics of the African, Turkish, and Irish populations pairwise comparisons of the number of repeat unit differences between haplotypes were made between all possible pairs of haplotypes within each population using the distance measures D_{asm} and D_{swasm}.

D_{asm} has been described above. The distance measure D_{swasm} (stepwise allele sharing measure) was used to calculate the number of mutational steps between two different haplotypes. Based on the stepwise mutation model of microsatellites D_{swasm} is calculated as the minimum number of mutational steps required to transform the haplotype of one individual into the haplotype of another individual. The distance metric is expressed as follows:

$$D_{swasm} = \sum_{j=1}^{r} \min \{\text{haplotype A} \leftrightarrow \text{haplotype B}\}$$

The pairwise difference matrices for D_{asm} and D_{swasm} were entered into an Excel™ spreadsheet on a Macintosh computer and a simple count established the frequency of each class of number of differences.

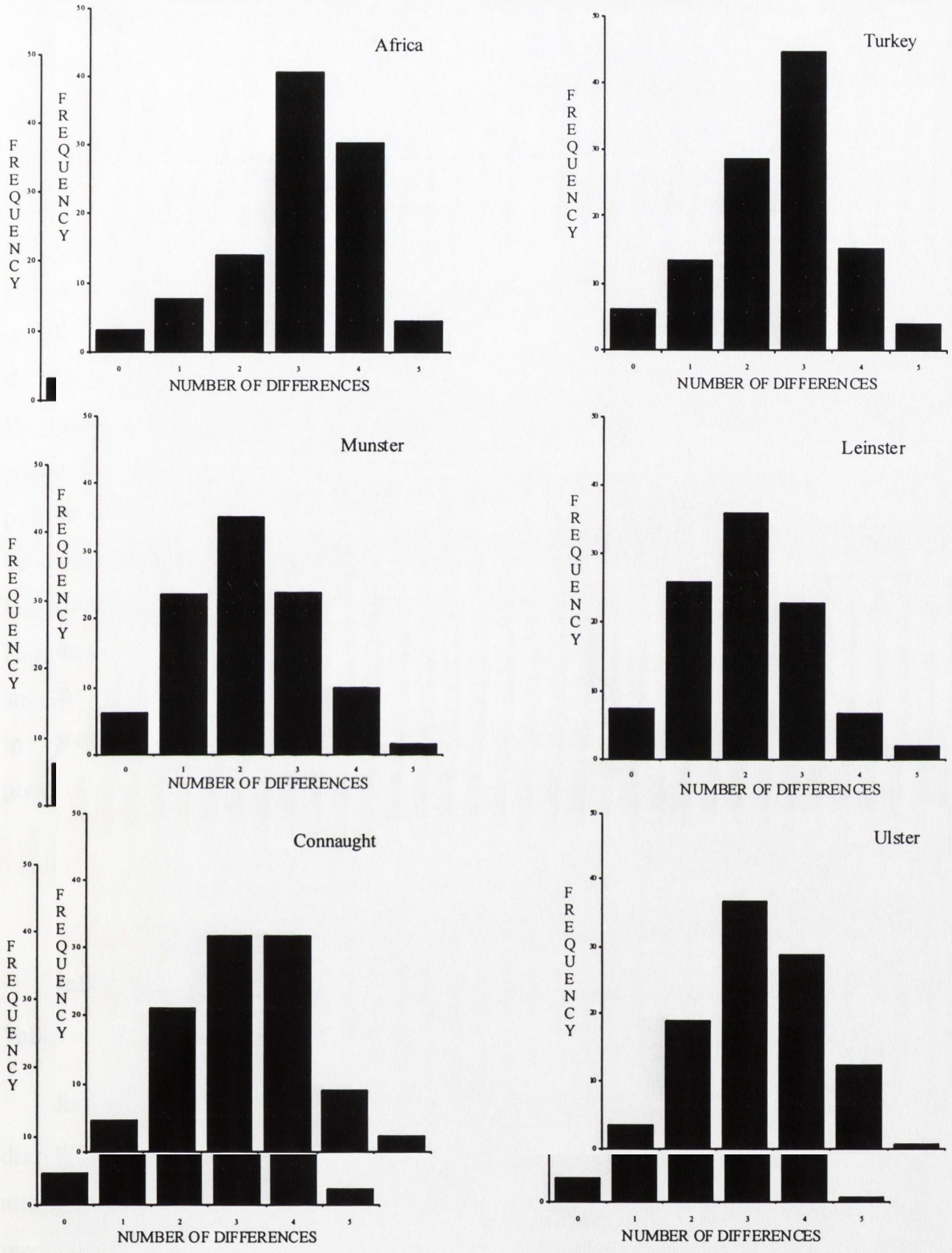


Figure 2.15: D_{asm} mismatch distributions for 5 loci X chromosome microsatellite haplotypes.

The D_{asm} mismatch distributions for all populations are similar, the modal number of differences being either 2 differences or 3 differences **(figure 2.15).** It is expected that in an older population the modal point of the distribution would be more distal from the

vertical axis. The results of all the above analyses have suggested that the African population is an older population, and additionally D_{asm} mismatches have detected this.

By analogy to Shriver *et al. (1997)* an analogous measure of pairwise differences, D_{swasm}, was used (taking only a random subset of the Irish population; Ireland n = 50) and the resulting distributions were plotted **(figure 2.16).** Simple summary statistics are shown in **table 2.18.** A comparison of the three mismatch distributions indicated a unimodal distribution for the African population with somewhat bimodal distributions for the Turkish and Irish populations. This suggests, in accordance with Rogers and Harpending *(1992),* that African populations have undergone expansions in the past whereas other global populations (including Irish and Turkish) have had long term fixed population sizes. However, the bimodal distributions in Turkey and Ireland are likely to be the result of the presence of both the 152 bp allele at locus F7 and the 105 bp allele at locus F13 which are present in both these populations but not in the African population. Haplotypes associated with these alleles will differ dramatically from those with the smaller allelic variants. The presence of alleles resulting from large mutations result also in the small bimodal tail of the African distribution which is likely to be arising from the presence of the 156 bp allele at locus F7.

	Mean	Mode	Median
Africa	7.154	8	7
Turkey	4.939	3	4
Ireland	3.828	2	3

Table 2.18: Summary statistics for D_{swasm} mismatch distributions

In order to attempt to quantify the differences between the three comparable distributions a Mann-Whitney test of significance was applied. This statistical test attempts to reject the null hypothesis (H_o) that the median of one population equals the median of another. For all three comparisons at $\alpha = 0.05$, the H_o was significantly rejected ($p = 0.0001$). Thus the distributions are significantly different from each other and this difference is most likely due to the disparate demographic histories of these populations.

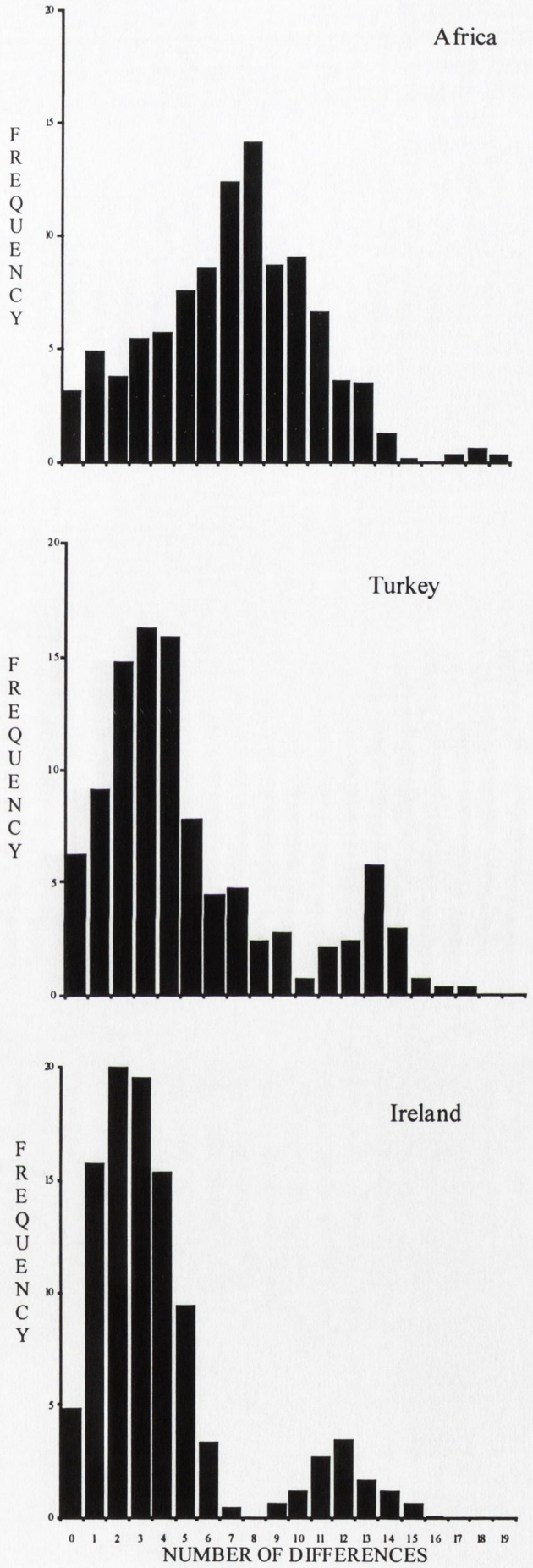


Figure 2.16: D_{swasm} mismatch distributions for X chromosome microsatellites

The modal points differ from 2.5 - 4 orders of magnitude, the African population being the most distal from the vertical axis. Additionally, the mean of the distribution is much greater in the African population (7.154) than in the Turkish (4.939) and the Irish (3.828) populations.

2.3.19. Dating the split between Africans and non-Africans

The average square distance (ASD) measure *(Goldstein et al 1995; Slatkin 1995)* and the D_{dm} measure *(Goldstein et al 1995)* are the distance measures of choice for estimating relative times of divergence from microsatellite data as they increase linearly with time. D_{dm} was adapted from ASD in order to reduce the effect of intrapopulation variation and to remove its dependence on population size. However this measure has been developed for the estimation of divergence times of lineages rather than populations and therefore the ASD measure is more appropriate for assessing the coalescence here.

$$ASD = V_A + V_B + (\mu_A - \mu_B)^2$$

where V_A, V_B, μ_A and μ_B are the variances and the means of the allele size in populations A and B.

The estimation of divergence times from ASD is calculated as

$$ASD = 2\mu t$$

where μ is the mutation rate and t is the time in generations.

It is necessary that an appropriate mutation rate for X chromosome microsatellites be used. A number of mutation rates have been estimated for autosomal microsatellites (for a combination of di- and tetranucleotide repeats) and range from 2.1×10^{-3} *(Weber and Wong 1993)* to 1.5×10^{-4} per generation *(Underhill et al. 1996).* For Y chromosome microsatellites mutation rate estimates ranging from 2.0×10^{-3} *(Heyer et al 1997)* to 1.2×10^{-4} per generation *(Cooper et al 1996)* have been described. Nucleotide substitution rate estimates for the X chromosome are intermediate between the rate for autosomes and the Y chromosome *(Weissenbach 1992, Zhong 1996)* although a faster mutation rate of 0.01 per generation for X chromosome microsatellites has been suggested by Mahtani and Willard *(1993).*

The ASD between African and non-African populations (averaged between Turkey, Denmark and Ireland) was calculated to be 4.9293. The estimated divergence times of African and non-African populations are presented in **table 2.19** and have been

calculated using the above ASD measure, a selection of mutation rates and a standard generation time of 25 years.

Mutation rate	Genetic system	Reference	generations)	(years)
2.1×10^{-3}	autosomes	*Weber and Wong 1993*	1,174	29,341
5.6×10^{-4}	autosomes	*Goldstein et al. 1995*	4,401	110,029
1.5×10^{-4}	autosomes	*Underhill et al. 1996*	16,431	410,775
1.0×10^{-2}	X chromosome	*Mahtani and Willard 1993*	164	4,107
1.5×10^{-3}	X chromosome	*Weissenbach et al. 1992; Zhong et al. 1994*	2,465	61,616
2.0×10^{-3}	Y chromosome	*Heyer et al. 1997*	1,232	30,808
4.8×10^{-4}	Y chromosome	*Cooper et al. 1996*	5,135	128,367
1.2×10^{-4}	Y chromosome	*Cooper et al. 1996*	20,539	513,468
Mean time of split between African and non-African X chromosomes			**6,443**	**161,064**

Table 2.19: Divergence times between African and non-African X chromosomes based on ASD = 4.9293 where ASD = $2\mu t$

All of the dates were averaged and an *ad hoc* average coalescence date for the split between African and non-African X chromosomes was estimated to be 161,064 *YBP*. The averaged coalescence date for the split between Irish and Turkish X chromosomes was younger, estimated to be 122,834 *YBP*. The split between Irish and Danish X chromosomes was even younger, estimated at 72,528 *YBP*.

Mutation rate	Turkey / Ireland	Denmark / Ireland
2.1×10^{-3}	22,702	13,404
5.6×10^{-4}	85,134	50,268
1.5×10^{-4}	317,833	187,667
1.0×10^{-2}	4,768	2,815
1.5×10^{-3}	31,783	18,767
2.0×10^{-3}	23,838	14,075
4.8×10^{-4}	99,323	58,645
1.2×10^{-4}	397,291	234,583
Mean time of split	***122,834***	***72,528***

Table 2.20: Divergence times between Turkish and Irish, and Danish and Irish X chromosomes based on ASD = 3.814 and ASD = 2.252 respectively, and where ASD = $2\mu t$

2.4 Discussion

The patterns of genetic diversity at X chromosome microsatellite loci, described here, are consistent with previous genetic studies of global human population history which have detected the deepest split in worldwide populations as being that between African populations and non-African populations *(Cann et al. 1987, Scozzari et al. 1988, Bowcock et al. 1994, Cavalli-Sforza et al. 1994, Mountain and Cavalli-Sforza 1994, Horai et al. 1995, Jorde et al. 1995, Batzer et al. 1996, Knight et al. 1996, Tishkoff et al. 1996, Pérez-Lezaun et al. 1997, Shriver et al. 1997, Nei and Rhoychoudhury 1993, Vigilant et al. 1991).* Additionally, the coalescence time (160,000 *YBP*) for the split between African populations and non-African populations is in accordance with the Out of Africa model for the recent emergence of modern human populations *(Cann et al 1987, Scozzari et al. 1988, Bowcock et al 1991, 1994, Cavalli-Sforza et al. 1994, Mountain and Cavalli-Sforza 1994, Horai et al 1995, Jorde et al. 1995, Batzer et al. 1995, Knight et al. 1995, Nei 1995, Tishkoff et al. 1996, Pérez-Lezaun et al. 1997, Shriver et al. 1997).*

2.4.1. The deepest evolutionary split is between Africans and non-Africans

Evolutionarily diverse population groups within a species have been shown to exhibit differential spectra of allele frequencies at variable microsatellite loci. Occasionally 'private' alleles are detected at loci that may serve as diagnostic markers for a subspecies or race within a species *(Neel 1973, MacHugh et al. 1997).* X chromosome loci detect a number of alleles specific to African and European populations but do not detect population specific alleles exhibiting differences between the European populations. Significantly, allele 152 bp at locus F7 and allele 105 bp at locus F13 are specific to Europe but are absent in the Togolese population. The most common alleles at each locus are shared between all populations except for the African population at locus F13. However, apart from these, the distributions demonstrate similar patterns, this being consistent with a recent common ancestry for all human populations.

Haplotype sharing is indicative of coancestry and/or recent migration among populations. Here, the greater proportion of shared haplotypes in the Irish and Turkish populations suggests a more recent common ancestry and/or greater level of population

admixture of these two populations than that between either of them and the African population. The association of similar haplotypes is exemplified in the haplotype neighbour-joining tree **(figure 2.13)** which shows the close relationship between all Togolese haplotypes. An ethnically specific monophyletic clade is apparent that contains haplotypes present in the African population only. Those haplotypes present in Africa but not associated with the African-specific clade tend to be placed on much longer branches than haplotypes present in the other populations suggesting that these haplotypes diverged long before the other European X chromosomes. The major African clade may represent a major X chromosome lineage, the other African haplotypes belonging to another or perhaps more than one further lineage. That these haplotypes cluster with European haplotypes might suggest that they may all be found within one separate lineage which may be a subset of the total diversity present on African X chromosomes. This may tentatively suggest a bottleneck during the emergence of Africans from Africa. The association of Irish and Turkish haplotypes to the exclusion of African haplotypes demonstrates, however, the clear distinction between African and European X chromosomes. This also demonstrates the limited power of X STRs in revealing close interpopulation differences. Scozzari *et al. (1997)* also suggest that there is a low level of geographic population substructuring at X STR loci.

Of all species, human phylogenies are possibly the most numerous due to the persisting interest in human origins and ethnic relationships *(Pérez-Lezaun et al. 1996, Bowcock et al. 1994, Forster et al. 1998, Deka et al. 1996, Cooper et al. 1996, Richards et al. 1998, Hammer et al. 1998).* General trends exist in these phylogenies, most notably the separation of African populations from other global populations. In this study also, at the population level, the greatest difference between Africans and non-Africans is detected. The relationship between the African and non-African populations is therefore compatible with what is known of human population history *(Cavalli-Sforza et al. 1994).*

2.4.2. The greater antiquity of the African population

Studies of human population history have indicated that the split between African populations and all other worldwide populations is older than the split between Europeans and Asians *(Bowcock et al. 1994, Cann et al. 1986, Nei 1995, Goldstein 1995, Deka et al.. 1995),* signifying that the variance of the allele distributions should be greater in the

African population than in the European populations. The variances of the allele length repeats do vary in the different populations and, most importantly, the overall variance is highest in the African population. Variance is lowest in the Irish population and the Turkish population is intermediate to the two. Demographic factors, other than coalescence, may influence allele length variance in a population and include the presence of ancient population bottlenecks, population size expansions, migration, population subdivision, selective sweeps or genetic hitchhiking and a high variance in reproductive success *(Hammer et al. 1997).*

Not surprising, considering the large number of alleles at the locus, the variance at locus F7 is the greatest in all populations. At this locus the highest variance is detected in the Turkish population, presumably due to the presence of the smaller 152 bp allele which is absent in the African population. This allele is also present in Ireland but at much reduced frequency.

In addition to higher levels of allele length variances, higher gene diversity levels were detected in the African population. In the African population the genetic diversity was calculated as 0.602, whereas the diversity in the Turkish population was 0.500 and the diversity averaged across all provincial Irish populations was 0.435. Greater levels of heterozygosity in African populations have been shown before *(Bowcock et al. 1994, Relethford and Harpending 1994, Deka et al. 1995, Jorde et al. 1995, 1997, Tishkoff et al. 1996, Relethford 1997).* In particular greater diversity has been detected in African populations at two X linked STR loci located on Xp22 *(Scozzari et al. 1997).* At locus DXS8175 diversity measures were found to range between 0.682 – 0.811 in West African populations, they were lower in Northern European populations (0.685 and 0.668 in Britain and Denmark respectively) and lowest in the Turkish population (0.543). At locus DXS8174 diversity levels were more similar between African and European populations although the greatest diversity was found in the African Daba population. In general the highest heterozygosities and variances were found in Sub-Saharan Africans.

An interpretation is that this relatively greater level of gene diversity, or heterozygosity, is an indication of an older origin for African populations compared to other global populations. Recently, however, it has been argued that the higher diversity detected in African populations may be indicative of a larger effective population size in Africa than elsewhere prior to or immediately following the split of African and non-African populations *(Rogers and Jorde 1995, Relethford and Harpending 1994,*

Relethford 1995, Harpending et al. 1996). Closely related populations are not expected to exhibit any significant differences in their heterozygosity measures, and indeed this is the case for the four Irish subpopulations.

An assessment of phylogenetic diversity was performed using a comparison of mismatch distributions. Mismatch distributions have been utilised for inference of population expansions in various populations using mtDNA sequence data and are illustrations of haplotypic diversity and not simply the diversity at individual loci. Distributions displaying a smooth unimodal curve are consistent with recent population expansions whereas rough multimodal distributions have been shown, by simulation, to be consistent with long periods of constant population size *(Rogers and Harpending 1992, Rogers 1995, Rogers et al. 1996, Harpending et al 1998, Reich and Goldstein 1998).* Mismatch distributions may also be employed to estimate the time depth of expansions and to estimate the effective size of the population immediately prior to expansion. In effect, the shape and properties of a distribution can be used to infer something of a population's past history and by direct comparison can infer relative diversities between populations.

There has been some debate surrounding the interpretation of such pairwise difference distributions *(Watson et al 1997, Bandelt and Forster 1997)*, however, the consensus interpretation from mtDNA studies suggests that the patterns of global mismatch distributions result from a Paleolithic population expansion ~60,000 *YBP* and it is calculated that the effective population size of the population immediately preceding the expansion was ~10,000 individuals *(Rogers and Harpending 1992, Sherry et al 1994, Rogers 1995, Harpending et al 1998).*

By analogy to mtDNA studies, the mismatch distributions of the African, Turkish and Irish populations have been compared and relative assumptions about their demographic histories can be made. The Irish and Turkish mismatch distributions are bimodal. The African distribution is somewhat more unimodal with a slight tail. The unimodality of the African distribution is suggestive of a population that has undergone an expansion in the past, whereas the bimodal distributions of the Irish and Turkish populations might indicate that these populations have not undergone similar expansions. This, however, may only be true if the haplotypes observed within each population are presumed to have arisen from the same chromosomal lineage. Any migratory influence incorporating further lineages into a population may result in bimodal peaks if the

haplotypes associated with recently integrated lineages differ significantly from the original lineage.

A more ancient distribution is that which displays a modal point more distal from the vertical axis of the mismatch chart. In a comparison of the three D_{swasm} distributions, the modal point of the African population is 2.5 and 4 times greater than that of the Turkish and Irish populations respectively. Additionally the mean number of pairwise mismatches is 1.5 and 1.8 times greater in the African population than in the Turkish and Irish populations. Assuming that most of the diversity in the Turkish population has resulted from a population expansion in the Upper Paleolithic, ~50,000 *YBP (Comas et al. 1996),* and that the diversity in Ireland and the rest of Europe has arisen since the Neolithic, ~10,000 *YBP (Cavalli-Sforza 1994, Chikhi et al. 1998),* then simple projections of these figures estimate from the modal values that the coalescence time for the African population falls somewhere between 40,000 - 125,000 *YBP* and from the mean values the coalescence time is estimated to be between 150,000 - 180,000 *YBP*. Although these are crude estimates, they fit well with previous data relating to global human population history.

Patterns of genetic diversity in each population have also been assessed by the construction of maximum parsimony networks relating haplotypes within populations. By employing a consensus approach to the construction of the networks, maximum consensus phylogenies have been created. From the patterns of the networks inferences may be made about the histories of the populations, particularly by making comparisons between the different population groups. In all of the Irish networks common core haplotypes are topologically central in the network and generally form tight clusters of more common closely related haplotypes. Most common haplotypes are likely to be ancestral *(Watterson and Guess 1977)* and here, the most frequent haplotypes lie at the core of the networks with less common haplotypes radiating from them in a star-like fashion. Star-like patterns in phylogenies occur with rapid growth and suggest expansion from the ancestral haplotype *(Castelloe and Templeton 1994).*

The lengths of the branches connecting each haplotype node are proportional to the number of mutations separating them. Although there are a few outlier haplotypes on the Irish networks the majority of the haplotypes lie in close proximity on the network suggesting that they are all recent haplotype derivatives. The African and Turkish networks, however, lack pronounced star-like patterns and phylogenetic density and the

haplotypes are widely distributed in the network. These patterns of greater diversity are suggestive of an older age for both the Turkish and African populations. In summation, the phylogenetic topologies and pairwise mismatch distributions suggest that most of the X chromosome diversity in Ireland has arisen as a result of rapid population growth, and by comparison to other populations, this expansion was recent. X STRs suggest that the African population is the most ancient population.

2.4.3. Evidence for multiple X chromosome lineages

Phylogenetic analysis illustrates a topology consistent with a population expansion in Ireland and Turkey. Hence, the bimodality of their mismatch distributions most likely indicates the presence of more than one distantly related X chromosome lineage which, as a result of gene flow and population admixture (which is more evident in Europe than Africa), have become integrated into the same populations. Phylogenetic information identifies haplotypes associated with the 152 bp allele at F7 and the 105 bp allele at F13 as being those of a separate chromosomal lineage. There are six haplotypes associated with the 152 bp allele (ht 6, ht 24, ht 25, ht 26, ht 27, ht 28 and ht 29) and two associated with the 105 bp allele (ht 12 and ht 91). These two divergent alleles are not found together in any sampled haplotypes suggesting that perhaps there may be three distinct lineages present in the Irish and Turkish populations. For example, one lineage may contain all haplotypes containing the 152 bp allele, another may consist of haplotypes containing the 105 bp allele, the main lineage being that containing all of the other closely related haplotypes. Additionally the haplotype associated with the 156 bp allele at locus F7 (ht 7) may represent a second African X chromosomal lineage.

A stepwise model for microsatellite mutation allows that the introduction of new lineages may be the most parsimonious explanation for the presence of bimodal patterns in the pairwise difference distributions of these populations. However, it is possible that the divergent allelic variants detected at these loci may be resulting from a departure from the stepwise mutation model for STR loci and could have been generated from large mutational deletions or insertions. If this were the case then they may not in fact be associated with distinct chromosomal lineages and may instead be closely related to the main core of haplotypes. However, in the D_{asm} mismatch distributions the tail, evident in the D_{swasm} mismatches, is not present. This might suggest that a large deletion/insertion

mutational event is responsible for the phylogenetic patterns observed here, rather than resulting from the accumulation of smaller mutations over time within diverse and differentiated lineages, although this cannot be certain.

One criticism relating to the use of microsatellite haplotypes in mismatch analysis pertains to the fact that most mismatch analysis models, based on mtDNA sequence data, assume an infinite sites model of mutation and suggest that mutational processes may affect the mean pairwise difference of a mismatch and also the number of segregating sites *(Bertorelle and Slatkin 1995, Aris-Brosou and Excoffier 1996).* The forward-backward stepwise mutational mechanism of microsatellite repeats may render them unsuitable, therefore, for this type of demographic analysis. The employment of the D_{asm} distance metric in the construction of pairwise difference distributions is unlikely to reflect the true nature of microsatellite mutation processes as it detects only whether there is or is not a difference between haplotypes and does not account for the scale of that difference. By employing the D_{swasm} metric, pairwise difference distributions are more likely to reflect the real demographic properties of a population and reflect haplotypic diversity rather than simply diversity at different loci. Despite the availability of microsatellites on the non-pseudoautosomal regions of the sex chromosomes, until now investigations of mismatch distributions of microsatellites in real populations have not been recorded. A simulation study of expanding unlinked autosomal microsatellite lineages, performed by Shriver *et al. (1997)* is in agreement with mtDNA studies and has detected expansions in simulated populations. Although the mismatches reflect the population structure associated with alleles as opposed to haplotypes, Shriver *et al. (1997)* have illustrated that the mutational mechanism of microsatellites does not hamper the determination of population dynamics using STR loci.

2.5 Summary

Previous studies of human genetic diversity have detected a greater diversity in African populations than in European and Asian populations. It is evident that X chromosome STR loci are powerful tools for the detection of genetic diversity within populations. Not only do they detect the ancient split between African populations and non-African populations but also they are effective in the detection of a greater Togolese diversity compared to European populations, illustrative of a greater antiquity for African populations. The date of divergence of Africans from non-Africans is in accordance with the model for the emergence of modern humans in Africa sometime between 100,000 - 200,000 *YBP*.

The proposed existence of at least three X chromosomal lineages in the Turkish and Irish populations and at least two lineages in the African population requires testing. Apart from the stepwise nature of microsatellite mutation the convergent nature of microsatellites means that two identical haplotypes may be identical by state but may be derived from different common ancestors. Therefore it is necessary that slower evolving mutations linked to the microsatellite loci be used to anchor the haplotypes with the assumption that they have arisen only once in the history of human populations. This will be discussed in more detail in **chapters 3 and 4.**

Chapter 3:

Y chromosome unique event polymorphisms

3.1 Introduction

The Y chromosome is paternally inherited in a haploid manner. Apart from the two small pseudoautosomal regions, the Y chromosome does not undergo recombination. The non-recombining portion of the Y chromosome therefore retains a record of mutational events that occurred in previous generations. The evolutionary history of male-specific lineages can, in this way, be traced. All extant Y chromosomes have a single paternal ancestor just as mtDNA sequences have a single female ancestor. A number of estimates for the time to the most recent common Y chromosome ancestor have been made, ranging from 37,000 - 49,000 *YBP (Whitfield et al 1995)* to ~170,000 *YBP (Underhill et al 1997),* ~150,000 *YBP (Hammer et al. 1998)* and ~188,000 *YBP (Hammer 1995).*

The first DNA polymorphisms on the Y chromosome were detected almost 15 years ago *(Casanova et al 1985, Lucotte and Ngo 1985*). Initial investigation of polymorphisms, however, yielded little variation between chromosomes indicating that the Y chromosome has the lowest nucleotide diversity of all chromosomes *(Spurdle and Jenkins 1992, Oakey and Tyler-Smith 1990).* Estimates of DNA sequence diversity have been found to range from < 1/18,000 nucleotides *(Jakubiczka et al 1989)* to < 1/46,515 nucleotides *(Malaspina et al 1990).* The reasons for the apparent lack of variation are uncertain although the small effective population size compared to autosomes, the lack of recombination and the possible effects of selection resulting in chromosomal sweeps may be contributing factors.

More recently a large number of polymorphic loci have been identified, which include both RFLPs *(Underhill et al 1997, Seielstad et al 1994, Hammer 1995, Whitfield et al. 1995, Kwok et al 1996, Hurles et al. 1998, Mathias et al 1994, Hammer et al 1997, Veitia et al. 1997 Zerjal et al 1997)* and STRs *(see de Knijff et al 1997 and Kayser et al 1997).* For example, Underhill *et al. (1997)* detected ~20 new polymorphisms on the Y chromosome using the denaturing high-performance liquid chromatography technique. Genealogical analysis of these mutations indicated that the most ancestral Y chromosome is found in African populations only. They estimated that a one nucleotide site difference between two randomly chosen Y chromosomes is expected every 3.2×10^3 - 3.8×10^3

nucleotides. This is in contrast to the initial low estimates of Y chromosome diversity *(Dorit et al. 1995, Whitfield et al 1995).*

3.1.1. The highly polymorphic p49a,f system

The first highly polymorphic Y chromosome locus detected was the p49a,f (DYS1) RFLP locus *(Ngo et al 1986).* Two sub-clones of cosmid 49, probes p49f and p49a, hybridise to DYS1 and detect up to 18 Y specific restriction fragments (A - R) corresponding to low copy number sequence when digested with *Taq*I. At least 8 of the fragment bands (A, B, C, D, F, G, H and I) have been shown to be present, absent or variable in length and are considered to be independent of each other *(Ngo et al. 1986, Torroni et al 1990).* However, coinheritance of alleles from the A, D and F series of bands has been detected *(Spurdle and Jenkins 1992).* One hundred and forty-four haplotypes have been observed in combinations of alleles from the polymorphic bands *(Poloni et al. 1997).* Initially it was thought that the *Taq*I polymorphisms resulted from point mutations within *Taq*I restriction sites only *(Ngo et al 1986)*, but polymorphisms have also been detected with *Pvu*II, *Sst*I, *Bgl*II, *Hind*III and *Pst*I *(Spurdle and Jenkins 1993).* Recently, the DYS1 locus has been found to correspond to the DAZ (Deleted in Azoospermia) gene cluster which would have been transposed to the Y chromosome from an homologous gene on chromosome 3 *(Saxena et al 1996).* The DAZ gene is organised in tandem repeats of 2.4 kb segments. It may be that the polymorphisms at the DYS1 locus result either from point mutations within the 2.4 kb segments leading to fragmentation of the band series, variation in the number of tandem repeats or sequence divergence of the same repeats on different DAZ gene copies *(Poloni et al. 1997).*

Variation at the p49a,f locus has been used in a number of population studies to infer male evolutionary history *(Torroni et al 1990, 1994, Persichetti et al 1992, Spurdle and Jenkins 1992, Semino et al 1996).* Results from this system concord with results from conventional genetic markers, nuclear DNA and mtDNA. For example, genetic distance measures based on p49a / *Taq*I haplotype frequencies have detected the most ancient split in human populations as that between African and non-African populations *(Spurdle and Jenkins 1992).* Contrary to other markers, however, a lower within population diversity has been detected in Africans compared to non-Africans. This locus has also been used to establish a genealogy of the Y chromosome. By the assessment of loss of *Taq*I sites by

CpG methylation, the predicted ancestral haplotype has been found in African Pygmy populations.

Some of the haplotypes have been found to be either qualitatively or quantitatively population specific. For example, haplotype XV is found mainly in Europeans or those populations that have had contact with Europeans *(Lucotte et al 1990, Spurdle and Jenkins 1992, Torroni et al 1994).* A cline of haplotype XV frequencies has been detected within Europe, increasing in frequency from the Near East to populations in the northwest of the continent *(Semino et al 1996).* Haplotype XII has also been shown to be common in European populations, whereas haplotype XIV has been found to be most common in Indian populations *(Lucotte et al 1990).*

Although the highly polymorphic nature of this locus makes it advantageous for population studies, the assay requires the use of the Southern blot technique and high molecular weight DNA samples. Additionally, the complexity of the haplotypes means that direct deductions about relatedness are difficult to make. More recently, PCR-based assays have been developed for detection of variation at other Y chromosome polymorphic loci.

3.1.2. A hierarchical approach

Different genetic systems detect different facets of the past and it is beneficial to examine a variety of different systems in order to extract all of the information inherent in a sample. A hierarchical approach to the assessment of population diversity and history with genetic markers has been suggested *(Jobling and Tyler Smith 1995, Santos and Tyler Smith 1996, de Knijff et al 1997)* and has recently been applied in a number of informative studies *(Thomas et al 1998, Zerjal et al 1997, Hurles et al 1998, Underhill et al 1996, Jobling et al 1996, Jobling et al 1998).* Initially slowly evolving unique event polymorphisms, believed to have arisen only once in evolutionary history due to their slow mutation rate, are used to subdivide the sample into discrete chromosomal lineages. Following initial subdivision, the diversity within the defined lineages is assayed by employing a system with greater sensitivity using faster evolving loci such as microsatellites or minisatellites. Y chromosome STRs are discussed in **chapter 4.**

3.1.3. Compound haplotypes

Multiple Y chromosome polymorphisms, representing unique mutational events, can be combined to create compound haplotypes. Compound haplotypes can be used both for tracing Y chromosome evolution and for the detection of geographical patterns in their distribution. Most compound haplotypes have been found to be highly population specific. In a survey of long-range DNA polymorphisms on the Y chromosome compound haplotypes comprised from twelve polymorphisms were used to distinguish between all chromosomes within a population of 66 Y chromosomes *(Jobling 1994).* These haplotypes supported the existence of at least two major haplotype lineages and provided evidence for a third Y specific lineage, all of which were found to be population specific. Twenty-three unique event markers are now available for phylogeny reconstruction, ten of which have been developed for PCR *(Mark Jobling personal communication).* In this study nine unique event polymorphisms have been used to construct compound haplotypes which separate all Y chromosomes into 11 discrete haplogroup lineages.

3.1.4. Unique event polymorphisms

The most frequently employed Y chromosome unique event polymorphism in population genetic studies is the Y Alu polymorphic (YAP) element. The recent insertion of an Alu element on the long arm of the Y chromosome, Yq11, defines a simple and stable polymorphism. There is no evidence for the deletion of Alu elements and thus it is thought that such polymorphisms are stable over millions of years *(Sawada et al. 1985).* These short interspersed repeats are believed to have spread throughout primate genomes by a mechanism of transposition involving an RNA intermediate. The YAP element is a member of the polymorphic subfamily – 3 (PSF-3) family of Alu elements *(Hammer 1994).*

A few human Alu elements are polymorphic, defined by either the presence or absence of the element at a specific chromosomal site. Such polymorphic Alu elements are the most recently inserted. The YAP element is believed to be a recent insertion occurring after the divergence of humans from the great apes as the YAP element was absent at the homologous site in samples from 14 chimpanzees and one gorilla *(Hammer 1994).* The geographic distribution of the YAP element insertion polymorphism has been

widely used to study male human population history. One of the first global studies of YAP insertion frequencies was carried out in 340 individuals from Europe, Africa, the Far East and Oceania *(Hammer 1994).* The insertion was present on 31% of global chromosomes (48% using combined data from Persichetti *et al. (1992)*, Spurdle *et al. (1994)*). The highest frequency of the insertion was found in sub-Saharan African populations, particularly in South African Bantu-speakers and West Africans (78-86%). Frequencies were much lower in Europe (4-11%), intermediate in other African populations and absent in most of the Asian and Oceanasian populations. However, the YAP insertion element was detected in 24% of Japanese chromosomes surveyed. The geographic distribution of the YAP insertion makes it a useful tool for studying past population history.

More than three unique event polymorphisms at the sex-determining gene region have been detected *(Whitfield et al 1995, Kwok et al. 1996, Veitia et al. 1997),* the polymorphisms being found to be population specific. The sex-determining locus (SRY) has been mapped to approximately 5 kb of the short arm of the Y chromosome immediately proximal to the Yp pseudoautosomal boundary and contains a single 612 bp exon. Although it is generally agreed that this locus is necessary for the sex-determining function of the Y chromosome, it is possible that other genes located proximal to the 5 kb region are also involved *(Whitfield et al 1995).* Mutations in the SRY gene lead to failure in testis development. The SRY-2627 polymorphism was detected by screening for mutations in the SRY open reading frame and 3.8 kb of flanking sequence in patients presenting with 46,XY partial or complete gonadal dysgenesis *(Veitia et al 1997).* Two point mutations were detected in a 2.9 kb sequence located immediately 5' to the SRY open reading frame at positions -1532 and -2627. The SRY-1532 polymorphism involved an A $\rightarrow$ G transition. Analysis of chimpanzee DNA indicated that the A allele is the ancestral state *(Whitfield et al. 1995).* The SRY-2627 polymorphism involved a C $\rightarrow$ T transition within a 12 bp tandem duplication sequence, the C allele being ancestral. The SRY-1532 polymorphism was detected in 4 of 55 normal Caucasian males and more frequently in an isolated Indian population (17 / 55 normal individuals). In a population of mixed ethnic origin the frequency of the polymorphism was 3 in 43. The SRY-2627 polymorphism was detected in one of 55 normal individuals of Caucasian origin but was absent in both the Indian population and the population of mixed ethnic origin.

Compound haplotypes employing four markers within the YAP region, three markers within the SRY gene region and the 92R7 polymorphism *(M.E. Hurles and C. Tyler-Smith, unpublished)* were found to delimit ten distinct Y chromosome haplotypes unevenly distributed among human populations *(Hammer et al 1998).* The ancestral haplotype was found in African populations only, African populations also having the highest level of genetic diversity than other global populations. A nested cladistic analysis determined a number of global migrational events, the first being an initial range expansion out of Africa into the rest of the globe with the replacement of local populations. A secondary expansion from Asia resulted in Y chromosomes returning to Africa without replacement of the African genepool (see **section 3.4)**.

Two biallelic markers have been found to distinguish between Y chromosomes with the YAP insertion polymorphism. These include the SRY-8299 marker involving a G → A mutation at the SRY gene locus *(Whitfield et al. 1995)* and the sY81 marker involving an A → G transition at position 168 at the DYS271 locus *(Seielstad et al 1994).* The sY81 mutation involves the loss of an *Nla*III restriction site. The G allele was initially detected in 41% of Africans but on only one other single chromosome of Mayan origin. The almost exclusivity of the G allele in African populations suggests that this mutation arose within Africa after the introduction of YAP+ chromosomes (for a more detailed discussion see **section 3.4**).

The M9 polymorphism involving a C → G transversion was detected by Underhill *et al. (1997)* and defines a previously undetected male lineage. This lineage is prevalent in Eurasia but has not been found in any African populations except those that have undergone recent admixture with Europeans. The exclusion of this polymorphism from Africa has suggested that the mutation arose in a population outside Africa after the initial separation of Africans from non-Africans, although it is possible that drift has removed any traces of this lineage within African populations over time.

A T → C transition within the single-copy locus RBF5 on proximal Yq has been recently identified by Zerjal *et al. (1997).* The geographic distribution of the alleles at this locus (Tat) has determined that the C allele is found only in a subset of Asian and North European populations. The C allele was found to be absent in Africans, Southern Europeans, Oceanasians and Americans, but it was present in Asian and Northern European populations. The highest frequency of the C allele was detected in the Yakut

population of Siberia (86%) and it has been suggested that this mutation arose in Asia, most likely in the China / Mongolia region.

The LLY22g locus also demonstrates alleles, resulting from a C → A transversion, that are unevenly distributed in some populations *(E. Righetti and C. Tyler-Smith, unpublished).* Thirty-seven of 58 males with a deletion at the 50f2/C locus have the LLY22g T allele *(Jobling et al. 1996).* Those chromosomes with the T allele are relatively common and are geographically widespread, found in Siberian, Chinese, Japanese, Indian and Scandanavian populations. There is a high frequency of this allele in the Finnish population (55% 50f2/C deletion) suggesting a Northeast Asian influence. This chromosome type was also found in 3 of 9 Saami individuals.

Many other unique event polymorphisms (not analysed in this study) on the Y chromosome have also been found to be population specific. For example, a C → T transition was detected on the Y chromosome by comparative sequencing of Y specific sequence-tagged sites by denaturing high-performance liquid chromatography *(Underhill et al. 1996).* The T allele was found exclusively in 90.5% of South and Central American populations and was not detected in populations outside the Americas. The T allele was also found in North American Navajo and Eskimo populations at frequencies of 50% and 67% respectively. The distribution of the T allele suggest that the mutation might have arisen in a North American population, although it is possible that it arose in a Northeast Asian population prior to migrations to the American continent. Recent analysis of Y chromosome polymorphisms have confirmed the hypothesis that native Americans originated in a Northeast Asian population although which ancestral population and the timing of migrations remains controversial. In a study of the Chinese population, the C allele was detected in 100% of chromosomes, which excludes the possibility that a Chinese population was ancestral to native Americans unless the C → T mutation occurred very recently *(Wang et al 1999).* It has been suggested that the most likely ancestral source population originated in Siberia and may have been closely related to the Kets or Altaians *(Santos et al. 1999),* although it has been suggested that the presence of the Y specific 1T haplotype found at high frequencies in Native Americans and Siberians may be the result of a back migration from the Americas into Asia *(Karafet et al. 1997).* Additionally, a recent study of 12 biallelic polymorphisms defining 14 discrete haplogroup lineages suggests that there may be more than one original American founder haplotype in

addition to a number of haplotypes originating from recent admixture with non-American populations *(Karafet et al. 1999).*

The 12f2 locus, mapping to interval Yq11.1 - Yq11.22, also demonstrates polymorphism within global male Y chromosomes *(Casanova et al. 1985).* The 12f2 polymorphism results from a deletion in a 10 kb segment of DNA resulting in a smaller 8 kb fragment. The 8 kb fragment is found at relatively high frequencies in Caucasian populations but is absent in sub-Saharan African populations and other non-Caucasian populations. It has been shown, therefore, to be a Caucasoid specific allele *(Mitchell et al 1997),* decreasing in frequency from the Near East to Northwest European populations *(Semino et al 1996).* The 12f2 8 kb allele is a good marker, therefore, for determining Caucasian admixture in non-Caucasian populations. For example, 25.4% of the Ethiopian population have the 8 kb allele *(Passarino et al. 1998)* suggesting a considerable Caucasian paternal influence in that population. Recent analysis of the distribution of the 12f2 8 kb allele within Europe suggests that it may have been spread by sea-farers along the Mediterranean, rather than by Neolithics into continental Europe *(Mitchell et al. 1997).*

3.1.5. Y chromosome analysis to complement maternal history.

Population genetic studies of the Y chromosome have been used recently to substantiate predictions made by analysis of maternally inherited mtDNA. MtDNA data defines female lineages and elucidates facets of female migrational history. Y chromosome and mtDNA studies may coincide or be contradictory. For example, Sajantila *et al. (1996)* used mtDNA sequence analysis with Y chromosome data to establish the origin of the Finnish population and to establish a date for the putative bottleneck during the establishment of the Finnish population. Conversely, Passarino *et al. (1998)* have shown by parallel analysis of mtDNA and Y chromosome polymorphisms that the Ethiopian population, although containing mainly African mtDNA lineages, contains 25.4% Caucasian male lineages and that admixture levels between Africans and Caucasians on the Y chromosome reach up to 57.7%. Additionally, Hurles *et al (1998)* have shown that mtDNA and Y chromosome studies establish different origins for male and female lineages in Polynesian peoples; mtDNA lineages coming from native Polynesian ancestors and at least one third of Polynesian Y chromosomes stemming from a European origin. Conflicting mtDNA and Y chromosome origins have also been

determined in Arab tribal groups *(Salem et al 1996),* Finns *(Zerjal et al 1997)* and Basques *(Poloni et al 1997).*

3.1.6. Surnames

Surnames are paternally inherited in the same manner as Y chromosomes, although in cases of non-paternity some anomalies may arise. The hereditary use of surnames has been active in Ireland since *c* 950 AD and is thought to exist as one of the oldest applications of the surname system. The surname system in Ireland was not introduced, but rather it is thought that locative names, toponymics and nicknames were adopted *(MacLysaght 1997).* Post-marriage residence in Ireland is generally patrilocal suggesting that the name derived from the eponymous ancestor has remained in the homeland of the associated clan for generations. Surnames, except in the infrequent case of non-paternity, are therefore an indication of family history and, on a larger scale, of population history *(Zei et al. 1993, Guglielmino et al. 1998).* In this study, surnames have been used to subdivide the Irish population into components of known historic and prehistoric (Gaelic) origin, to which 1,000 year old geographical information may also be assigned.

The statistical correlation between surnames and genetics has been estimated before as 0.66 from a large Sardinian data set. Surnames were found to be more highly correlated with language (0.89) than genetics, however, and less with geography. Importantly, geography and genetics demonstrated a lower correlation than surnames did with genetics *(Cavalli-Sforza et al 1994).*

	Surnames	Genetics	Geography
Languages	0.89	0.68	0.76
Surnames		0.66	0.65
Genetics			0.64

Table 3.1: Congruence coefficients between linguistic, surname, genetic and geographic distances. From Cavalli-Sforza *et al (1994)*, pg 275.

3.2 Materials and methods

3.2.1. Population samples and subdivision based on surname origins.

Samples were collected with informed consent as for the X chromosome study. For individuals sampled in Ireland, additional information regarding the subject's surname was collected.

Irish samples were classified into population cohorts according to the origin of their surname. Eight population cohorts were identified in the Irish sample and were as follows: Ulster, Munster, Leinster, Connaught, England, Scotland, Norman and Norse. As only a small number of Norse samples were identified, they were subsequently grouped with Norman samples. Names originating in the four provinces of Ireland are of prehistoric Gaelic origin and have been associated with local septs and clans since the adoption of the surname system in the 10th century AD. Names sampled in Ireland which have English, Scottish, Norman and Norse origins are relics of more recent historic influences ranging from the 9th century to the 16th and 17th centuries AD.

Population classification according to surname origin was determined by reference to MacLysaght *(1997).* If names were hyphenated then the last name was taken. If there was more than one possible origin for the name then the region that correlated with information about paternal family history was used. If a common origin and a less common origin for a name were given then it was assumed that the name came from the more common of the two. As far as possible, the counties from where it is thought the names originated were taken and not the counties from where the names are found most commonly today. Where the name did not appear in MacLysaght *(1997)* but a phonetically similar name appeared then that was used, for example Greavy = MacGreavy = MagRiabhaigh and Herron = Heron.

Connaught surnames	Ulster surnames	Leinster surnames	Munster surnames	English surnames	Norman surnames	Norse surname	Scottish surnames	Asian surnames
Boyle	Bonner	Banville	Bohannon	Bagnall	Barry	Doyle	Beattie	Chan
Brogan	Colton	Black-Macken	Callanan	Ball	Bourke		Birrane	Ho
Colleran	Coyle	Breen	Carney	Berry	Bryan		Boyd	Mar
Conroy	Devine	Brophy	Cronin	Brady	Fitzgerald		Craigie	Mohd Zim
Conway	Donnelly	Byrne	Curry	Clarke	Harper		Creighton	Varadakar
Culkin	Gallagher	Carroll	Dunlea	Collier	Henderson		Ferguson	Zainuddin
Curran	Gilsenan	Conlan	Dwyer	Field	Lynch		Grant	
Devers	Hegarty	Ennis	Geaney	Forde	MacNicholas		Hamilton	
Dolan	Hussey	Gargan	Hallinan	Fox	Marshall		Johnston	
Duggan	Kee	Geoghegan	Hannon	Garrett	Morrissey		Knox	
Egan	Lynn	Hallahan	Harrington	Gorey	Roche		Latimer	
Fallon	Maguire	Heaton	Hearne	Harris	Wall		MacKenzie	
Flynn	McCafferty	Kavanagh	Herlihy	Harrison			Orr	
Foody	McNelis	Kinsella	Hickey	Hughes			Shaw	
Gannon	Mullen	McEvoy	Hogan	Jacob			Sheehy	
Ginty	Mulligan	Molloy	Kearney	Kent			Simpson	
Greavey	O'Reilly	Nolan	Kelleher	King				
Hynes	Quinn	Nowlan	Kennedy	Langford				
Kelly	Rooney	Phelan	Malone	Leonard				
Kenny	Tully	Quinlan	McCarthy	Mills				
Killeen		Rafter	Meagher	Mitchel				
Killian		Rickard	O'Brien	Moore				
Loftus		Sheane	O'Connell	Robson				
Lyons		Sheridan	O'Donoghue	Rowan				
MacDermott		Synott	O'Grady	Salisbury				
Madden		Searle	O'Keefe					
Mahon		Smith	O'Meara					
McHale		Smyth	O'Regan					
McKenny		Spencer	Regan					
McHugh		Stokes	Ryan					
McKeown		Tibbits	Salter-Townshend					
Munnelly		Tucker	Stuart-Dinneen					
O'Boyle		Wagstaff	Sweeney					
O'Hara		Warner						
O'Malley		Weld-Moore						
Pull								
Quigley								
Reape								
Ruane								
Travers								
Treacy								
Tuffy								
Tunney								

Table 3.2: Surnames sampled in Ireland listed in associated population cohorts.

Figure 3.1: Approximate locative origins of surnames sampled in this work. Locations were identified from the Heraldic scroll and map of family names and origins of Ireland, published by Mullins of Dublin (Dublin, Ireland). Those names that were sampled but not found on the reference map are not plotted. On visual inspection, a representative sample from all geographic regions of the country has been used.

Two hundred and twenty-one Y chromosomes, 146 of which were of Gaelic origin, were sampled in Ireland and fully genotyped. The Irish samples were subdivided as follows:

Population	n
Ulster	22
Munster	37
Leinster	30
Connaught	57
England	40
Scotland	17
Norman/Norse	18

An additional 105 non-Irish Y chromosomes were fully genotyped.

Population	n
Denmark	12
Togo	37
Turkey	50
India	1
Hong Kong	1
China	2
Malaysia	2

In total 326 Y chromosomes were fully genotyped.

3.2.2. Unique event polymorphisms

Eight single nucleotide polymorphisms (SNPs) and 1 YAP insertion/deletion polymorphism, located on the non-recombining portion of the Y chromosome, were employed in this study. These polymorphisms identify 11 discrete haplogroups, 8 of which were detected here. Haplogroup genotypes are given in **table 3.3** and their relationships are illustrated in **figure 3.2.** All 326 Y chromosomes were assayed for the YAP, M9, 92R7 and SRY-1532 polymorphisms. By reference to the maximum parsimony tree **(figure 3.2)** developed by Mark Jobling from a wide data set *(personal communication)*, it was unnecessary to type all of the samples for all of the polymorphisms. For example, only YAP+ chromosomes were typed for the SRY-8299 and sY81 polymorphisms. Additionally, only hg 26 chromosomes were typed for the LLy22g and Tat polymorphisms, and similarly only hg 1 chromosomes were typed for the SRY-2627 polymorphism.

Hg	YAP	SRY-8299	92R7	SRY-1532	SRY-2627	Tat	sY81	M9	LLY22g
1	0	0	1	1	0	0	0	1	0
2	0	0	0	1	0	0	0	0	0
3	0	0	1	0	0	0	0	1	0
4	1	0	0	1	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0
8	1	1	0	1	0	0	1	0	0
12	0	0	0	1	0	0	0	1	1
16	0	0	0	1	0	1	0	1	1
21	1	1	0	1	0	0	0	0	0
22	0	0	1	1	1	0	0	1	0
26	0	0	0	1	0	0	0	1	0

Table 3.3: Haplogroup genotypes defined by 0 – ancestral state 1 – derived state. Only haplogroup 1, hg 2, hg 3, hg 8, hg 12, hg 16, hg 21 and hg 26 were detected in the populations surveyed here.

The convention for allelic states is 0 - ancestral, 1 - derived *(Jobling and Tyler-Smith 1995)*. The ancestral states were determined by comparison with non-human ape DNA sequences *(Mark Jobling, personal communication)*. The ancestral state is not known for the 92R7 or SRY-2627 polymorphisms. For 92R7 the T allele is taken as the derived state and for SRY-2627 the T allele is taken as the derived state. Control DNA samples and primer sets were kindly supplied by Mark Jobling *(Department of Genetics, University of Leicester, Leicester)*.

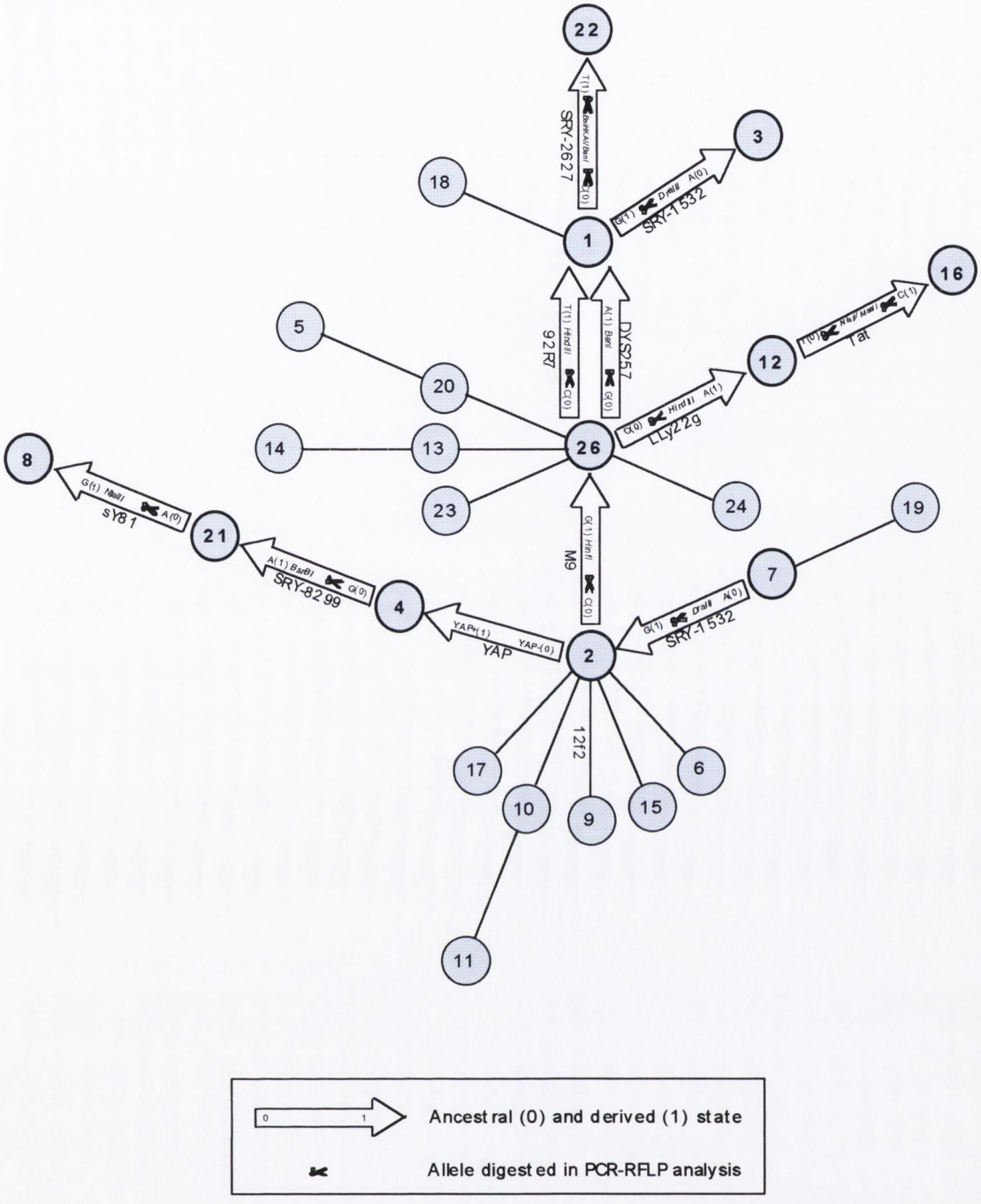


Figure 3.2: Unrooted maximum parsimony tree of Y chromosomal haplogroups. Haplogroups are present as nodes on the tree and are linked by branches representing each mutational event. The assays for haplogroups that have no defined mutational event leading to them have not yet been developed for PCR. Ancestral states for denoted polymorphisms were deduced by reference to non-human ape sequences *(Mark Jobling, personal communication).*

3.2.3. Design of a new primer set for SRY-2627 assay

PCR amplification by the primer set for SRY-2627 from Veitia *et al. (1997)* and supplied by Mark Jobling was made troublesome by degraded DNA samples. The large amplicon (1242 bp) could not be achieved with this primer pair. A new primer set was designed in order to achieve a smaller PCR amplicon which would amplify more easily in degraded DNA samples.

SRY-2627

Taatatatgcatctgggaggtctttttgccttcttaaaaacatatagatggttggacatatgtat
Ataagaatataaaattcaccactttatcttttgtgaatgtgtgctgtgaagaactcctttactggg
Gtgatggaaccagtggctacaaagtaaggagctggtttactgctgtaaagggttcgcggctttgaa
Tttcaagctctggttctgtgtccttgggcacctgcgcgtgaatcgttgccgcgaggctgggccaag
Ttaaggccccacgcagtttggcttccgggccaaggaagccccacagggtgccccacagggtgaagc
Cccatgccctacagggtgaagcggctgaagctggtagtggttccgaggaagcggtcaaagtcccgc
Tccagaggttcctcttcttggttgtcactcccggaaccccgccaggggtctggcttccccatcgac
Acctcctcctgttcagtcaccatcacgtatcccacgaggccggacggcactgccacctcgtcgccc
Cgtagactgcggccccgaaactacacttaagagtccctcggggccctggcggatggcaggcgtgat
Gaagcgccacaccgtggtgggccgcagggtatcaggtgtagtgtggcgggggcggcgtcctgcagt
Gtggctgggcgcaactggacgctgtacctctccatagccgcctagtcgccgctctccatcctcgct
Cctgtgcgacgccactttatccctcttaaactggacgattcagtagtaccggggaacgaaggcaac
Aagacctaaccatagcaacaacattttgtttttgaaacccatcttgcaactgcctgactgcaaaat
Gtacttgcagattctgcagacaaaaattgctgccacttattaaaagacttcatagtgaaaactgac
Tttttacccaaaccactgattcaaggacttttgcgtagaaactcatgaactgttaagataacccctt
Cattaccaattaagataataggttgtaagtaacctaatgagtaccaactttctcttatttccttta
Aaaactgtctctcaaaacagccggtcacgacacaatgtcagatgcttctggaatctgtttctctaa
Ttacagttctaaaactgccaacaaactccttatttggacttcagtttctctgactctttggttcac
Catgttgtgcagccatcacctctctctagttccagagcatattttatcaatcctcaaaagaaaccg
Tgcatccacatatgcttctgctatgt

Original primer pairs

New primer pairs

BanI restriction sites

Possible allele sizes resulting from *BanI* restriction digest:

C allele – 111bp, 87bp, 115bp

T allele – 111bp, 202bp

The SRY-2627 locus is located 2627 bp upstream from the SRY gene open reading frame. The sequence surrounding this polymorphic site was gained from the human genome database (accession no. CM 981849, *S. Whitfield)* and is shown above. The new PCR amplification primers amplify a product of 313 bp. Restriction enzyme cutting in the presence of the C allele results in 3 bands (111 bp, 87 bp and 115 bp). When the T allele is present at the site then only two restriction products result (111 bp and 202 bp). An internal *BanI* restriction site is present.

3.2.4. PCR assays for biallelic polymorphisms

Summaries of PCR and restriction enzyme digestion conditions for the restriction fragment length polymorphism (RFLP) assays are given in **table 3.4.** PCR amplification reactions were performed using components as in **table 2.4.** Amplification conditions were optimised for each primer set and these are given in **table 3.4.**

Restriction digest analysis was performed in 96-well microtiter plates in 40 µl final reaction volumes (20 µl PCR product and 20 µl digest mix). For each assay, 1U of the appropriate restriction enzyme was used. Corresponding restriction digest reagents and their quantities in each reaction were used according to specifications from New England BioLabs Inc. Restriction digests were carried out for > 2 hours at 37 °C.

Following digestion 6X electrophoresis tracking buffer was added to each sample. 10 µl of the digest product was then electrophoresed on a 1.5 % - 2.5 % agarose gel with 0.5 µg/ml ethidium bromide in 1X TBE buffer. Controls were run in each row of samples and the presence or absence of enzyme cutting was recorded for each sample. Photographs were made of all gels for later referral.

RFLP assay	Forward and reverse primers 5' - 3'	PCR cond.	$MgCl_2$	Restriction enzyme	Incubation temp.	Control DNA
YAP	cag ggg aag ata aag aaa ta act gct aaa agg gga tgg at	94C 30s 54C 30s 72C 30s 33 cycles	17.5 or 25	na	na	m19 0 m125 1
SRY-1532	tcc tta gca acc att aat ctg g aaa tag caa aaa atg aca caa ggc	94C 30s 59C 30s 72C 30s 34 cycles	15 or 20	*DraIII*	37C	m109 0 m125 1
SRY-8299	aca gca cat tag ctg gta tga c tct ctt tat ggc aag act tac g	94C 30s 62C 30s 72C 60s 33 cycles	25	*BsrBI*	37C	m121 0 m125 1
SRY-2627	gaa ctc ctt tac tgg ggt g gtg aca acc aag aag agg a	94C 30s 52C 30s 72C 120s 33 cycles	15	*BsiHKAI* *BanI*	65C	m35 0 m148 1
92R7	gac ccg ctg tag acc tga ct gcc tat cta ctt cag tga ttt ct	94C 30s 62C 30s 72C 60s 33 cycles	15	*HindIII*	37C	m19 0 m35 1
Tat	gac tct gag tgt aga ctt gtg a gaa ggt gcc gta aaa gtg tga a	94C 30s 60C 30s 72C 30s 33 cycles	15 or 20	*NlaIII*	37C	m148 0 m121 1
sY81	agg cac tgg tca gaa tga ag aat gga aaa tac agc tcc cc	94C 30s 60C 60s 72C 60s 32 cycles	10	*NlaIII*	37C	m19 0 m118 1
LLy22g	cca ccc agt ttt atg cat ttg ata gat ggc gtc ttc atg agt	94C 30s 55C 60s 72C 60s 33 cycles	15	*HindIII*	37C	m125 0 m121 1
M9	gca gca tat aaa act ttc agg aaa acc taa ctt tgc tca agc	94C 30s 58C 60s 72C 60s 33 cycles	20	*HinfI*	37C	m118 0 m35 1

Table 3.4: PCR forward and reverse primers (5' – 3') for Y chromosome UEP assays. PCR reaction conditions for the YAP assay and the SNP assays, restriction digest conditions for the SNP assays and the control DNA samples used as references for ancestral and derived states, supplied by Mark Jobling *(Department of Genetics, University of Leicester, Leicester).*

RFLP assay	Polymorphism	Restriction enzyme	Incubation temp.	Product size	Haplogroup and reference
YAP	Alu+ or Alu-	*na*	na	YAP+ 455bp YAP- 150bp	Alu+ defines hg 4 and subgroups hg 21 and hg 8 Hammer and Horai (1995)
SRY-1532	A -> G	*Dra* III	37C	A 167bp G 112bp and 55bp	A allele defines hg 3 (with 92R7 T allele and M9 C allele) or hg 7 (with 92R7 C allele and M9 G allele) Hg 3 is thus defined by a reversion of SRY-1532 Whitfield *et al.* (1995) and Kwok *et al.* (1996)
SRY-8299	G -> A	*BsrB* I	37C	A 509bp G 362bp and 147bp	G -> A defines hg 21, a sub-group of hg 4 (YAP+) Whitfield *et al.* (1995)
SRY-2627	C -> T	*Ban* I	37C	C 111bp, 87bp and 115bp T 212bp and 202bp	C -> T allele defines hg 22, a sub-group of hg 1 There is an internal control *BanI* site Veitia *et al.* (1997)
92R7	C -> T	*Hind* III	37C	C 709bp T 197bp and 512 bp	C -> T defines hg 1 and its subgroups Mathias *et al.* 1994 and M.E. Hurles and C. Tyler-Smith, unpublished
Tat	T -> C	*Nla* III	37C	T 112bp C 85bp and 27bp	T -> C defines hg 16 on hg 12 background (LLY22g A allele) Zerjal *et al.* (1997)
sY81	A -> G	*Nla* III	37C	A 209bp G 144bp and 65bp or 102bp, 65bp and 42 bp	A -> G defines hg 8, sub-group of hg 21 (YAP+, SRY-8299 A allele) Seielstad *et al.* (1994)
LLy22g	C -> A	*Hind* III	37C	C 500bp, 230bp and 120bp A 650bp, 500bp 230bp and 120bp	C -> A defines hg 12. The A allele is also present in hg 16. This is a repeated sequence, and two copies are amplified. One copy contains the polymorphism. E. Righetti and C. Tyler-Smith, unpublished
M9	C -> G	*Hinf* I	37C	G 248bp and 93bp C 182bp, 93bp and 66bp	C -> G defines hg 26 and its subgroups There is an internal *HinfI* control site Underhill *et al.* (1997)

Table 3.5: RFLP assays with conditions and references

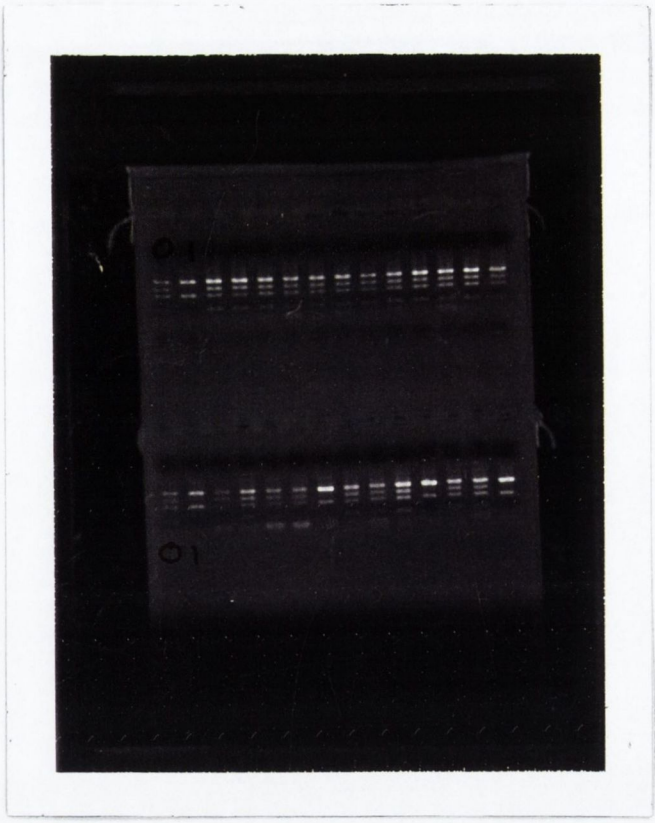


Figure 3.3: Gel photograph for RFLP assay 92R7. 0 is the ancestral state (C allele, hg 26) and 1 is the derived state (T allele, hg 1). The DNA samples are from the Turkish population.

3.2.5. New primer set for LLY22g

Due to degradation of the DNA samples over time, large amplicons were difficult to amplify successfully. Primers for a short amplicon at locus LLY22g were designed and kindly supplied by Tatiana Zerjal *(Oxford University, Oxford).*

The following primers amplify a product of ~210 bp.

LLY22g B2 5' ATA GAT GGC GTC TTC ATG AGT 3'

LLY22g S2 5' GAT GTT GGC CTT TAC AGC TC 3'

PCR cycling conditions were as follows:

30-38 cycles of 94 °C, 45 secs

55 °C, 30 secs

72 °C, 30 secs

$MgCl_2$ concentration: 7.5 mM

A C → A mutation at the LLY22g locus destroys a *Hind*III cutting site resulting in partial digestion in A allele samples and total digestion in C allele samples. Digestion with *Hind*III results in two products of 85 bp and 125 bp.

3.3 Results

3.3.1. Haplogroup frequencies

The 9 Y chromosome unique event polymorphisms (UEPs) detect 11 discrete paternally inherited haplogroups *(Jobling and Tyler-Smith 1995; Hurles et al. 1998, Kwok et al. 1996, Underhill et al. 1997, Hammer and Horai 1995, Whitfield et al. 1995, Seielstad et al. 1994, Mathias et al 1994; Veitia et al 1997, Zerjal et al 1997, M.E. Hurles & C. Tyler-Smith, unpublished).* Eight of the possible haplogroups (hg) were detected in the sample of 326 Y chromosomes of Caucasian and African origin. The frequencies of the haplogroups occurring in each population are given in **table 3.6.** Their distributions in each population are shown in **figures 3.4 a, 3.4 b and 3.4 c.**

For those chromosomes sampled in Ireland two frequency distributions were created. One distribution considers samples that were classified into population cohorts assigned by surname origin. The other distribution considers samples classified into cohorts assigned by using information on the individuals knowledge of their paternal origin (denoted 'geography'). Most of the analyses, henceforth, consider populations delimited by surname origin unless otherwise specified in the text. For example, when the English population is referred to, this relates to those individuals sampled in Ireland having a surname of English origin. This is also the case for the Scottish and Norman/Norse populations. The Irish population, unless otherwise stated in the text, refers to individuals sampled in Ireland that have a surname of Gaelic origin.

The haplogroup distributions of populations, classified according to surname origin and according to individual knowledge of paternal family history (geography), were compared. The distributions are broadly similar. However, when populations are classified according to surname origins, outlier haplogroups, assigned geographically, completely disappear in Munster and decrease in Ireland as a whole. In Leinster hg 16 disappears but a hg 26 chromosome appears in the sample. In general, the distributions become more coherent when populations are defined by surname origins.

When all of the chromosomes that were sampled in Ireland (including the English, Scottish, Norman/Norse samples of exogenous origin) were pooled as one population then 6 haplogroups were detected; hg 1, hg 2, hg 3, hg 16, hg 21 and hg 26.

Hg	Ireland (All) *(221)*	Ireland (Gaelic) *(146)*	Ulster *(22)*	Munster *(37)*	Leinster *(30)*	Connaught *(57)*	England *(40)*	Scotland *(17)*	Norman / Norse *(18)*	Denmark *(12)*	Turkey *(50)*	Togo *(37)*
1	0.781	0.897	0.818	0.946	0.733	0.982	0.630	0.529	0.830	0.417	0.080	...
2	0.170	0.089	0.182	0.054	0.200	0.018	0.330	0.353	0.056	0.333	0.580	...
3	0.059	...	...	...	...	...	...	0.059	...	0.167	0.020	...
4	...	...	...	...	...	...	...	...	...	...	...	...
7	...	...	...	...	...	...	...	...	...	...	...	...
8	...	...	...	...	...	...	...	...	...	...	...	0.838
12	...	...	...	...	...	...	...	...	...	...	0.020	...
16	0.008	...	...	...	...	...	...	...	0.056	...	...	...
21	0.028	0.007	...	...	0.033	...	0.050	0.059	0.056	0.083	0.220	0.162
22	...	...	...	...	...	...	...	...	...	...	...	...
26	0.005	0.007	...	...	0.033	...	...	...	...	...	0.080	...

Table 3.6: Frequencies of haplogroups in global populations. Samples collected in Ireland are separated into population cohorts by surname origin.

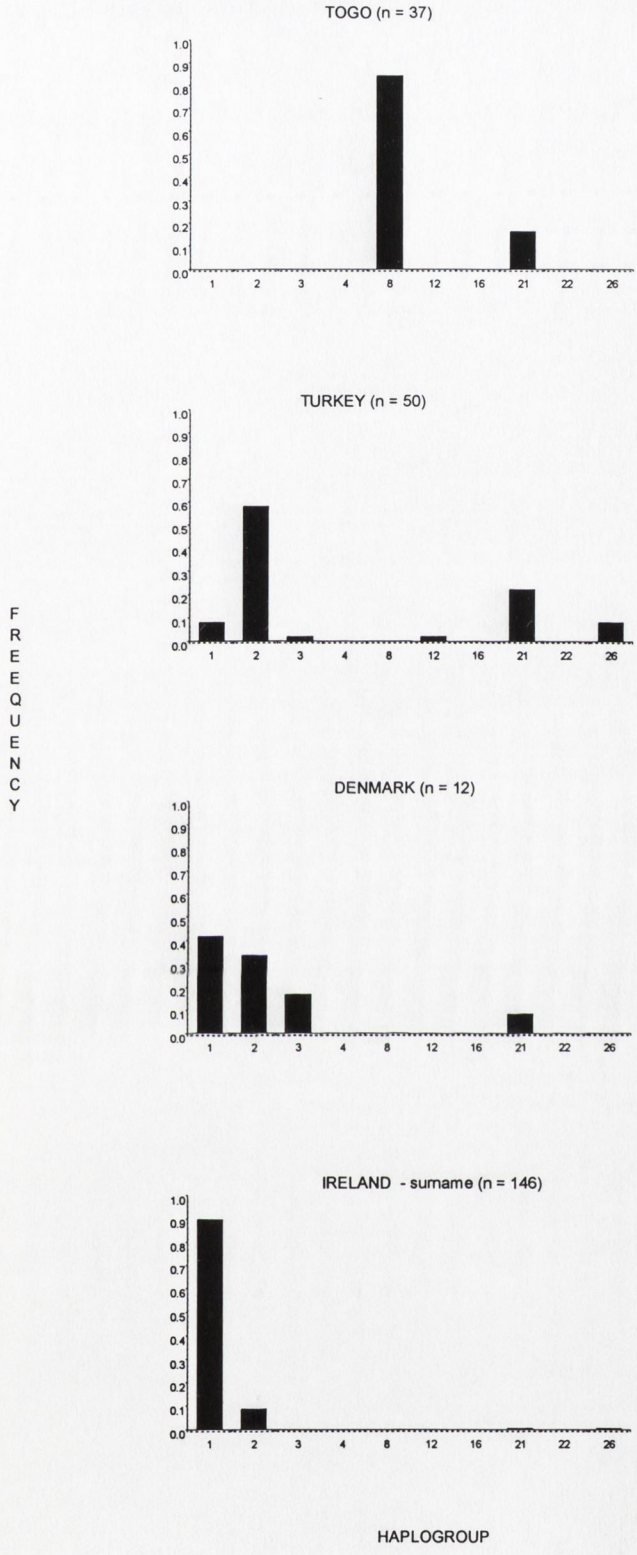


Figure 3.4 a: Frequency distributions of haplogroups sampled in the Togolese, Turkish, Danish and Gaelic Irish populations. The Turkish population is the most diverse containing six different haplogroups. One of the two Togolese haplogroups, hg 8, is not sampled in any of the other populations. Haplogroup 1 is prominent in the Gaelic irish population reaching a frequency of almost 90%.

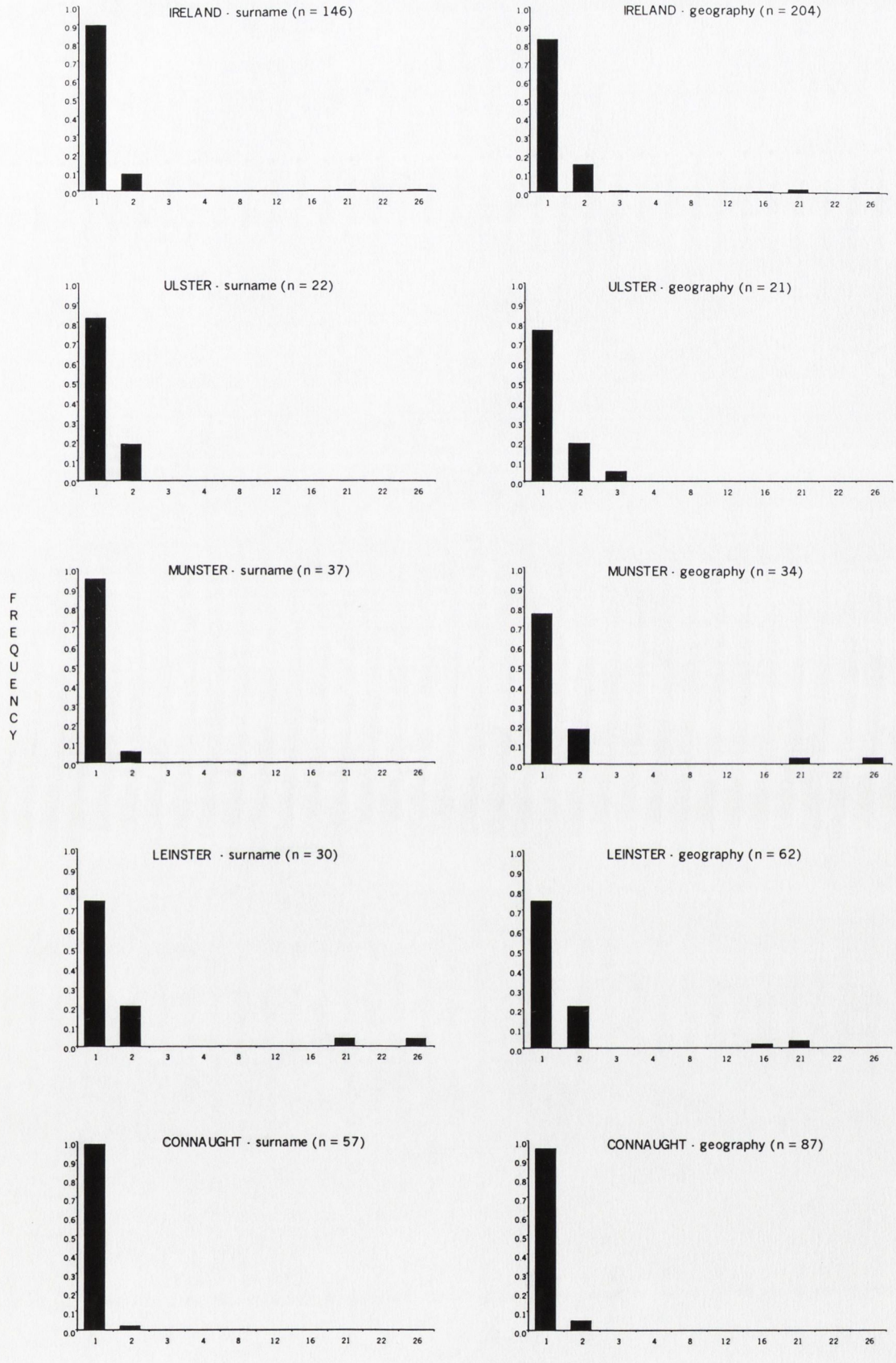


Figure 3.4 b: Frequency distributions of haplogroups sampled in Ireland. When samples are classified according to the origin of their surname more marked differences in distributions are evident than when classified according to the geographic origin of the sample.

Figure 3.4 c: Frequency distributions of haplogroups sampled in Ireland and classified into population cohorts according to the geographic origin of the sample and the origin of the surname. The distributions for the English and Scottish samples are similar when classified by either geography or surname.

When the Irish samples were separated into different surname cohorts then only 4 haplogroups were found to be present in the Gaelic Irish sample. The haplogroups that were absent in the Gaelic sample but were present in the other populations sampled in Ireland (i.e. those of exogenous origin) were hg 3 and hg 16. One Y chromosome of

Scottish origin was a hg 3 chromosome and one Norman Y chromosome was of hg 16 type.

Only two haplogroups, hg 1 and hg 2, were found in Ulster, Munster and Connaught. In Leinster two additional haplogroups, hg 21 and hg 26 were detected in one chromosome each. English chromosomes were found to consist of 3 haplotypic types and included hg 1, hg 2 and hg 21, although hg 21 was detected on only two chromosomes. Individuals of Scottish origin were found to have 4 haplogroups that included hg 1, hg 2, hg 3 and hg 21. Again hg 21 occurred rarely and was found on only one chromosome. Haplogroup 3 also occurred on one chromosome only. The Norman and Norse chromosomes contained 4 types, hg 1, hg 2, hg 16 and hg 21. One Norman individual was the only individual of the 326 samples to have a hg 16 chromosome. Danish chromosomes were comprised of hg 1, hg 2, hg 3 and hg 21. The Turkish samples were the most diverse. Six of the 8 haplogroups were detected in the Turkish population. Only hg 16 and hg 8 chromosomes were not found. Only two chromosomal types were found in the Togolese sample and these were hg 8 and hg 21.

3.3.2. Haplogroup 1

Haplogroup 1 Y chromosomes are defined by 3 allelic states: a T allele at the *HindIII* restriction site at locus 92R7 *(Mathias et al. 1994; M.E. Hurles and C. Tyler-Smith, unpublished)* (or an A allele at the *BanI* restriction site at locus DYS257 *(Hammer et al. 1998)*); a C allele at the *BsiHKAI or BanI* restriction site at locus SRY-*2627 (Veitia et al. 1997);* and a G allele at the *DraIII* restriction site at locus SRY-1532 *(Whitfield et al. 1995; Kwok et al. 1996).* Found to be widespread in Europe, hg 1 chromosomes reach their highest frequency in geographically extreme and isolated populations in the northwest of the continent *(this study, Semino et al 1996).* The most striking feature of the distribution of hg 1 chromosomes is the evident contrast between the southeast of Europe and the extreme northwest. In the Turkish population only 4 (8%) of 50 Y chromosomes are hg 1 chromosomes whereas in northwest populations the occurrence of hg 1 chromosomes ranges from 5 (41.7%) of 12, in Danes, through 25 (62.5%) of 40, in English, reaching their highest frequency in Irish and most notably in the Connaught subsample, comprising 131 (89.7%) of 146, and 56 (98.3%) of 57, of the Y chromosomes respectively.

In Munster two chromosomes were non-hg 1 chromosomes and in Ulster four chromosomes were found to be non-hg 1 types. In Leinster eight chromosomes were non-hg 1 chromosomes. The frequencies given in **table 3.6.** demonstrate the predominance of hg 1 chromosomes in European populations, particularly in Ireland, and the near fixation of this Y chromosome haplogroup in the west of the country.

The modal haplogroup in all populations, except for Turkey and Togo, was hg 1. Haplogroup 1 was not found at all in the African sample.

3.3.3. Haplogroup 2

Haplogroup 2 Y chromosomes are defined by 3 allelic states: the absence of the Alu insertion element at the YAP locus *(Hammer et al. 1994);* a G allele at the *DraIII* restriction site at locus SRY-1532 *(Whitfield et al. 1995; Kwok et al. 1996);* and a C allele at the *HinfI* restriction site at locus M9 *(Hurles et al. 1998; Underhill et al. 1997).* Haplogroup 2 chromosomes were found in all of the European populations but were not detected at high frequencies except in Turkey. In the Turkish population hg 2 was found to be the most common haplogroup where it was found in 29 (58%) of 50 chromosomes. The occurrence of haplogroup 2 chromosomes ranged from 1 (5.6%) of 18, in the Norman and Norse chromosomes through 13 (8.9%) of 146, in the Gaelic Irish chromosomes to 13 (33%) of 40, 4 (33.3%) of 12, and 6 (35.3%) of 17, respectively for English, Danish and Scottish chromosomes.

3.3.4. Haplogroup 3

Haplogroup 3 is a subgroup of hg 1 and is defined by an A allele at the *DraIII* restriction site at locus SRY-1532 *(Whitfield et al. 1995; Kwok et al. 1996).* In total only five hg 3 chromosomes were detected. In the Danish population 2 (16.7%) of the 12 Y chromosomes were hg 3 chromosomes. One hg 3 chromosome was found in Turkey and one was found in the Scottish population. One hg 3 chromosome of Indian origin was also found.

3.3.5. Haplogroup 8

Haplogroup 8 is a subgroup of hg 21 and is defined by a G allele at the *NlaIII* restriction site at the sY81 locus *(Seielstad et al. 1994).* Haplogroup 8 chromosomes were restricted to the Togolese population. The high frequency of hg 8 chromosomes in this population (83.8%) represented 31 out of the 37 African chromosomes.

3.3.6. Haplogroup 12

Haplogroup 12 is a subgroup of hg 26 and is defined by an A allele at the *HindIII* restriction site at locus LLY22g *(Righetti and Tyler-Smith, unpublished).* One hg 12 chromosome was detected in the Turkish sample only.

3.3.7. Haplogroup 16

Haplogroup 16 is a subgroup of hg 12 and is defined by a C allele at the *NlaIII* (or *MaeII*) restriction site at the Tat locus *(Zerjal et al. 1997).* Only one chromosome in the whole sample set was found to be hg 16.

3.3.8. Haplogroup 21

Haplogroup 21 is a subgroup of hg 4 and is defined by an A allele at the *BsrBI* restriction site at locus SRY-8299 *(Whitfield et al. 1995).* Haplogroup 21 is geographically widespread and was the only haplogroup common to all populations. Haplogroup 21 was found at the highest frequency in the Turkish population (22%). Togolese hg 21 chromosomes were found in 6 (16.2%) of 38 chromosomes. European populations were found to have low frequencies of hg 21 chromosomes ranging from 1 (< 1%) of 146 in the Gaelic Irish population to 2 (5%) of 40, 1 (5.9%) of 17, 1 (5.6%) of 18 and 1 (8.3%) of 12 in the English, Scottish, Norman/Norse and Danish populations respectively.

3.3.9. Haplogroup 26

Haplogroup 26 is defined by 3 allelic states: a G allele at the *HinfI* restriction site at locus M9 *(Hurles et al. 1998; Underhill et al 1995);* a C allele at the *HindIII* restriction site at locus 92R7 *(Mathias et al. 1994; Hurles and Tyler-Smith, unpublished)* or a G allele at the *BanI* restriction site at the DYS257 locus *(Hammer et al 1998);* and a C allele at the *HindIII* restriction site at locus LLy22g *(Righetti and Tyler-Smith, unpublished).* Haplogroup 26 chromosomes were found at low frequencies in the sample set as a whole. Haplogroup 26 comprised 4 (8%) of 50 of the Turkish chromosomes and was found in the Gaelic Irish population at < 1% (3.3% in Leinster, absent in all other provincial populations).

3.3.10. Geographic distributions of haplogroups

The geographic distributions of Y chromosome haplogroups are shown in **figures 3.5 and 3.6.** The most evident feature of the distribution of Y chromosome haplogroups

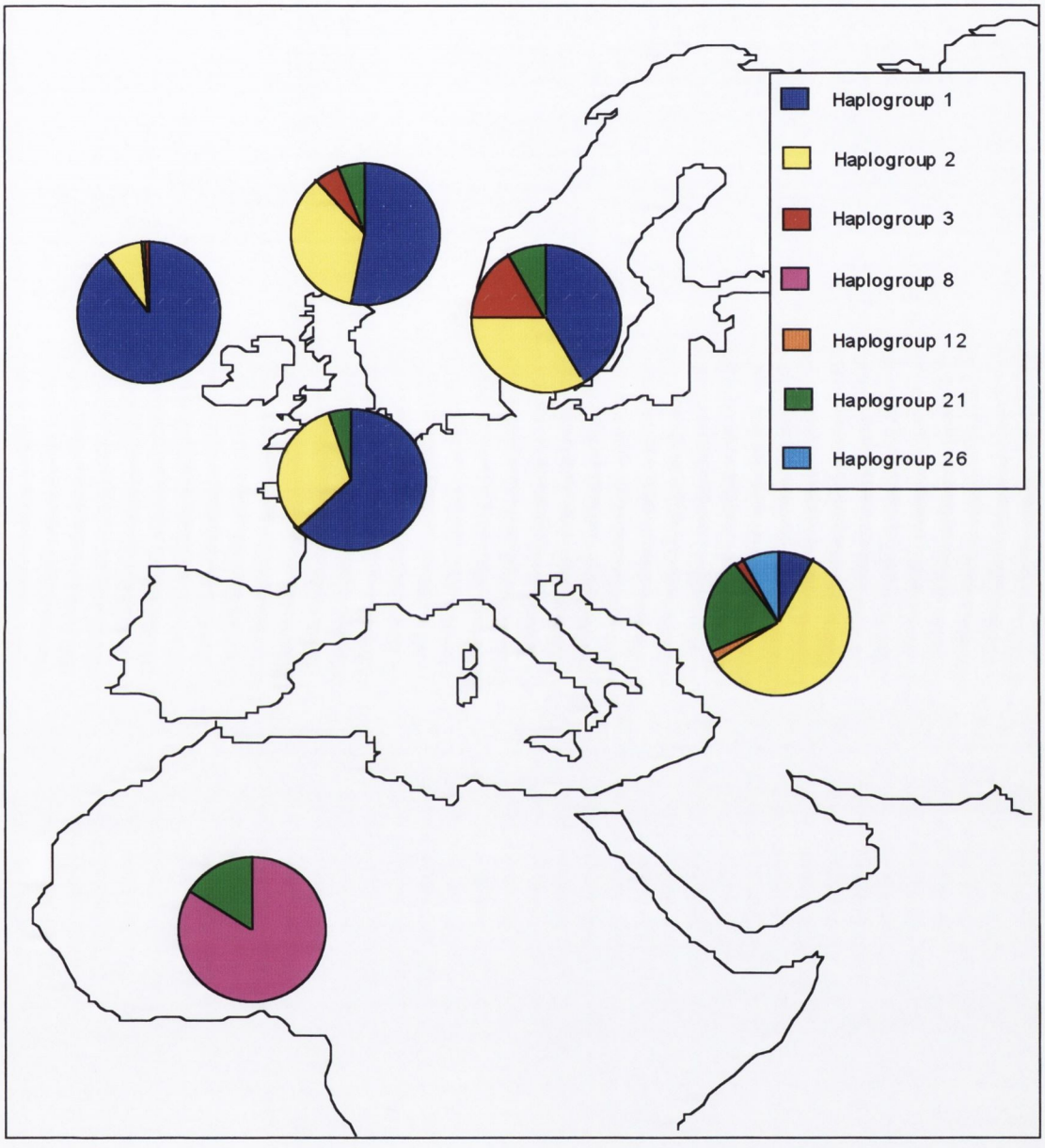


Figure 3.5: The geographic distribution of Y chromosome haplogroups in Europe and Togo, West Africa. The most geographically widespread haplogroup is hg 1, present at reasonable frequencies in all populations, but reaching its highest frequency in the northwest regions of Europe.

in Europe is the partial, symmetric cline of hg 1 and hg 2 between the southeast of Europe and the northwest. As hg 2 chromosomes decrease in frequency from southeast to northwest, hg 1 chromosomes increase in occurrence.

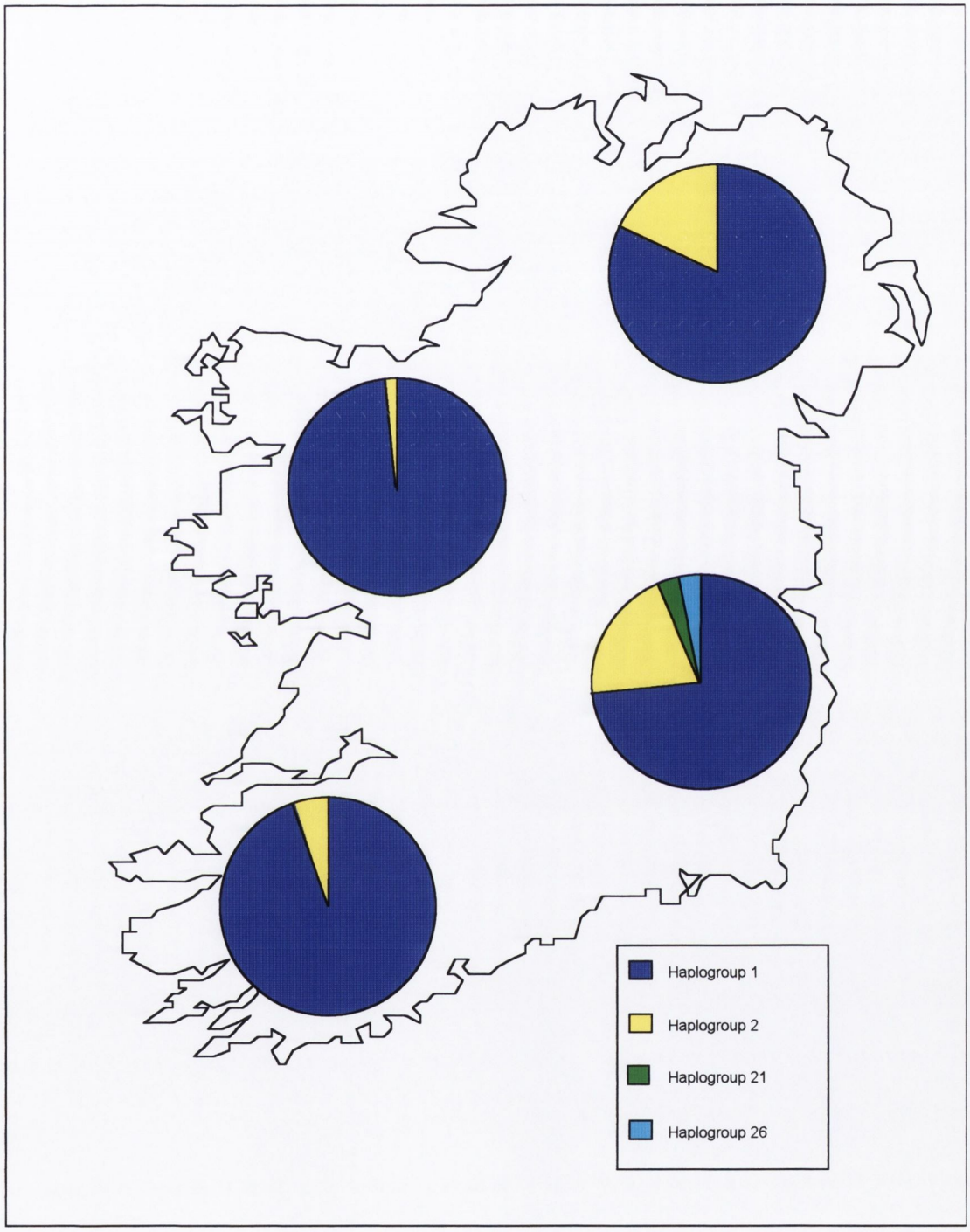


Figure 3.6: Within Ireland, the gradient of hg 1 frequencies continues culminating at near fixation (98%) in the western province, Connaught.

This cline is also present within Ireland itself. In Ireland the predominance of hg 1 is most evident. The near fixation of hg 1 chromosomes in Connaught is a very striking feature of this distribution.

The African sample contains hg 8 chromosomes which are not evident in any of the European populations. A small number of hg 21 chromosomes in the African population are also found. Haplogroup 21 frequencies decrease in numbers in the northwest of the continent but are found at reasonable frequencies in the southeast and Togo (22% and 16.2% respectively). The occurrence of the remaining haplogroups diminishes in the far northwest regions of Europe with the predominance of hg 1.

3.3.11. Assessment of population substructure in an AMOVA test

Results of analysis of molecular variance (AMOVA) tests are given in **table 2.7.** Three AMOVA tests using different population groupings were performed.

Population groups	[Connaught; Munster] [Ulster; Leinster] [Scotland; England; Norman/Norse]		[Connaught; Munster] [Ulster; Leinster]		[Ulster; Munster; Leinster; Connaught] [Scotland; England; Norman/Norse]	
Source of variation	**% variation**	**p value =**	**% variation**	**p value =**	**% variation**	**p value =**
Among groups	12.04	< 0.0099	14.49	< 0.0099	10.42	< 0.0099
Among populations within groups	1.77	0.3069	-1.27	0.6733	5.29	< 0.0099
Within populations	86.19	0 .9901	86.79	0 .9901	84.29	0 .9901

Table 3.7: Results of 3 AMOVA tests of population substructure.

Groupings were made by considering geographic factors, for example the eastern and northern Irish provinces were grouped and the western and southern provinces were grouped. Also exogenous Irish populations were grouped. When population cohorts, separated by surname origins, were tested (Gaelic names in [Ulster, Munster, Leinster and Connaught] against non-Gaelic names sampled in Ireland) the results indicated an among group partitioning of molecular variance which, in a permutation test, showed a significant partitioning of molecular variance ($p<0.0099$) between Gaelic and non-Gaelic surname samples. Similarly, when cohorts separated by surname origins within Ireland, relating to northern and eastern (Ulster and Leinster) and southern and western (Munster and

Connaught) populations, were tested the results also indicated a significant partitioning of molecular variance ($p<0.0099$). Additionally, when all sub-groupings were tested a significant partitioning of variance was detected ($p<0.0099$).

3.3.12. Assessment of population substructure with Fst genetic distances

A distance matrix of pairwise Fst distances was calculated for all population comparisons **(table 3.8).** The greatest genetic difference was found between the African and Connaught populations (Fst = 0.9682). The population found to be most closely related to the African population was Turkey (Fst = 0.6094). Within the Gaelic Irish

	Togo	Connaught	Munster	Leinster	Ulster	Denmark	England	Ireland (Gaelic)	Norman	Scotland
Connaught	0.9682									
Munster	0.9483	-0.0010								
Leinster	0.8601	0.2127	0.1041							
Ulster	0.9084	0.1754	0.0511	-0.0240						
Denmark	0.8180	0.4889	0.3181	0.0164	0.0800					
England	0.8021	0.3112	0.2048	-0.0004	0.0424	-0.0301				
Ireland (Gaelic)	0.9190	...	...	...	...	0.3075	0.2031			
Norman	0.8940	0.1114	0.0196	-0.0161	-0.0313	0.0745	0.0567	-0.0057		
Scotland	0.8230	0.4441	0.2864	0.0041	0.0629	-0.0671	-0.0385	0.2713	0.0704	
Turkey	0.6094	0.6599	0.5794	0.3663	0.4335	0.1844	0.2609	0.6412	0.4257	0.2050

Table 3.8: Pairwise Fst genetic distances for SNP data between global population groups. Samples in Ireland were separated into population cohorts by surname origins.

populations Connaught and Leinster were the most distantly related (Fst = 0.2127). Pairing between Connaught and Munster, and between Ulster and Leinster produced negligible genetic distances. Genetic distances increased roughly with geographical distance for all population comparisons.

3.4 Discussion

3.4.1. Evolutionary relationships of Y chromosome haplogroups

The biallelic markers assayed here define 11 possible haplogroups which are highly non-randomly distributed among human populations *(see also Jobling and Tyler-Smith 1995).* Based on information gained from a number of studies *(Oakey and Tyler-Smith 1990; Jobling et al 1994, Mathias et al 1994, Jobling et al 1996)* an unrooted maximum parsimony tree was developed relating all possible Y chromosomal haplogroups detected thus far *(Mark Jobling, personal communication).* The evolutionary relationships of these Y chromosomal lineages are shown in **figure 3.2.**

3.4.2. Population subdivision according to surname origins

Surnames have been used here to delimit population groups. That significant differences were found between chromosomes of Gaelic Irish origin and non-Gaelic Irish origin suggests that the subdivision of samples based on this criterion is valid.

Surnames contain two sources of information, ethnic and geographical, and it is therefore possible to reconstruct a 1,000 year old geographical sample of Irish variation by this means. By grouping chromosomes according to this ancient, prehistoric, classification the genetic variation in each population is concentrated and made more coherent. For comparison, chromosomes were grouped according to individual knowledge of paternal family history. In total 204 chromosomes were characterised, in this way, as being geographically Irish. Irish samples, denoted by geography, were found to be comprised of 6 haplogroups. Subsequent grouping according to surname (146 individuals) decreased this number to 4 haplogroups. All of the Irish provincial haplogroup distribution charts demonstrate this concentration of diversity that is indicative of a coherent history of Y chromosome lineages within Gaelic populations. One anomaly however is seen in the Leinster sample where hg 16 disappears but hg 26 appears in the sample. In Ireland, Leinster is the most variable region, possibly a result of its geographic proximity to Britain and susceptibility to immigration. Additionally, because many of the sample donors were uncertain of their family history and the majority of samples were collected in the Leinster

area a false perception of diversity in geographical Leinster might be expected. By separating samples into population cohorts according to the origin of their surname a less biased account of localised Y chromosome diversity has been achieved.

3.4.3. European distribution of haplogroup 1 Y chromosomes

Eighty percent (21/26) of European hg 1 Y chromosomes *(Jobling 1994)* belong to haplotype XV (ht XV), which is defined using the 49f/*Taq*I polymorphic system *(Ngo et al. 1986).* Some haplotypes of the 49f/*Taq*I system have previously been found to be population specific *(Semino et al. 1996),* ht XV being associated, in the most part, with Caucasians and those populations which have had contact with Caucasians *(Ngo et al. 1986, Torroni 1990, Spurdle and Jenkins 1992, Persichetti et al 1992, Jobling 1994).* The European distribution of hg 1/ht XV *(Semino et al. 1996)* chromosomes is shown in **figure 3.7.** Haplotype XV frequencies were converted, by mutiplying by 1/0.8 *(Jobling 1994),* to equivalent hg 1 frequencies from previously published data *(Semino et al. 1996 and references therein, Lucotte and Hazout 1996).* The frequencies of ht XV chromosomes and equivalent hg 1 frequencies are given in **table 3.8.** Although widespread in Europe, hg 1/ht XV chromosomes reach their highest frequency in geographically extreme and isolated populations in the northwest of the continent. Notably hg 1 chromosomes are almost fixed in Connaught, the westernmost region in Ireland. This is a remarkable observation that has been mirrored in only one other population, the putatively Mesolithic Basque population *(Lucotte and Hazout 1996).*

3.4.4. A Neolithic gradient

The most striking feature of the hg 1 distribution is the evident genetic cline extending from the southeast of Europe to the extreme northwest of the continent. This southeast - northwest cline closely mirrors other genetic gradients in Europe, particularly that discerned in the first principal component of a large classical data set *(Cavalli-Sforza et al. 1994)* and is most easily explained by the demic diffusion of Neolithic farmers from the Near East into Europe from ~10,000 *YBP (Ammerman and Cavalli-Sforza 1984).* It is evident from this and other studies *(Menozzi et al 1978; Sokal et al. 1991; Cavalli-Sforza et al. 1994; Semino et al. 1996; Zschocke et al 1997),* that migrations of these peoples

have left a mark on the European genetic landscape. This *wave of advance* model implies a large scale population movement including absorption of the indigenous Mesolithic genepool by the migrating Neolithics. However, it has been argued from mtDNA variation that this genetic gradient in fact represents a more ancient migration of Paleolithic peoples into Europe *(Richards et al. 1996).*

3.4.5. A Mesolithic gradient

One striking feature of the distribution of hg 1 chromosomes is the high estimated frequency (90.3%) of these chromosomes in the Basque region and an even higher frequency in Connaught and Munster (98.3% and 94.6% respectively) with median occurrences in France (48.5%), Scotland (52.9%), England (29.8% and 63%) and The Netherlands (64.4%). Elevated, extreme frequencies of alleles may result from genetic isolation from neighboring populations contributing to increased genetic drift. The analysis of many genetic systems (blood groups, enzymes, proteins, mtDNA sequences and Y chromosome 49f/*Taq*I haplotypes) has indicated that the isolated and genetically different Basque peoples may be representative of autochthonous Mesolithic hunter-gatherer populations who have had little exposure to more recent Neolithic peoples in Europe *(Bertranpetit and Cavalli-Sforza 1991; Lucotte and Hazout 1996; Semino et al. 1996; Richards et al. 1996; Torroni et al 1998).* Evidence from linguistics also indicates that the Basques are an unusual isolate in Europe, speaking a non-Indo-European language unlike any other spoken in Europe *(Ruhlen 1987).*

One hypothesis to account for this north-south cline attests that during the last Glacial Maximum, ~18,000 BP, indigenous hunter-gatherer populations, inhabiting northern regions of the continent, took refuge in milder southern areas *(Joachim 1983; Gamble 1986; Zschocke et al. 1997)* The Iberian Peninsula may have been one such

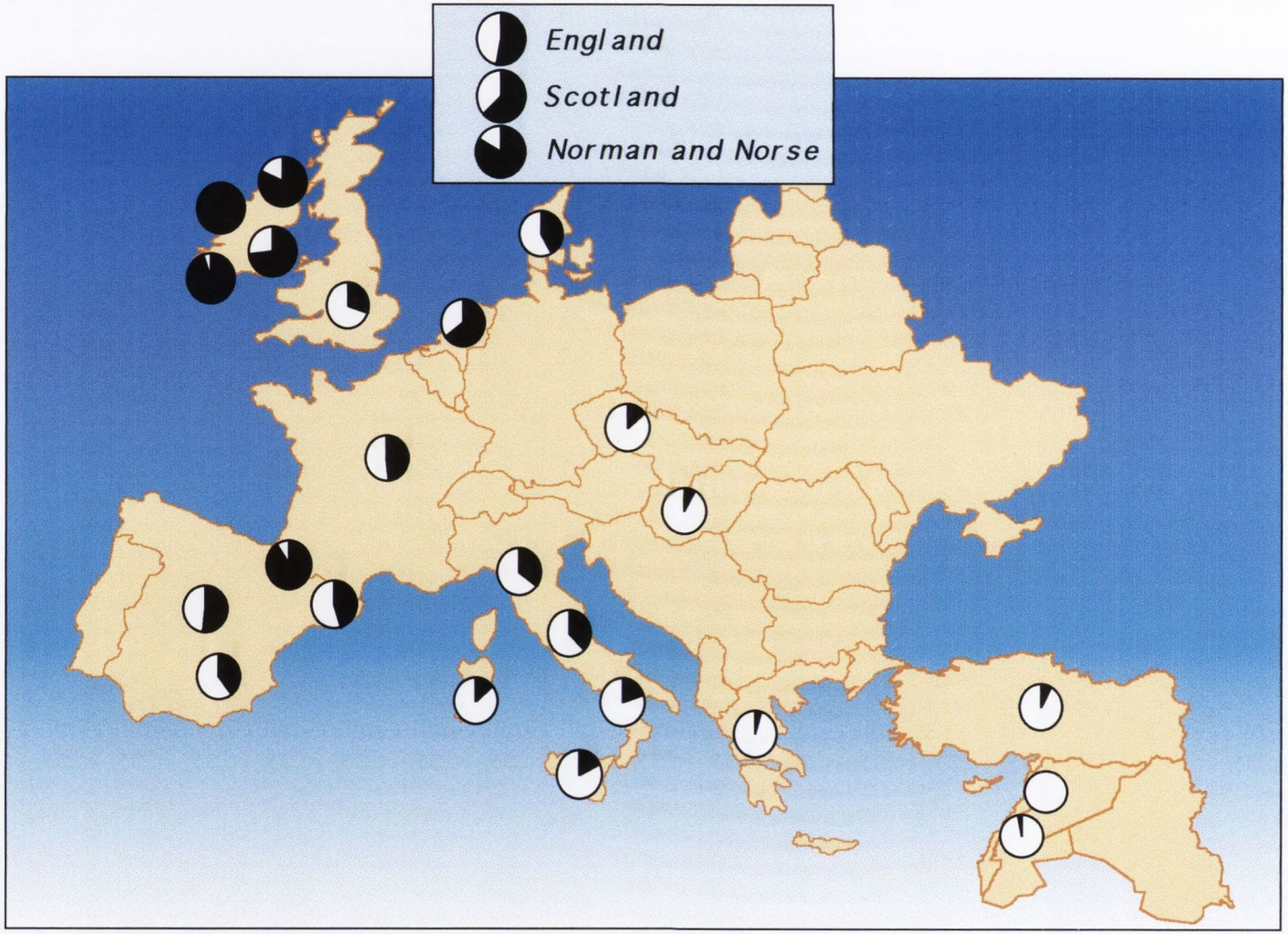


Figure 3.7: Haplogroup 1 / haplotype XV frequencies in Europe. Hg XV data was drawn from Semino *et al. (1996)* and Lucotte and Hazout *(1996)*. A cline of hg 1 frequencies across Europe is evident increasing in frequency from the southeast of the continent to the northwest of the continent reaching the highest frequencies in geographically extreme populations in the west of Ireland and in the Basque country.

Population	ht 15	hg 1
Connaught*	...	**98.3**
Munster*	...	**94.6**
French Basque[L]	**72.2**	**90.3**
Irish- Gaelic*	...	89.7
Norman/Norse*	...	**83.0**
Ulster*	...	**81.8**
Irish - all*	...	78.1
Leinster*	...	**73.3**
Bearnais[S]	56	70.0
Spanish Basque[S]	**53.8**	**67.3**
Dutch[S]	**51.5**	**64.4**
English*	...	63.0
Northwestern Europe[S]	49.1	61.4
Scottish*	...	**52.9**
Spanish[S]	**41.2**	**51.5**
French[S]	**38.8**	**48.5**
Central Italian[S]	**37.5**	**46.9**
Catalan[S]	35.7	44.6
Danish*	...	**41.7**
Andalusian[S]	**31.4**	**39.3**
Northern Italian[S]	**28.1**	**35.1**
Continental Italian[S]	25.0	31.3
English[S]	**23.8**	**29.8**
Southern Italian[S]: Calabriar	**17.6**	**22.0**
Southern Italian[S]: Apulian	**15.7**	**19.6**
Sardinian[S]	**13.7**	**17.1**
Southern Italian[S]: Sicilian	13.1	16.4
Czechoslovak[S]	**11.4**	**14.3**
Indian[S]	11.1	13.9
Ashkenazi Jewish[S]	7.2	9.0
Hungarian[S]	**6.4**	**8.0**
Turkish*	...	**8.0**
Algerian[S]	3.4	4.3
Greek[S]	**3.3**	**4.1**
Sephardi Jewish[S]	2.4	3.0
Tunisian[S]	2.3	2.9
Turkish[S]	1.4	1.8
Other Jewish Communities[S]	0.9	1.1
Ethiopian[S]	0	0
Togolese*	...	0
Lebanese[S]	0	0
Egyptian[S]	0	0
Albanian[S]	0	0
Hindu[S]	0	0
South African Indian[S]	0	0
Tharus[S]	0	0

Table 3.9: Haplogroup 1 frequencies. Data from Semino *et al. (1996)*[S] and Lucotte and Hazout *(1996)*[L] converted from 49a,f*Taq*I haplotype XV frequencies according to Jobling *et al. (1994)* { hg1 = 1/0.8 * ht XV frequency}; present study *. Only those populations in bold type are included in **figure 3.7**. The Calabrian and Apulian South Italian frequencies were averaged for the purpose of **figure 3.7.**

refuge area. It is possible that when the icesheets retreated the Mesolithic peoples migrated northwards, carrying with them their pre-Neolithic hg 1 Y chromosomes.

Evidence for this movement also comes from the distribution of the I65T allele of the phenylketonuria (PKU) gene, which is prevalent in both Ireland and in the Iberian Peninsula, with intermediate frequencies present in geographically intermediate locations *(Zschocke et al 1997).*

Further evidence for this cline is illustrated in the distributions of mtDNA haplotypes V and H in Europe which have high frequencies in extreme continental regions and lower frequencies in intermediate geographic regions *(Torroni et al 199).* Additionally, the fifth principal component of Cavalli-Sforza *et al. (1994),* which accounts for 5.3% of the total genetic variance, may represent the recolonisation of northern Europe following the last glaciation ~18,000 YBP.

3.4.6. East-West gradient detected within Ireland

Whatever the origin of hg 1 chromsomes in Ireland, it is interesting that not only are clines present across Europe but they are also maintained through the British Isles and, more strikingly, across Ireland itself. The distribution of Y chromosome hg 1 exhibits a stark east-west cline across the island **(figure 3.6).** Clines within Ireland have been detected before, the most evident being an east-west gradient of gene frequencies from Leinster to Connaught. Univariate analysis of among group variation in anthropometric data has inferred an east-west gradient in biological variation in Ireland *(Hooton. 1940; Hooton et al. 1955).* The distribution of alleles of the ABO blood group system has also suggested an east-west gradient with a high frequency of blood group O in the west and low in the east, mirrored by a relatively higher frequency of blood group A in the east and lower in the west. The distribution of the d allele of the Rhesus system also exhibits a clinal dispersion across the country *(Hackett and Dawson 1958).* Relethford and Crawford (*1995*) used principal components analysis to detect population substructuring within Ireland, also using anthropometric data. The first principal component separates Norway and Denmark from England and Ireland and accounts for 51% of the total variation. The second principal component accounts for 34% of the total variation and separates England from the rest. The plot of the second principal component illustrates an east-west gradient across Ireland.

3.4.7. Male and female migration

The observation that hg 1 chromosomes exhibit near maximal differences between extreme geographic regions of the gradient is consistent with the suggestion that human populations show higher between-group variability (Fst) for paternally inherited variation than for other genetic systems *(Seielstad et al 1998).* Between-population variation has been found to be much greater between the Y chromosomes of diverse population groups than between both autosomes and maternally inherited mtDNA. It has been suggested that this difference may be related to the migration patterns of females who, in 70% of societies, move from their birthplace to that of their spouse. This cultural phenomenon of patrilocality may account for the large differences between geographically distant population groups. Seielstad *et al.* (1998) compared Fst genetic distances between mtDNA, autosomes and Y chromosomes. MtDNA Fst distances were found to be 8.22 ($P<0.05$) times greater than equivalent distances for Y chromosomes. This suggests that female migration rates are up to 8 fold greater than male migration rates *(Seielstad et al 1998).*

3.4.8. Putative European Paleolithic/Mesolithic Y chromosome type

The correlation of the demic diffusion hypothesis with the European cline of hg 1, together with its virtual fixation in the western Gaelic fringe and in Basque peoples, suggests that earlier Paleolithic and Mesolithic inhabitants of Europe were fixed for this haplogroup. The presence of other haplogroups in extreme, isolated populations might have resulted from later introgressions possibly associated with the Neolithic movement of peoples throughout Europe.

3.4.9. Distribution of haplogroup 2 Y chromosomes

Haplogroup 2 Y chromosomes are geographically widespread being found in both Caucasian and Asian populations *(Jobling et al 1994).* More specifically, hg 2 chromosomes have been found in Chinese, Indian and European populations at reasonable frequencies *(Jobling and Tyler-smith 1995).* In this study the majority of hg 2 chromosomes are found in the Turkish population where they constitute more than 50% of

all chromosomes. Additionally, two hg 2 chromosomes were found in Asians from Hong Kong and China.

Haplogroup 2 chromosomes in Ireland

There is a clear pattern of hg 2 distribution in Ireland. The presence of this haplogroup in the eastern and northern regions of Ireland and their absence in the west and south suggests that hg 2 Y chromosomes are intrusive. As a result of their proximity to Britain, eastern regions of Ireland have been more susceptible to population migration. The intrusion, or intrusions, must have been relatively recent, as gene flow has not yet facilitated the introgression of hg 2 chromosomes into more distal parts of the island.

Three historical intrusions into Ireland are well documented. The Viking invasions from the Nordic countries in the 9th and 10th century AD resulted in a number of settlements most notably in Dublin and Waterford. The Anglo-Norman invasion, prompted by the invasion of the Normans into Britain in the 12th century, is also well documented. Most recently the influence of the Tudor plantations in Ulster and Leinster in the 16th-17th centuries AD has played a large role in the demography of the eastern parts of this island *(Smyth 1997).*

However, the subdivision of samples into specific population cohorts on the basis of surname history attempts to filter out migrations occurring after the 11th century BC. It is unlikely, therefore, that these historical events have heavily contributed to the distribution of hg 2 in the Gaelic Irish populations unless a substantial breakdown in the association between surnames and Y chromosome through non-paternity can be assumed. It is most likely that hg 2 Y chromosomes, in Ireland, predate historical intrusions and can only be explained by older prehistoric influences.

A commonly stated possibility, but with a lack of firm supporting evidence from archaeology and previous genetic studies is that the Celts, arriving from mainland Europe, made an impact on the indigenous Irish population during the 1st millennium BC *(Waddel 1998; Zschocke 1997).* It is widely accepted that a Celtic Iron Age culture arrived in Ireland about 2,000 - 2,500 *YBP*, and many artifacts of this time remain *(Harbison 1988).* However, if the presence of hg 2 in Ireland does indeed represent the survival of Celtic lineages, then the question arises: why have these lineages not penetrated the western and southern regions of the county?

3.4.10. Y Alu polymorphism: an indicator of global population movement

The Y Alu polymorphism (YAP) is a stable polymorphism resulting from the recent insertion of an Alu element on the Yq arm of the Y chromosome within the past 29,000 - 334,000 years *(Hammer 1994; Hammer 1995).* Comparative sequence analysis of a 2.6 kb fragment of the YAP element with human, chimpanzee and gorilla Y chromosomes has established that this insertion event occurred only once in human population history and that the absence of the Alu element is the ancestral state *(Hammer 1995).*

In a recent study of Y chromosome polymorphism, the ancestral YAP- haplotype, found also in chimpanzees, was found to occur in African populations only. The highest frequency of this ancestral haplotype was found in Khoisan peoples *(Hammer et al. 1998).* In a recent study by Hammer *et al. (1998)* other YAP- haplotypes were detected in most global populations which suggested that the polymorphic Alu insertion could have arisen anywhere in the world. Previously, studies of YAP+ global distributions determined the origin of the YAP+ chromosome to be in Africa because high frequencies of YAP+ chromosomes were found in Africa and rarely elsewhere, except in some populations of East Asia *(Hammer 1994).*

The high frequency of YAP+ chromosomes in some Asian populations, however, and results from a nested cladistic analysis of Y chromosome haplotypes has suggested that in fact the origin of YAP+ chromosomes may have been in Asia and not, as suggested, in Africa *(Hammer et al 1998).* This had previously been suggested by others *(Excoffier et al. 1987).* A greater genetic diversity in African populations at non-Y chromosomal loci has inferred that Africans are the oldest modern humans and that global genetic diversity is a subset of the diversity found in Africa *(Stoneking and Cann 1989; Armour et al 1996, Tishkoff et al 1996).* Therefore in order to explain the present day distribution of global YAP+ chromosomes Hammer *et al.* (1998), proposed that ancestral African Y chromosomes (YAP-) were carried out of Africa into Asia and the rest of the world, perhaps in a number of separate migration events, and that these migrations were responsible for the replacement of the indigenous *Homo erectus* populations. Frequency analysis of a polymorphic G → A transition at locus SRY-4064 has suggested, on the basis of a high frequency of this polymorphism in Asian populations, that the YAP insertion arose somewhere in Asia after the initial colonisation event by modern humans, and was

subsequently brought back to Africa by later migrations without a total population replacement event. Therefore a proportion of the African genepool has a largely Asian influence. Additionally, DNA sequencing studies have shown that eight geographically diverse chromosomes (from African and Asian populations) share the same two base pairs at the insertion site, and therefore YAP+ chromosomes in Asia and Africa are identical by descent *(Hammer and Horai 1995).*

A range expansion back into Africa from Asia has been detected before in studies of the β-globin locus. Introgression of South-East Asian sequences into Africa were detected without complete replacement *(Harding et al. 1997).* The number of bidirectional migrations is unknown, but it is possible that the same expansion back into Africa from Asia brought both the Asian β-globin sequences and YAP+ Y chromosomes.

Conversely, in an assessment of diversity in global populations using the most polymorphic marker detected to date on the Y chromosome, MSY1, Jobling *et al. (1998)* do not find evidence for an Asian YAP+ influence in African populations, although they accept that the high mutation rate of the MSY1 locus could be obscuring such a deep rooted relationship.

The present day distribution of YAP+ chromosomes is given in **table 3.9**. In this study, the highest frequency of YAP+ chromosomes was found in the Togolese population (100%) which is consistent with previous studies of YAP+ chromosomes *(Hammer 1994, Hammer 1995, Seielstad et al 1994, Passarino et al 1998, Hammer et al 1998; Spurdle et al 1994).* The West African populations, including Togolese and Senegalese populations have the highest frequency of YAP+ chromosomes and in fact YAP+ chromosomes are fixed in these populations. Few YAP+ chromosomes have been detected in non-African populations, the exception being East Asian populations where 42% of Japanese chromosomes are YAP+ *(Hammer and Horai 1995).*

Population	% YAP$^+$ chromosomes
AFRICAN	
Saudi Arabian^P	10.0
Egyptian^P	53.1
North Africans^H	49.5
East Africans^H	88.0
Ethiopian^P	50.0
West Africans^H	87.5
Mandenka Senegalese^P	98.2
Wolof Senegalese^P	100.0
Togolese*	100.0
South African Bantu speakers^H	76.6
Pygmies^H	53.5
Khoisan^H	26.9
Nama and Sekele San^P	46.0
Tsumkwe San^P	11.0
ASIAN	
Australasians^H	1.0
West Asians^H	8.6
East Asians^H	22.3
EUROPEAN	
Europeans^H	13.8
Turkish*	22.0
Finns^S	0
Estonians^S	5.0
Saami^S	0
Swedes^S	0
Swiss^S	14.0
Basques^S	8.0
Irish*	0.7
English*	5.0
Scottish*	5.9
Norman/Norse*	5.6
Danish*	8.3

Table 3.10: YAP+ chromosome frequencies. Data from Passarino *et al. (1998)*^{P} ; Hammer *et al. (1997)*^{H} ; Sajantila *et al. (1996)*^{S} ; present study *.

3.4.11. Haplogroup 21 and haplogroup 8 distributions

An A → G transition at the sY81 (or DYS271) locus has been shown to exhibit maximal linkage disequilibrium with the Y Alu insertion element and has been found at high frequencies in African YAP+ chromosomes *(Seielstad et al 1994).* This polymorphism must have evolved relatively recently as it is present only on YAP+ chromosomes of African origin and therefore must have occurred following the introgression of Asian Y chromosomes back into Africa.

Haplogroup 8 chromosomes carry the sY81 G allele, whereas hg 21 chromosomes have the A allele. The apportionment of hg 8 and hg 21 chromosomes on YAP+ lineages

in African populations is shown in **table 3.10.** In the Togolese population 100% YAP+ chromosomes, of which 84% had the G allele, were detected. In the Senegalese population there were 99% YAP+ chromosomes of which 85% were hg 8. In the Ethiopian population there were 52% YAP+ chromosomes, none of which were hg 8.

Population	% hg 8	% hg 21
Mixed African[P]	41	18
Ethiopians[P]	0	52
East Africans[H]	86	11
East African Bantu speakers[H]	55	15
West Africans[H]	66	24
Togolese*	84	16
Senegalese[P]	85	14
Wolof Senegalese[P]	71	29
Mandenka Senegalese[P]	93	5
West African Bantu speakers[H]	78	8
Sub-Saharan Africans[H]	57	12
Dama[H]	65	15
Khoisan[H]	20	10
Pygmies[H]	50	5

Table 3.11: The apportionment of hg 8 and hg 21 chromosomes on YAP+ lineages in African populations. Data from Hammer *et al.* (1998), Passarino *et al.* (1998) and present study*.

Haplogroup 8 is an African-specific lineage which has not been detected in non-African global populations *(Seielstad et al 1994).* In the Togolese sample 84% of chromosomes belong to this group. West African populations in particular have high frequencies of hg 8 chromosomes. Conversely the East African Ethiopian population has been found to contain no hg 8 chromosomes *(Passarino et al. 1998).* A comparative analysis with mtDNA and Y chromosome haplotypes led Passarino *et al. (1998)* to propose that the Ethiopian population has been greatly influenced by male-mediated Caucasoid geneflow and female-mediated African geneflow. They have proposed that hg 8 chromosomes were spread by Bantu expansions which did not influence the Ethiopian population at all. The Bantu expansion causing the introgression of YAP+/sy81-G chromosomes has influenced much of west and southern Africa and indeed some of east Africa.

Haplogroup 8 chromosomes are rare in Khoisan and Pygmy peoples, two ancient tribes of southern Africa. Khoisan peoples have also been found to contain the rare A

allele at the M42 locus *(Underhill et al 1996)* at a frequency of 15%. The A allele is present in primates and suggests it as the ancestral allele at this locus. In Ethiopians and the Sudanese the A allele is found at a frequency between 5-10%. The polymorphic T allele is found in non-African populations *(Gibbons 1997).* That ancient African populations do not contain hg 8 chromosomes at a high frequency suggests that its presence is due to a relatively recent evolutionary event which occurred after the separation of African populations.

Haplogroup 21 chromosomes on the other hand are geographically widespread. Haplogroup 21 chromosomes either evolved in Africa following Asian reintroduction of Y chromosomes and then were spread via later migrations out of Africa into the rest of the world, or the hg 21 lineage evolved before the reintroduction into Africa of YAP+ chromosomes. Recently Hammer *et al. (1998)* have shown that hg 21 chromosomes originated in North Africa ~20,000 *YBP* and moved into the Middle East ~10,000 *YBP* from where they were spread throughout Europe following the trend of the Neolithic demic diffusion. This cline is evident in the populations in this study. Turkish chromosomes have the greatest number of hg 21 chromosomes (22%). From this geographically central location hg 21 chromosomes are found at decreasing frequencies, into the Togolese population (16.2%) and into Europe where there is a decreasing gradient of hg 21 terminating in Ireland (0.7%).

3.4.12. Haplogroup 3 chromosomes are restricted to Indian populations

Haplogroup 3 chromosomes were initially found to be restricted mainly to populations from the Indian subcontinent *(Jobling et al 1994; Kwok et al 1996).* Haplogroup 3 chromosomes, equivalent to haplogroup 1D in Hammer *et al. (1998) (Hurles et al 1998),* have, however, more recently been found in both European and Asian populations albeit at low frequencies in the European populations. In Indians hg 3 chromosomes reach a frequency of 30.9% *(Veitia et al. 1997).* Haplogroup 3 chromosomes were expected to be rare, and possibly absent, in the populations surveyed here and this is the case. The one Indian sample in this study was found to have a hg 3 Y chromosome.

3.4.13. Haplogroup 22

No hg 22 chromosomes were detected in any of the populations in this study. Haplogroup 22 chromosomes have been found in only 7.3% of Caucasian males.

3.4.14. Northern European predominance of hg 26, hg 12 and hg 16 chromosomes

By reference to **figure 3.2**, hg 26, hg 12 and hg 16 are related by one and two evolutionary events respectively *(E. Righetti and C. Tyler-Smith, unpublished; Zerjal et al. 1997; Underhill et al 1997, Jobling et al 1996).* They are found at very low frequencies in the populations studied here. Haplogroup 26 is found only in the Turkish population and in Leinster. Haplogroup 12 is found once in the Turkish population and hg 16 is found once in the Norman sample. By drawing on previously published literature, however, these chromosomal haplogroups can be put into an evolutionary context.

The smaller deletion at locus 50f2/C, determined to be representative of hg 12, was assayed in global populations *(Jobling et al. 1996).* Haplogroup 12 chromosomes were found in Asian and Northern European populations. In Mongolians 5 (41.6%) of 12 chromosomes were hg 12. In Indians 1 (100%) of 1 chromosome was hg 12. In Chinese 3 (60%) of 5 chromosomes were hg 12. In Finnish, Norwegian, Yakut, Saami, Greek, Altai and Japanese populations 11 (100%) of 11, 1 (100%) of 1, 5 (100%) of 5, 3 (100%) of 3, 1 (50%) of 2, 2 (40%) of 5 and 1 (50%) of 2 chromosomes were found to be hg 12, respectively. Haplogroup 12 chromosomes were not detected in African populations and it has been suggested that therefore this group may have evolved after the migration of humans from Africa into Asia. Although the dating of the most recent common ancestor of this chromosome is uncertain it has been estimated that hg 12 chromosomes arose in Asia recently, some time around ~2,000 - 4,000 BP *(Zerjal et al. 1997).*

Recently, a number of polymorphisms, assayable by PCR, have been described which separate hg 26, hg 12 and hg 16 *(E. Righetti and C. Tyler-Smith, unpublished; Zerjal et al. 1997; Underhill et al 1997, Jobling et al 1996).* Haplogroup 26 has been found to be present in all geographic populations except for Africans *(Underhill et all 1997).* This widespread geographic distribution outside Africa suggests that the most parsimonious explanation for the occurrence of the mutation is that it occurred outside

Africa after the separation of Africans and non-Africans and subsequently dispersed *(Underhill et al 1997).* Indeed, in this study it is not present in the Togolese sample.

Population	% hg 16
EUROPEAN	
Norman/Norse*	5.6
Norwegian[Z]	3.8
Finn[Z]	52.4
Saami[Z]	25.0
Estonian[Z]	47.4
Mari/Morinsky[Z]	50.0
Mari/Gornomariysy[Z]	35.0
Mordva[Z]	22.2
Russian[Z]	15.0
ASIAN	
Buryat[Z]	57.7
Khalkh[Z]	2.1
Mjangad[Z]	50.0
Khalimag[Z]	100.0
Yakut[Z]	85.7
Japanese[Z]	< 1.0

Table 3.12: Haplogroup 16 frequencies in Europe and Asia. Data from Zerjal *et al. (1997)*[Z] ; and present study *

A T → C transition has been identified which delimits hg 16 chromosomes when present on a hg 12 background *(Zerjal et al. 1997).* The highest frequencies of the C allele have been found in populations from northern Europe and Asia, specifically in Finns (52.4%), Yakuts (85.7%) and Buryats (57.7%). Genetic diversity analysis has indicated an origin for this haplogroup in Asia *(Zerjal et al 1997).* Populations with hg 16 chromosomes and their frequencies are given in **table 3.12.** It has been suggested that hg 16 chromosomes arrived in Europe with Uralic speaking populations from Asia. Uralic languages are spoken in Finnish, Saami and Estonian populations. The correspondence between language and Y chromosomes, however, was not found to hold true and has been explained by genetic drift or language replacement in some populations *(Zerjal et al. 1997).* Although language affinities with Y chromosomes may not be robust, that the one hg 16 chromosome detected in this study was of Norman origin, rather than having its origin in one of the more numerically represented populations, supports the hypothesis that surnames may be strong indicators of Y chromosome lineage history. The Normans, a people descended from Viking raiders who had settled in north-west France in the 9th

century, arrived in Ireland in the 12th century. The one hg 16 chromosome detected here is therefore of northern European origin.

3.4.15. Surnames as indicators of Y chromosome lineages

Of the thirty surnames sampled more than once in Ireland only 3 were found to have different haplogroups in different individuals. These were one Norman, one English and one Leinster Gaelic. The Norman name originated in two different geographic regions according to paternal family history. The other originated in similar regions according to this classification. That 90% of all shared surnames share the same haplogroup further suggests that subdividing names into population cohorts according to surname origin has validity.

The only non-hg 1 chromosome in Connaught was detected on a chromosome associated with the surname McKeown. According to MacLysaght, *"No less than seventeen variants and synonyms of MacKeown have been officially recorded....The main sept is of north Connaught with a branch in Co. Galway. The MacKeowns of the Glens of Antrim are mainly descendants of the Scottish Bissets"*. It is conceiveable that this sample has been erroneously grouped, and may accordingly be grouped in the Scottish cohort. If this were the case, then 100% of western Irish Y chromosomes would be hg 1 chromosomes. However, this would require further investigation of the individual sample history in order to discern the exact location of origin for this line of McKeowns.

The only hg 16 chromosome detected was on a chromosome of Norman origin which was associated with the surname Bryan. MacLysaght states that Bryan is *"The name of a prominent Anglo-Norman family settled in Co. Kilkenny"*. Further discussion of hg 16 chromosomes can be found in **section 3.4.13**.

Chapter 4:

Genetic diversity on the Y chromosome

4.1 Introduction

The hierarchical approach to the assessment of human genetic diversity used here involves initial subdivision of individual samples into discrete haplogroup lineages using unique event polymorphisms, and subsequent analysis within those lineages by employing faster evolving loci. Unique event polymorphisms have a relatively low mutation rate. Base substitutions mutate at a rate of approximately 5×10^{-7} per site per generation *(Hammer 1995)*, and therefore they are believed to have arisen only once in human evolutionary history. Faster evolving loci such as microsatellites have been shown to have much higher mutation rates, estimated to be 2.1×10^{-3} for Y chromosome microsatellites *(Heyer et al. 1997)*. Minisatellites have an even higher mutation rate, estimated for Y chromosome minisatellites as between 2% and 11% per generation *(Jobling et al 1998)*.

4.1.1 Microsatellites in population genetic studies

Microsatellites have been widely employed in the determination of human demographic history. Bowcock *et al. (1994)* were the first to use autosomal STR loci for the construction of an evolutionary tree of human populations. Thirty microsatellites on chromosomes 13 and 15 were analysed in 10 individuals from each of 14 ethnically diverse human populations. Using a genetic distance measure based on the sharing of alleles at each locus a phylogenetic tree was constructed which showed the geographic clustering of individuals within populations. This was found to contrast with trees constructed from mtDNA sequence data in which individuals were rarely found to cluster according to their geographic origin *(Cann et al. 1987, Vigilant et al 1991)*. The deepest split in the tree was that between Africans and non-Africans, which was supported in 100% of bootstrap resamplings. Subsequent separations were that of Europeans, East Asians and Pacific populations. The greatest genetic diversity was detected within African populations compared to other non-African populations supporting an African origin for modern humans. Although the sample sizes in the study by Bowcock *et al. (1994)* were quite small and STR loci were taken from only two chromosomes, a more recent study employing 60 autosomal STR loci from 14 autosomes and using larger population sizes has confirmed the result that the deepest split is that between Africans and non-Africans

and that Africans have higher diversities at STR loci than non-Africans *(Jorde et al 1997).* On average, Africans were found to have ~20% greater genetic diversity than Asian and European populations. Higher genetic diversity in Africans was also detected in 20 autosomal STR loci in a global sample of 16 populations. Lowest diversity was detected in Asians with Europeans intermediate to the two. Phylogenetic analysis and the greater diversity in African populations also concord with an African origin for human populations *(Pérez-Lezaun et al. 1996).*

4.1.2. Y specific microsatellite loci in population studies

Recently great attention has been paid to the use of Y specific microsatellite loci in studies of human evolutionary history. More than 20 Y STRs have been characterised so far *(Jobling and Tyler-Smith 1995, Mathias et al. 1994, Jobling et al. 1996, Kayser et al 1997).* Y specific microsatellites have an advantage over autosomal STRs in that they are inherited together as haplotypes. This allows for the application of powerful phylogenetic analyses for population genetic studies *(Hammer 1995).*

Furthermore, large numbers of autosomal microsatellite loci are required to resolve microevolutionary events between closely related populations. Zhivotovsky and Feldman *(1995),* for example, have estimated that as many as several hundred STR loci are required for fine-scale resolutions of population relatedness. In contrast, only four Y chromosome microsatellites have been shown to be required to detect significant differences between very closely related populations. Roewer *et al. (1996)* analysed seven Y STR loci in two closely related populations (German and Dutch). An analysis of molecular variance of haplotypes determined from four of the loci revealed a larger mean proportion of between population variance (66.8%) than within population variance (33.2%). Similarly it was shown in a further study that Y STRs had the resolving power to discriminate between closely related populations. Significant statistical differences were detected between Dutch, German, Swiss and Italian populations in an analysis of molecular variance incorporating haplotype information generated from seven Y STR loci *(de Knijff et al. 1997).* Pairwise comparisons could not, however, discriminate between either the Dutch and Swiss or the Swiss and Italian populations. On average 3.5% of the total genetic variation was found between the populations, the highest variability being that between the Dutch and German populations (8.12%).

The most comprehensive study, to date, on the geographic distribution of Y STR alleles comes from de Knijff *et al. (1997),* in which 3,825 males from 48 population groups from Europe, Asia, America, Africa and Oceania were analysed for one or more of the following loci: DYS19, DYS389-1, DYS389-2, DYS390, DYS391, DYS392 and DYS393. Allele frequency distributions for alleles at these loci demonstrate that Y STRs conform to the stepwise mutation model for microsatellites. Furthermore, some population specific alleles were detected (for details see **section 4.4**). Phylogenetic trees were constructed from the Dc *(Cavalli-Sforza and Edwards 1967),* Ds *(Nei 1972)* and Ddm *(Goldstein et al 1995)* genetic distance measures. The most striking feature of all three trees is the relatively short branch length leading to the African population. This is in contrast to trees based on mtDNA data and autosomal loci but was also found in another study of Y STRs *(Deka et al. 1996).* De Knijff *et al. (1997)* suggest that this may be resulting from the fast mutation rate of Y STRs and the mode of mutation which results in recurrent mutations. They further suggest that Y STRs may not be suitable for the inference of evolutionary events but may be more suitable for more recent historical events due to their mutation - time linearity being maintained only up to 49,000 years ago.

Y STRs have been used in a number of other human evolutionary studies. For example, Pérez-Lezaun *et al. (1997)* analysed eight Y STR loci in the Basque and Catalan populations. By comparison to autosomal loci in the same samples, analysis of Fst values and genetic diversity measures in the 8 loci suggested that Y STR diversity follows a similar pattern as for autosomes in these two populations. Haplotype and phylogenetic analysis using a 5-locus subset of the STR loci (DYS19/388/390/391/392) detects two possible ancestral haplotypes, *ht 1* and *ht 2* (their nomenclature; for further discussion see section **4.4**). Using the Bertranpetit and Calafell *(1996)* method of dating (outlined in **section 4.3**) the diversity within the two populations was estimated to have arisen between 15,000 and 61,000 *YBP* which is in agreement with dates obtained from mtDNA D-loop sequences *(Bertranpetit et al 1995, Comas et al 1997).* In the phylogenetic tree, constructed using maximum parsimony, haplotypes were scattered randomly with a lack of population specific clustering. They suggest that this may be because the diversity within these two populations arose prior to the separation of the populations which may predate the differentiation of European populations.

Deka *et al. (1996)* analysed five Y STR loci in 15 ethnically and geographically diverse human populations belonging to African, Caucasian, Mongoloid and American

population groups. The network constructed from haplotype data was characterised by a lack of clustering of haplotypes in geographically proximal populations. This might be explained by the recurrent nature of microsatellite mutation coupled with the high mutation rate of Y STRs. In a further study of Y STR haplotypes in East Anglian, Sardinian and Nigerian populations Cooper *et al. (1996)* constructed a network relating haplotypes. In addition to microsatellite loci the presence or absence of the YAP element was assayed and used to create compound haplotypes. This resulted in the almost complete separation of East Anglian haplotypes from haplotypes sampled in the other populations, although the resolution of the network is not optimal. Furthermore, a network relating 77 haplotypes detected in Dutch and German individuals demonstrated no population specific structuring in the clustering of haplotypes *(Roewer et al 1996).*

4.1.3. Combining unique event polymorphisms and STRs

Utilising both UEPs and STRs gives greater resolution of Y chromosome relationships than utilising each system independently. For example, Underhill *et al. (1996)* using the polymorphic C → T transition at the DYS199 locus in Native American populations, subdivided the population into those individuals with the T allele and those without the T allele (see **section 3.1**). Analysing polymorphisms at the DYS19 STR locus revealed significant linkage disequilibrium between the T allele and the DYS19 186 bp allele. This suggested a recent single origin for Native American populations with recent European and African admixture. Analysis of the T / 186 bp linkage estimated that the C → T mutation occurred approximately 30,000 *YBP*.

The two most frequently employed systems in this approach have been the YAP element and the DYS19 STR locus. These loci are easily assayable and were two of the first Y chromosome polymorphisms to be identified. Some initial studies employing both types of markers were not very informative *(Santos et al. 1996, Ruiz-Linares et al. 1996).* Ruiz-Linares *et al. (1996)* used two UEPs (including the YAP element) and three STRs (including DYS19) to assess phylogenetic relationships in a global sample. UPGMA and neighbour-joining trees separated Africans from non-Africans, but demonstrated no further structuring. Santos *et al. (1996)* also employed the YAP element and the DYS19 locus (along with an alpha-heteroduplex system) in a study of global populations. They detected significant differences in the geographic distributions of the haplotypes and higher genetic

diversities in Asians than in either Africans or Caucasians. A very crude estimate of the relative antiquity of the haplotypes was determined from comparison of DYS19 diversity. This idea has more recently been successfully developed and used in numerous studies to make good estimates about the diversity within lineages and the dates of origin of chromosomal lineages. More recently the same authors have used a genealogical approach in the determination of past paternal ancestry in Native Americans. A much greater resolution was determined and demonstrates the increasing success of using such markers *(Santos et al. 1999, Ruiz-Linares et al 1999).*

Jobling *et al. (1997)* used UEPs to distinguish between European, Indian and Sri Lankan chromosomes. Haplogroup 1 was found in 47% of the European sample and 22% of the Indian and Sri Lankan sample. Microsatellite diversity estimates within hg 1 chromosomes in these populations demonstrated a greater diversity in Indian and Sri Lankan samples than Europeans. This suggested that European hg 1 diversity has arisen more recently than the diversity in hg 1 chromosomes of Indian and Sri Lankan origin.

Thomas *et al. (1998)* used a similar approach in a study of Cohanim and Levite Jews. Employing 6 UEPs and 6 STRs they determined genetic diversities within different groups of Jews. They found that the Levite Jews were more diverse than Cohanim Jews, having significant numbers of different chromosomes within three separate Y chromosome lineages. Levites are thought to all be descendents of the tribe of Levi, of which Moses was a member. Genetic data, however, suggest that contemporary Levites are not all direct patrilineal descendants of one paternally related tribal group as believed. Assessment of the diversity within chromosome lineages using the ASD method of dating *(Goldstein et al 1995)* estimated that the time to the most recent common Cohen ancestor was about 3,000 *YBP.*

Zerjal *et al. (1997)* and Hurles *et al. (1998)* have clearly demonstrated the advantages of using a system combining information from UEPs and STRs. Zerjal *et al. (1997)* identified a T → C transition prominent in Asians and Northern Europeans (see **section 3.1**). In order to obtain a better understanding of the origin of this mutation they analysed 10 STR loci in a sample of 60 C allele chromosomes and a small sample of 7 T allele chromosomes. By constructing a network of the STR haplotypes they identified the ancestral haplotype in a Mongolian sample. This suggested that the T → C mutation occurred somewhere in Mongolia. In order to date the origin of the mutation, they employed two different dating methods, the variance method *(Goldstein et al 1996)* and

the Bertranpetit and Calafell method *(Bertranpetit and Calafell 1996).* The diversity within the T allele chromosomes was estimated to have originated between 2,000 and 4,000 *YBP.*

Hurles *et al. (1998)* employed eight UEPs, seven STR loci and one minisatellite locus (MSY1) to distinguish between Polynesian paternal influences. In chromosomes of Papua New Guinean, Polynesian and Melanesian origin, three discrete chromosome lineages were detected, two of which were represented in 55% of all Polynesian samples. Assessment of the STR and MSY1 diversity within these lineages estimated that the time of origin of the majority (64%) haplogroup (hg 24) in Papua New Guinea originated ~4,400 *YBP*, and probably originated within Papua New Guinea itself. This predates the time estimated for the migration into Polynesia as determined by archaeology. However, hg 24 chromosomes are not detected in Polynesian samples. It has been shown that the majority of Polynesian Y chromosomes can be traced to a Melanesian origin, but that there has been substantial European admixture, with 33% of Polynesian chromosomes seemingly originating in Europe. This is in contrast to mtDNA studies *(Richards et al. 1996)* and emphasises the potential power of using not only different genetic systems, but also demonstrates the advantage of using a genealogical approach for the determination of past paternal population histories.

4.2 Materials and Methods

4.2.1. Samples

The same samples and population classifications as in **chapter 3** were used.

4.2.2.Y chromosome microsatellite loci

Primer pair sequences for 6 Y STR loci were as in Kayser *et al. (1997)* and de Knijff *et al. (1997),* and are shown in **table 4.1.** Five were tetranucleotide repeats (DYS19, DYS389-1, DYS390, DYS391 and DYS393) and one was a trinucleotide repeat (DYS392). Microsatellite genotyping was performed as outlined in **section 2.2. Table 4.1** details the PCR conditions for the 6 STR loci used.

Locus	Repeated motif*	Size range of alleles (bp)*	PCR primer sequence (5' - 3')*	Locus GDB-ID*	$MgCl_2$ mM	Annealing temp. °C
DYS19	CTAT/C	174-210	cta ctg agt ttc tgt tat agt atg gca tgt agt gag gac a	G00-121-409	1.5	56
DYS389-1	CTG/AT	239-263	cca act ctc atc tgt att atc tat tct tat ctc cac cca cca ga	G00-366-108	2.0	58
DYS390	CTG/AT	191-227	tat att tta cac att ttt ggg cc tga cag taa aat gaa cac att gc	G00-366-115	1.5	58
DYS391	CTAT	275-295	cta ttc att caa tca tac acc ca gat tct ttg tgg tgg gtc tg	G00-366-118	2.0	58
DYS392	ATT	236-263	tca tta atc tag ctt tta aaa aca a aga ccc agt tga tgc aat gt	G00-456-509	1.75	55
DYS393	GATA	108-132	gtg gtc ttc tac ttg tgt caa tca aac tca agt cca aaa aat gag g	G00-456-649	1.5	58

Table 4.1: Primer sequences and PCR conditions for 6 Y chromosome STR loci. Primer sequences from Kayser *et al. (1997)* and de Knijff *et al. (1997).*

4.3 Results

4.3.1. Allele frequencies

A total of 32 alleles were detected in the 6 Y chromosome STR loci in the Irish, English, Scottish, Norman/Norse, Asian, Turkish, Danish and Togolese populations. The mean number of alleles at each locus was therefore 5.33. The number of alleles at each locus ranged from 4 repeat variants at DYS391 to 6 repeat variants at each of DYS19, DYS390 and DYS392. The frequency of each allele in each population is given in **table 4.2.** The most frequent alleles (MFAs) are not shared between all populations for all loci.

At DYS19 the MFA for the Irish[1], Danish, English, Scottish, Norman/Norse and the Turkish populations (190 bp) was the same, but allele 194 bp was the MFA at this locus for the African population. Allele 190 bp was found at a very low frequency (0.054) in the African population but was found at frequencies of 0.839, 0.583, 0.750, 0.824, 0.833 and 0.460 respectively in the other populations. Similarly, whereas allele 194 bp was found at 0.568 in the African population it was found at lower frequencies in all of the other populations (0.136, 0.083, 0.175, 0.176, 0.111 and 0.160 respectively). Allele 190 bp therefore is predominantly a Caucasian allele demonstrating high frequencies in European populations and low frequencies in the African population.

At DYS389-1 the MFA (251 bp) was shared between all populations (see **table 4.2** for frequencies).

The MFA at DYS390 in each of the diverse populations was different. Allele 203 bp was found most frequently in the Togolese population (0.667) but was absent in the Turkish, English and Norman/Norse populations and rare in the Irish population (0.008). Allele 203 bp was found in the Scottish sample at a frequency of 0.059. Allele 211 bp was the MFA in the Turkish, English and Danish populations (0.480, 0.375 and 0.417 respectively) but was absent in the Togolese population and less frequent in Ireland (0.202). Allele 215 bp was the MFA found in Ireland and the Scottish and Norman/Norse populations (0.430, 0.294 and 0.500 respectively) and was absent in the African

[1] The Irish population refers to all Gaelic Irish samples which include samples from Ulster, Munster, Leinster and Connaught, unless otherwise specified. Scottish, English, Norman and Norse denote Irish samples with exogenous surnames.

population and rare in the Turkish sample (0.260). Allele 219 bp was found at a similar frequency to allele 215 bp in the Scottish population.

Allele 283 bp at DYS391 was the most common allele found in the Turkish, English, Norman/Norse and Togolese samples (0.660, 0.575, 0.556 and 0.811 respectively). Allele 287 bp was the MFA found in Ireland, Scotland and Denmark and was sampled at a frequency of 0.593 in Ireland, 0.588 in Scotland, 0.280 in Turkey, 0.162 in Togo and 0.500 in Denmark.

DYS392 also demonstrated variable frequencies of alleles between populations. Allele 248 bp was found at the highest frequency in the Turkish, Scottish and African populations (0.780, 0.353 and 0.892 respectively) but was rare in the Irish population (0.067) and uncommon in the Norman/Norse and English populations (0.111 and 0.300 respectively). Allele 254 bp was the most common in Ireland and in the Norman/Norse population (0.554 and 0.611 respectively). Both of these alleles (248 bp and 254 bp) were found at frequencies of 0.417 in the Danish population. The intermediate allele between these common alleles (251 bp) was found rarely in all populations with frequencies of 0.063, 0.020, 0.054, 0.167, 0.075 and 0.118 in the Irish, Turkish, Togolese, Danish, English and Scottish populations respectively. Allele 251 bp was absent in the Norman/Norse sample.

Allele 124 bp at DYS393 was the MFA in all populations with the exception of Turkey. Allele 124 bp was found at frequencies of 0.459, 0.333, 0.833, 0.750, 0.889, 0.882 and 0.895 in the Togolese, Danish, English, Norman/Norse, Scottish and Irish populations respectively. Allele 120 bp was the MFA in the Turkish population with a frequency of 0.460.

The allele frequency distributions of all 6 loci in all of the populations (except Denmark n = 12 and Asia n = 6) are illustrated in **figures 4.1 - 4.6.** All of the Y STR loci, with the exception of DYS392, demonstrate clear unimodal distributions with one common allele and less common alleles differing by a series of single repeat units. DYS392 shows an uneven distribution with alleles 248 bp and 254 bp most common in the West European populations whereas allele 251 bp is rare. Overall, the allele frequency distributions support conformation of a stepwise mutation model for STR loci.

Locus	Allele size (bp)	Africa (37)	Asia (6)	Denmark (12)	England (40)	Norman / Norse (18)	Scotland (17)	Turkey (50)	Ulster (22)	Connaught (57)	Leinster (30)	Munster (37)	Ireland Gaelic (106)
	182	0	0	0	0	0	0	0	0	0	0.033	0	0.008
	186	0.027	0	0.083	0.050	0.056	0	0.280	0.000	0.018	0.033	0.000	0.013
DYS19	190	0.054	0	**0.583**	**0.750**	**0.833**	**0.824**	**0.460**	**0.864**	**0.807**	**0.767**	**0.919**	**0.839**
	194	**0.568**	**0.500**	0.083	0.175	0.111	0.176	0.160	0.136	0.158	0.167	0.081	0.136
	198	0.243	0.333	0.167	0.025	0	0	0	0.000	0.100	0.000	0.018	0.029
	202	0.108	0.167	0.083	0	0	0	0	0	0	0	0	0
	247	0.216	0.167	0.333	0.175	0.111	0.235	0.180	0.136	0.018	0.133	0.108	0.099
	251	**0.676**	**0.833**	**0.667**	**0.700**	**0.722**	**0.588**	**0.700**	**0.773**	**0.737**	**0.633**	**0.784**	**0.732**
DYS389-1	255	0.108	0	0	0.125	0.111	0.176	0.100	0.091	0.228	0.233	0.108	0.165
	259	0	0	0	0	0	0	0.056	0	0	0	0.018	0.004
	263	0	0	0	0	0	0	0	0	0.020	0	0	0.005
	203	**0.667**	0	0	0	0	0.059	0	0	0	0.033	0	0.008
	207	0.303	**0.333**	0	0.100	0.056	0.118	0.140	0.045	0.035	0.033	0.027	0.035
DYS390	211	0	0.167	**0.417**	**0.375**	0.389	0.235	**0.480**	0.227	0.158	0.233	0.189	0.202
	215	0	**0.333**	0.333	0.350	**0.500**	**0.294**	0.260	0.227	**0.386**	**0.567**	**0.541**	**0.430**
	219	0.030	0	0.250	0.150	0.056	**0.294**	0.100	**0.500**	**0.386**	0.100	0.216	0.301
	223	0	0.167	0.000	0.025	0	0	0.020	0	0.035	0.033	0.027	0.024
	279	0.027	0	0.083	0.025	0.389	0	0.020	0	0	0	0.027	0.007
DYS391	283	**0.811**	**1.000**	0.333	**0.575**	**0.556**	0.412	**0.660**	0.364	0.316	**0.500**	0.405	0.396
	287	0.162	0	**0.500**	0.375	0.056	**0.588**	0.280	**0.636**	**0.667**	**0.500**	**0.568**	**0.593**
	291	0	0	0.083	0.025	0	0	0.040	0	0.018	0	0	0.004
	245	0.027	0	0	0	0	0	0.040	0	0	0	0	0
	248	**0.892**	**0.500**	**0.417**	0.300	0.111	**0.353**	**0.780**	0.091	0.018	0.133	0.027	0.067
DYS392	251	0.054	0	0.167	0.075	0	0.118	0.020	0.091	0.035	0.100	0.027	0.063
	254	0.027	0.333	**0.417**	**0.575**	**0.611**	0.294	0.040	0.318	**0.509**	**0.633**	**0.757**	**0.554**
	257	0	0.167	0	0.050	0.278	0.235	0.040	**0.409**	0.421	0.133	0.189	0.288
	260	0	0	0	0	0	0	0.020	0.091	0.018	0	0	0.045
	116	0	0	0	0	0	0	0.040	0	0	0	0	0
	120	0.135	**0.333**	0.083	0.025	0	0	**0.460**	0.045	0.053	0.033	0	0.033
DYS393	124	**0.459**	**0.333**	**0.833**	**0.750**	**0.889**	**0.882**	0.440	**0.864**	**0.877**	**0.867**	**0.973**	**0.895**
	128	0.297	0.167	0	0.175	0.111	0	0.060	0.045	0.053	0.067	0	0.041
	132	0.108	0.167	0.083	0.050	0	0.118	0	0.045	0.018	0.033	0.027	0.031

Table 4.2: Allele frequencies at 6 Y STR loci in a number of geographically diverse populations

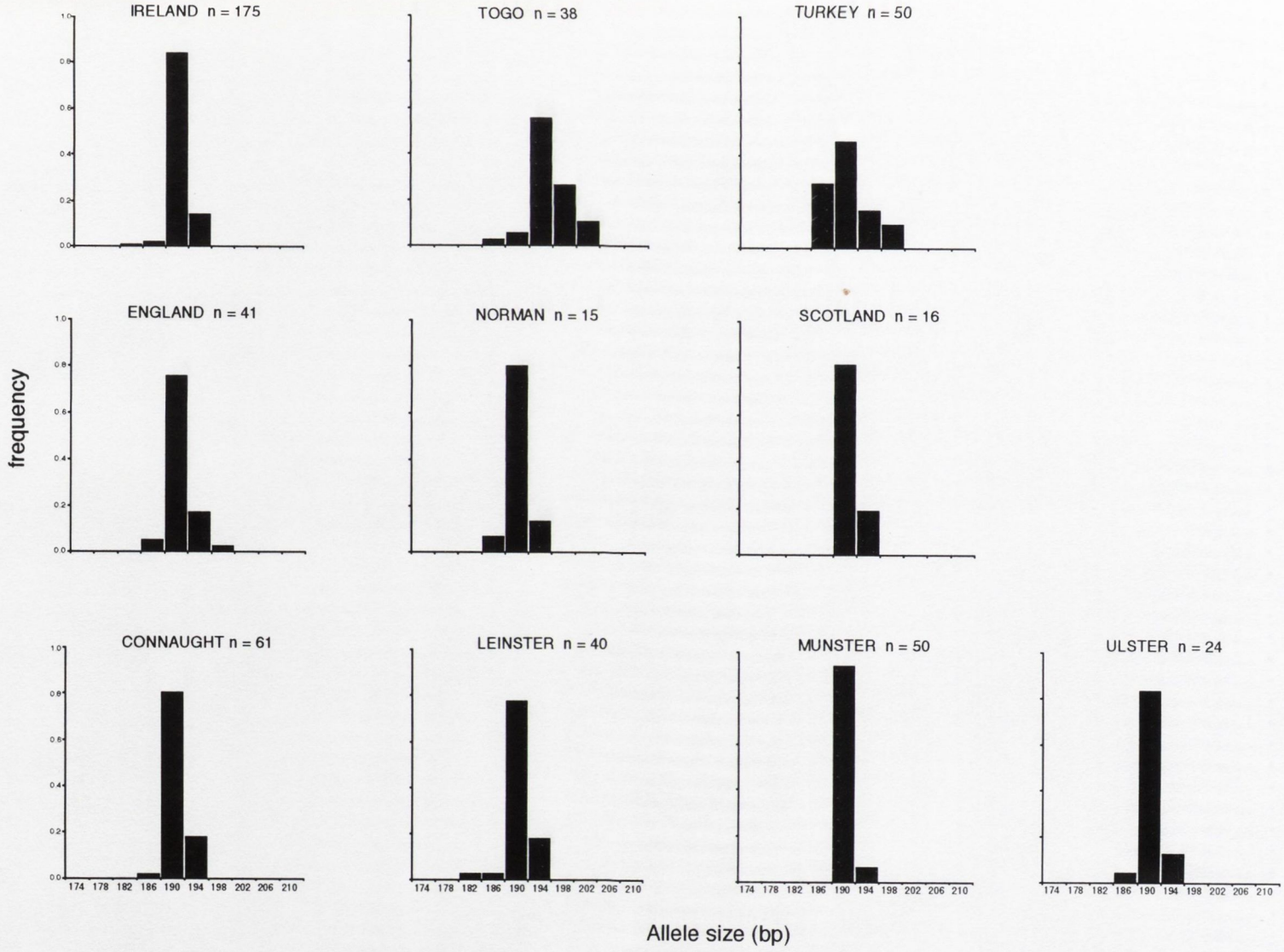


Figure 4.1: Allele frequency distributions at Y chromosome microsatellite locus DYS19.

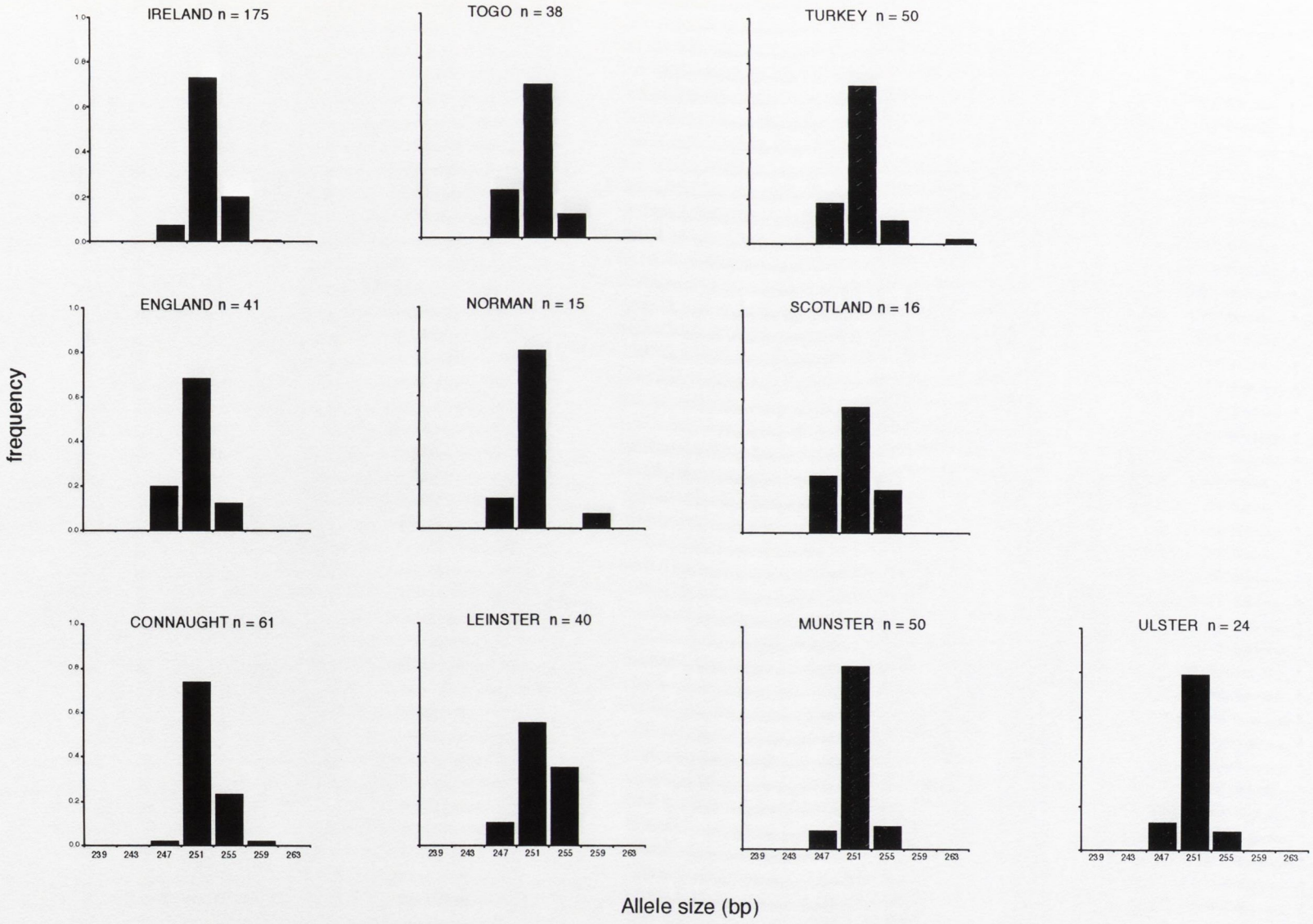


Figure 4.2: Allele frequency distributions at Y chromosome microsatellite locus DYS389-1.

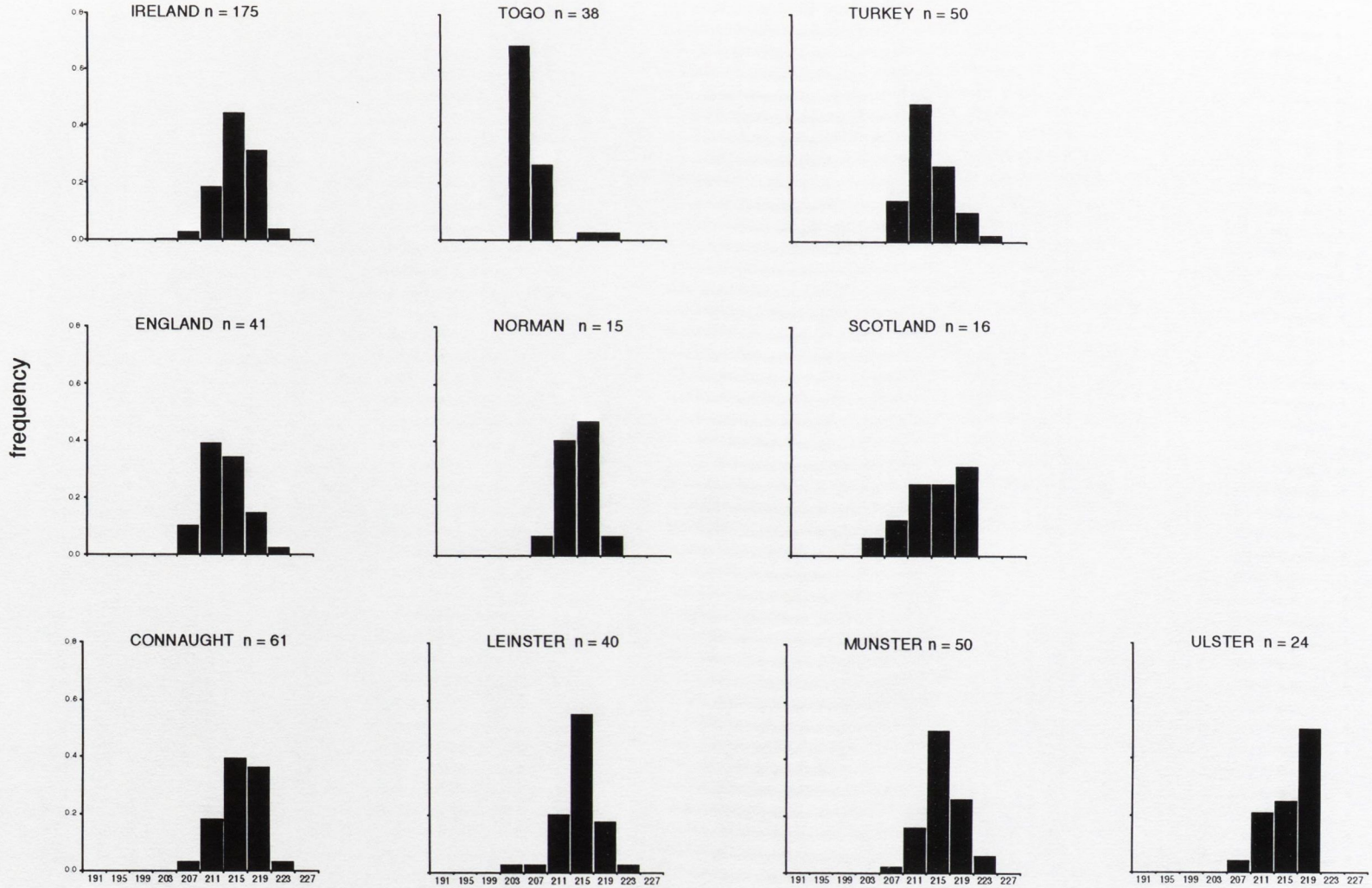


Figure 4.3: Allele frequency distributions at Y chromosome microsatellite locus DYS390.

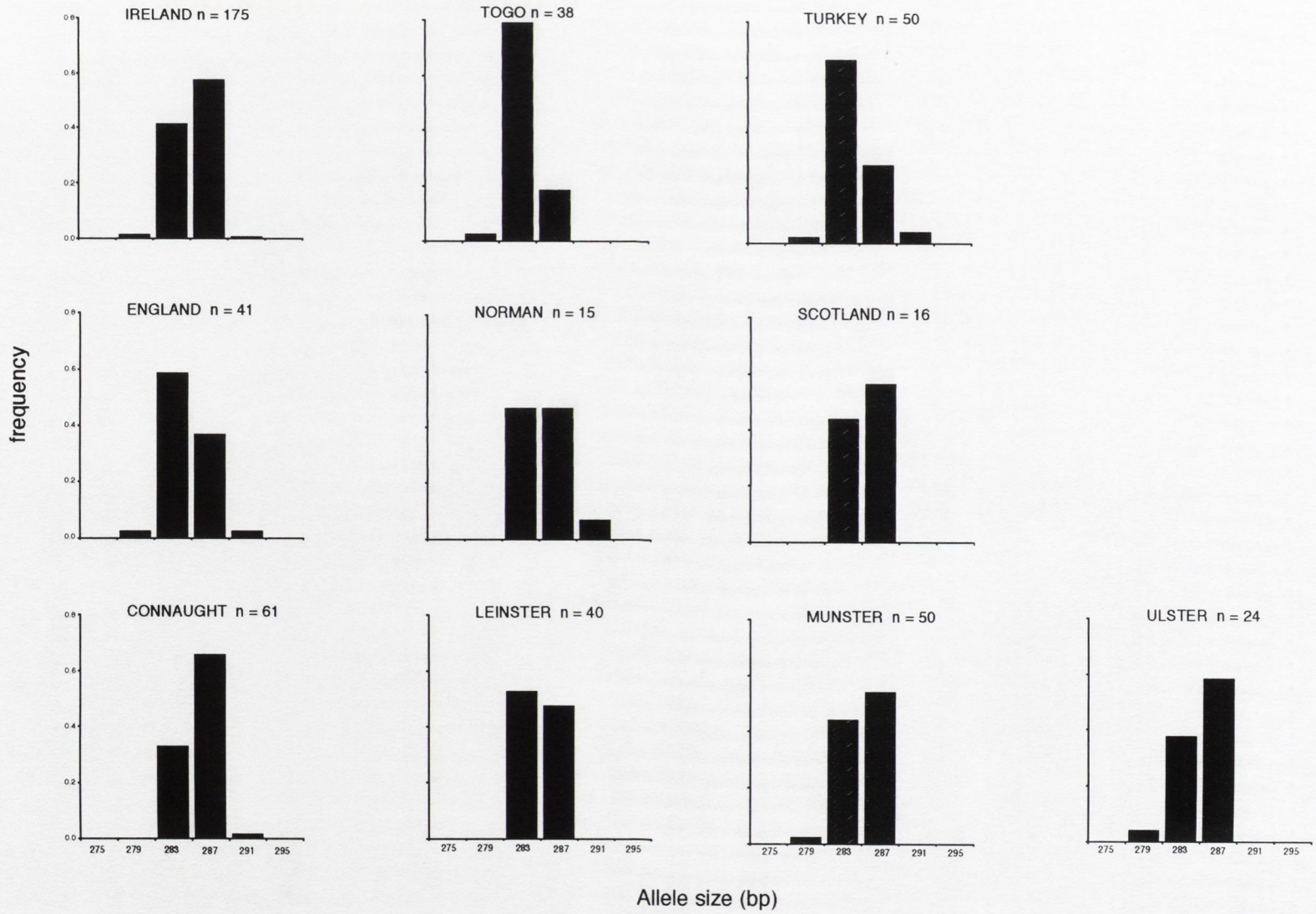


Figure 4.4: Allele frequency distributions at Y chromosome microsatellite locus DYS391.

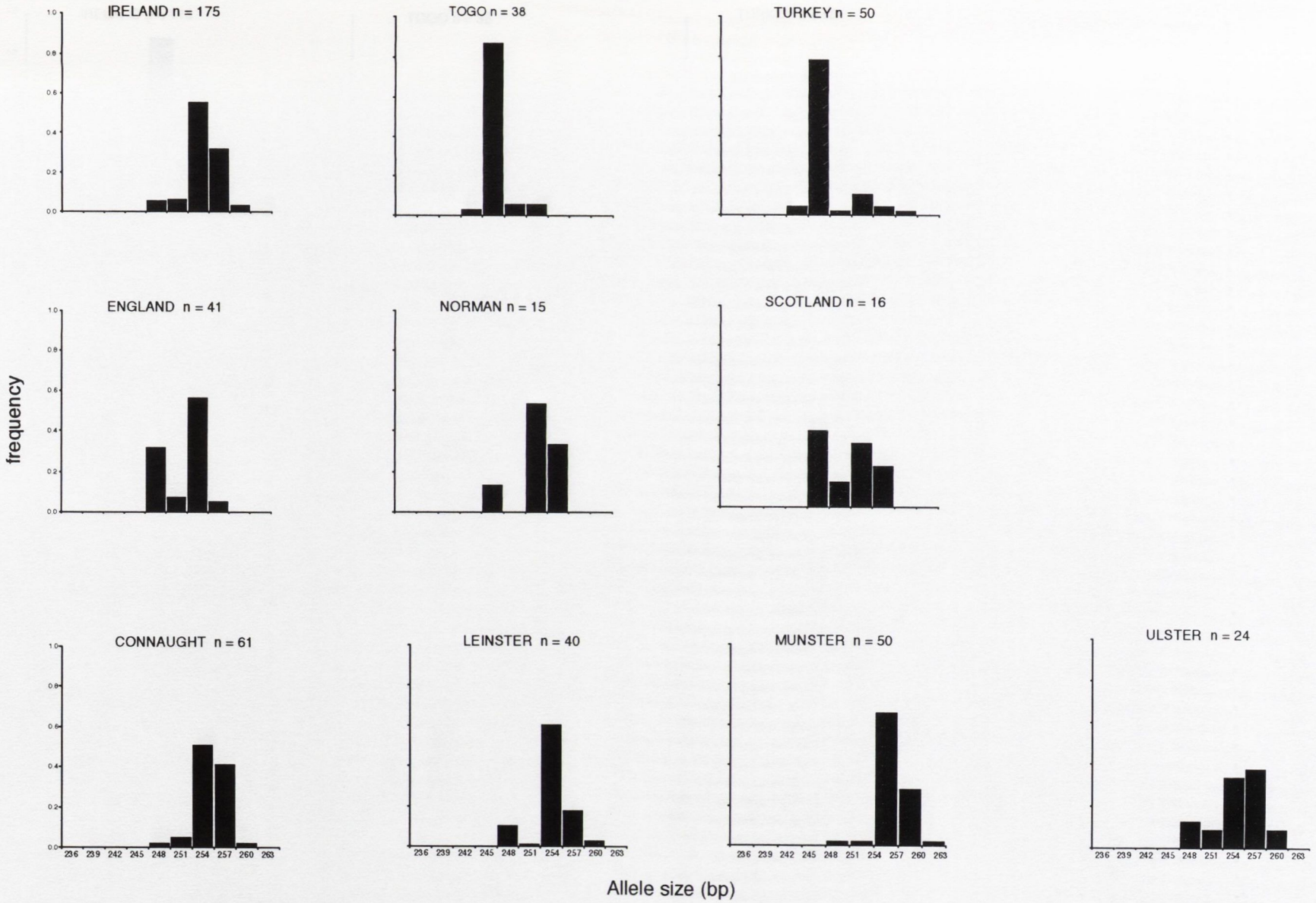


Figure 4.5: Allele frequency distributions at Y chromosome microsatellite locus DYS392.

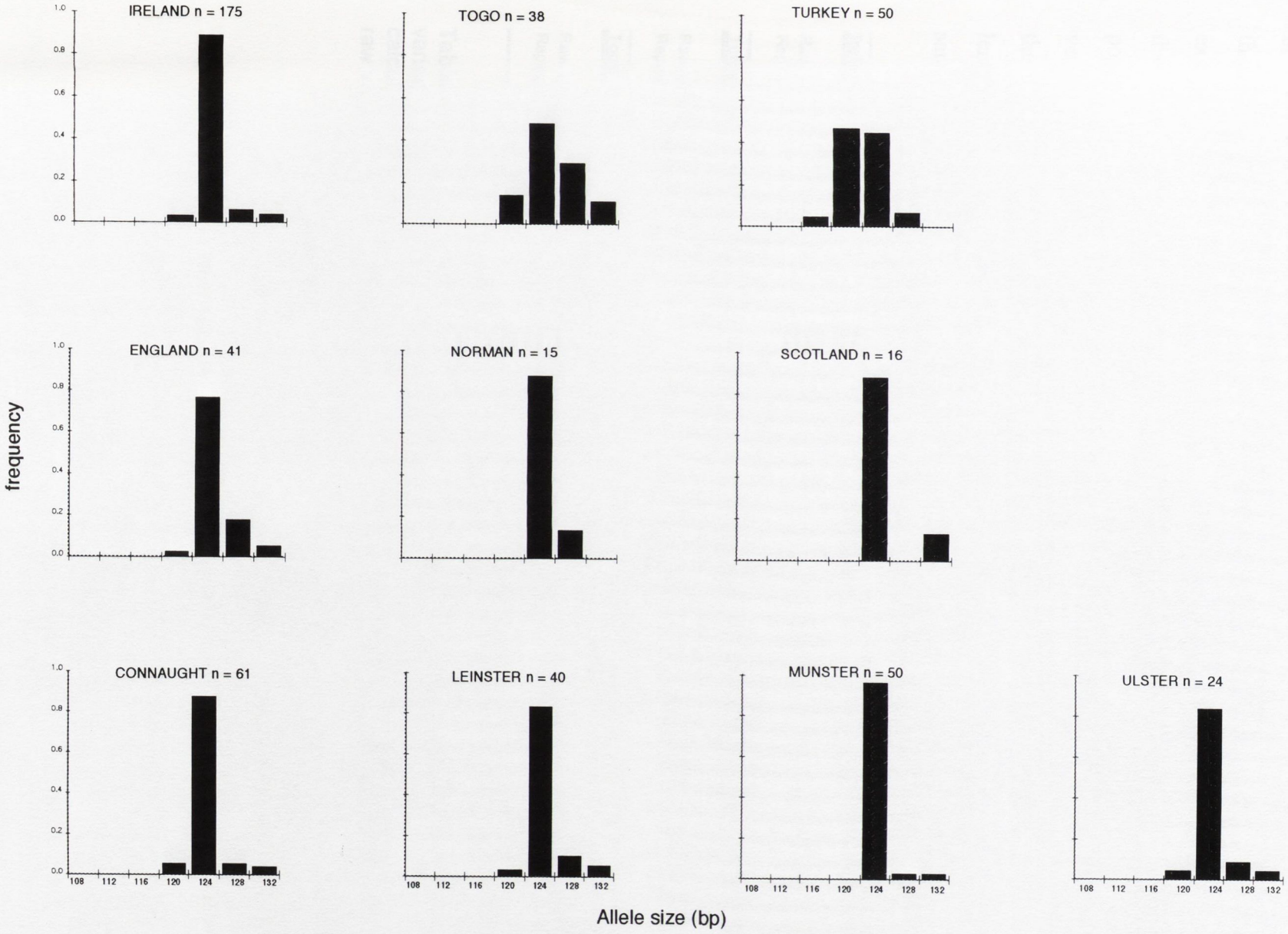


Figure 4.6: Allele frequency distributions at Y chromosome microsatellite locus DYS393.

4.3.2. Microsatellite allele length variance

Allele length variance, an indication of genetic diversity, at each of the 6 STR loci in each of the Irish, Turkish and Togolese populations, is given in **table 4.3.** Surprisingly, the overall mean variance is highest in the Turkish population and not, as might be expected, in the Togolese population. At 4 of the 6 loci, the highest variance is found in the Turkish population. At DYS393 the highest variance is found in the African population. At DYS390 the highest variance is found in the Irish population. The variance of repeat number in the African population averaged across all loci is 0.4560, in the Turkish population 0.6664 and in Ireland 0.4816. The overall variance in allele lengths is therefore higher in both the Irish and the Turkish samples than in the Togolese sample.

Ireland							
	DYS19	**DYS389-1**	**DYS390**	**DYS391**	**DYS392**	**DYS393**	**Mean variance**
Raw variance	3.501	5.1605	14.0428	4.8638	8.3623	3.7977	6.6213 nucleotides
Repeat variance	0.2188	0.3225	0.8777	0.3040	0.9291	0.2374	0.4816 repeat units
	tetra	tetra	tetra	tetra	tri	tetra	
Turkey							
	DYS19	**DYS389-1**	**DYS390**	**DYS391**	**DYS392**	**DYS393**	**Mean variance**
Raw variance	13.6098	7.5037	13.6424	5.6229	9.1433	7.3404	9.4771 nucleotides
Repeat variance	0.8506	0.469	0.8527	0.3514	1.0159	0.4588	0.6664 repeat units
	tetra	tetra	tetra	tetra	tri	tetra	
Togo							
	DYS19	**DYS389-1**	**DYS390**	**DYS391**	**DYS392**	**DYS393**	**Mean variance**
Raw variance	11.7477	5.411	9.2012	2.8108	1.6892	11.8679	7.0763 nucleotides
Repeat variance	0.7342	0.3213	0.5751	0.1757	0.01877	0.7417	0.4560 repeat units
	tetra	tetra	tetra	tetra	tri	tetra	

Table 4.3: Allele length variance statistics for 6 Y STR loci in 3 populations. The raw variance is calculated as the total variance at each locus, the repeat variance is calculated as the variance of the repeat lengths (ie.for trinucleotides, repeat variance = raw variance/3^2 and for tetranucleotides, repeat variance = raw variance/4^2)

4.3.3. Microsatellite haplotypes and compound haplotypes

A total of 142 Y 6-locus STR haplotypes were observed on the 326 chromosomes of African, Asian, Irish, Danish and Turkish origin. A small number of these were shared

Ht	DYS19	DYS389-1	DYS390	DYS391	DYS392	DYS393	hg
1	186	251	207	279	248	120	21
2	190	255	207	279	248	120	2
3	198	251	211	279	248	120	2
4	186	251	219	279	248	128	21
5	190	251	215	279	257	124	1
6	194	251	203	283	245	124	8
7	194	255	211	283	245	128	1
8	198	247	207	283	245	128	2
9	186	247	219	283	248	120	2
10a	190	247	207	283	248	120	21
10b	190	247	207	283	248	120	2
11	190	247	211	283	248	120	2
12	190	251	207	283	248	120	2
13	190	255	215	283	248	120	2
14	194	247	207	283	248	120	21
15	194	247	215	283	248	120	2
16	194	251	215	283	248	120	2
17	198	251	211	283	248	120	2
18	186	247	215	283	248	124	21
19	186	251	215	283	248	124	21
20	186	251	219	283	248	124	21
21	190	247	207	283	248	124	2
22	190	247	211	283	248	124	2
23	190	251	211	283	248	124	2
24	190	251	215	283	248	124	21
25	194	247	203	283	248	124	8
26	194	247	207	283	248	124	2
27	194	247	211	283	248	124	2
28	194	251	203	283	248	124	8
29	194	251	207	283	248	124	8
30	194	251	223	283	248	124	3
31	194	255	203	283	248	124	8
32	194	255	207	283	248	124	8
33	194	255	215	283	248	124	2
34	198	247	203	283	248	124	8
35	198	247	215	283	248	124	2
36	198	251	203	283	248	124	8
37	198	263	223	283	248	124	3
38	186	251	215	283	248	128	21
39	190	247	207	283	248	128	2
40	190	251	207	283	248	128	2
41a	194	247	207	283	248	128	8
41b	194	247	207	283	248	128	2
42	194	251	203	283	248	128	8
43	198	251	203	283	248	128	8
44	202	251	203	283	248	128	8
45	202	251	211	283	248	128	2
46	198	251	203	283	248	132	8
47	198	251	207	283	248	132	2
48	202	251	207	283	248	132	8
49	190	247	211	283	251	124	2
50	194	255	211	283	251	124	2

Table 4.4: Microsatellite haplotypes detected in 326 ethnically diverse individuals.

between populations, but the majority were unique. Ninety-eight haplotypes were unique to individuals and were not shared between or within populations. Of all the 142 haplotypes, therefore, 69% were encountered only once.

Haplotype	DYS19	DYS389-1	DYS390	DYS391	DYS392	DYS393	hg
51	194	251	215	283	251	128	2
52	194	255	211	283	251	128	2
53	198	251	203	283	251	128	8
54	190	255	211	283	251	132	2
55	194	247	211	283	251	132	2
56	194	251	215	283	251	132	2
57	194	255	211	283	251	132	2
58	198	251	203	283	251	132	8
59	194	251	211	283	254	120	1
60	198	247	215	283	254	120	26
61	190	247	215	283	254	124	1
62	190	251	207	283	254	124	1
63	190	251	211	283	254	124	1
64	190	251	215	283	254	124	1
65	190	251	219	283	254	124	1
66	190	255	215	283	254	124	1
67	194	251	203	283	254	124	8
68a	194	251	215	283	254	124	26
68b	194	251	215	283	254	124	1
69	194	255	211	283	254	124	1
70	194	255	215	283	254	124	1
71	198	251	211	283	254	124	26
72	190	251	211	283	254	128	1
73	190	251	215	283	254	128	1
74	194	251	207	283	257	120	26
75	182	255	211	283	257	124	26
76	190	247	207	283	257	124	26
77	190	251	211	283	257	124	12
78	190	251	215	283	257	124	1
79	190	251	219	283	257	124	1
80	190	251	223	283	257	124	1
81	190	255	215	283	257	124	1
82	190	259	219	283	257	124	1
83	198	251	215	283	260	116	26
84	190	255	219	283	260	124	1
85	190	251	211	287	248	120	2
86	194	247	207	287	248	120	21
87	186	251	211	287	248	124	21
88	186	251	215	287	248	124	21
89	190	247	203	287	248	124	2
90	190	247	211	287	248	124	2
91	190	251	211	287	248	124	2
92	190	251	215	287	248	124	2
93	190	251	219	287	248	124	21
94	194	251	203	287	248	124	8
95	194	251	211	287	248	124	2
96	194	251	219	287	248	124	3
97	194	255	203	287	248	124	8
98	194	255	211	287	248	124	2
99	202	251	219	287	248	124	3
100	194	251	203	287	248	128	8

Table 4.4 *cont.:* Microsatellite haplotypes detected in 326 ethnically diverse individuals.

When the SNP haplotypes (see **chapter 3**) were added to the STR haplotypes to produce compound haplotypes 3 new haplotypes emerged. Haplotype (ht) 10, ht 41 and ht 68 were found to be associated with more than one haplogroup. Haplotype 10 was found on both a hg 21 and a hg 2 chromosome. Haplogroup 21 and hg 2 are separated by 2 mutational events (see **figure 3.2**) and therefore ht 10a and ht 10b, although they share the same microsatellite haplotype, are separated by at least 2 mutational events. Haplotype 41 was found on a hg 8 and on a hg 2 chromosome and therefore ht 41a and ht 41b are separated by three at least evolutionary events. Haplotype 68 was found on a hg 26 and

Ht	DYS19	DYS389-1	DYS390	DYS391	DYS392	DYS393	hg
101	190	251	219	287	251	124	1
102	194	255	215	287	251	124	2
103	190	255	215	287	251	128	1
104	194	247	211	287	251	132	2
105	194	255	211	287	251	132	2
106	194	251	211	287	254	116	26
107	190	251	211	287	254	120	1
108	190	251	215	287	254	120	1
109	186	255	211	287	254	124	1
110	190	247	211	287	254	124	1
111	190	247	215	287	254	124	1
112	190	251	203	287	254	124	1
113	194	255	215	287	254	128	1
114	190	251	211	287	254	124	1
115	190	251	215	287	254	124	1
116	190	251	219	287	254	124	1
117	190	255	207	287	254	124	1
118	190	255	215	287	254	124	1
119	190	255	219	287	254	124	1
120	190	259	215	287	254	124	1
121	194	251	211	287	254	124	1
122	198	251	215	287	254	124	1
123	194	255	215	287	254	128	1
124	194	255	211	287	254	132	1
125	190	247	219	287	257	124	1
126	190	251	211	287	257	124	1
127	190	251	215	287	257	124	1
128	190	251	219	287	257	124	1
129	190	251	223	287	257	124	1
130	190	255	215	287	257	124	1
131	194	251	219	287	257	124	1
132	190	251	219	287	257	128	1
133	194	251	211	287	257	128	16
134	190	251	211	287	260	124	1
135	190	251	219	287	260	124	1
136	198	251	219	291	248	124	3
137	190	251	207	291	254	120	1
138	186	247	211	291	254	124	1
139	190	247	215	291	254	124	1
140	190	251	215	291	254	128	1
141	190	251	219	291	257	124	1
142	190	251	211	283	248	120	2

Table 4.4 ***cont:*** Microsatellite haplotypes detected in 326 ethnically diverse individuals.

also on a hg 1 chromosome and therefore ht 68a and ht 68b are separated by a minimum of one evolutionary event. That STR haplotypes can arise on divergent Y chromosome lineages demonstrates the convergent nature of microsatellite evolution.

4.3.4. Compound haplotypes in each population

The absolute frequencies of the Y chromosome STR microsatellite haplotypes sampled in each population are given in **table 4.5**.

Togo

In the Togolese population 25 haplotypes were found in 37 individuals. Sixteen haplotypes were sampled only once in the population. The most common haplotype was ht 28 (10.8%). Haplotype 10a (associated with hg 21) and ht 41a (associated with hg 8) were both present in this population. No African haplotypes were found in any of the other global populations.

Turkey

The Turkish population contained 33 haplotypes in a sample of 50 individuals. Of the 33 haplotypes 26 were sampled only once. The most common haplotype was ht 12 (14%). Haplotype 12 differs from the most common Togolese haplotype, ht 28, by 1 mutational step at each of 3 STR loci (DYS19, DYS390 and DYS393) and by 3 UEPs. Haplotype 12 was not found in any of the other populations.

Ireland

In the Gaelic Irish population 58 haplotypes were detected in 146 individuals. Of the 58 haplotypes 17 were sampled more than once. Of the remaining haplotypes, 33 were found to be unique to individuals within the popualtion. Of the haplotypes detected only once in the Irish population but present also in other populations, 3 were shared with Turkey, 5 were shared with England and 2 were shared with Scotland.

haplotype	Ulster 22	Munster 37	Leinster 30	Connaught 57	England 40	Scotland 17	Norman 18	Norse 3	Denmark 12	Turkey 50	Togo 37	Asia 6
1											1	
2										1		
3									1			
4					1							
5		1										
6											1	
7										1		
8										1		
9										1		
10a											1	
10b	1		1									
11										1		
12										7		
13										2		
14											2	
15										1		
16										1		
17										1		
18									1			
19			1									
20										2		
21					1	2						
22		1			2	1				1		
23	1									1		
24						1						
25											1	
26				1						1		
27			1		2							
28											4	
29											1	
30												1
31											2	
32											1	
33					1							
34											2	
35					1							
36											1	
37										1		
38					1					1		
39					1							
40					1							
41a											1	
41b							1					
42											2	
43											3	
44											2	
45												1
46											1	
47												1
48											2	
49									1			
50										1		
51					1							
52	1		1									
53											1	
54					1							
55	1											
56			1									
57		1				2						
58											1	
59										1		
60												1
61		1										
62					1							
63		1	2	2	2		2		1			
64		7	2	6	4	1	1		1			
65	1				1							
66		1	2		1							
67											1	
68a												1
68b		1										
69				3								
70				1								
71										1		
72					1							
73			1									
74												1
75			1									
76										1		
77										1		
78	2						2					
79	1	2	1	3								
80			1									
81				1								
82							1					
83										1		
84				1								
85										5		

Table 4.5: Absolute frequencies of haplotypes sampled in each population. The most common haplotypes in each population are shown in bold type.

haplotype	Ulster	Munster	Leinster	Connaught	England	Scotland	Norman	Norse	Denmark	Turkey	Togo	Asia
90									1			
91										1		
94											1	
95										1		
96						1						
97											1	
98			1									
99									1			
100											2	
101				1								
102			1									
103				1								
104									1			
105					1							
106										1		
107				1								
108				2								
109				1								
110	1											
111		1	2									
112			1									
113		1										
114		3	1	1	5	1	1	1				
115	2	6	6	4	2	1	3		2			
116	2	4	1	1	2	1			1			
117				1								
118		1	1	2		1		2				
119	1				1							
120				1								
121		1										
122				1	2							
123				1								
124				1								
125		1										
126							1					
127	1			2		1						
128	4	1	1	12	1	3						
129		1		2	1							
130		1										
131	1			2								
132				1								
133							1					
134	1											
135	1											
136									1			
137										1		
138										1		
139							1					
140					1							
141				1								
142					1							

Table 4.5 *cont*:: Absolute frequencies of haplotypes sampled in each population. The most common haplotypes in each population are shown in bold type.

The two equally most frequent haplotypes present on Gaelic Irish Y chromosomes were ht 115 and ht 128. Haplotype 115 and ht 128 differ by one mutational step at each of DYS390 and DYS392 and are therefore 2 mutational steps different from each other. The next most common haplotype in the Irish population, ht 64, differs from ht 115 at DYS391 by one mutational step and from ht 128 by one mutational step at each of DYS390 and DYS392.

Ulster

Sixteen haplotypes were detected in the Ulster sample of 22 individuals. Four of these haplotypes were detected more than once. The most common haplotype was ht 128 (18.2%). Haplotype 64 was absent and ht 115 was sampled only twice.

Munster

Twenty haplotypes were detected in the Munster sample of 37 individuals. Five of these haplotypes were detected more than once. The most common haplotype was ht 64 (18.9%). Haplotype 115 was also common, but ht 128 was uncommon.

Leinster

In Leinster, 21 haplotypes were detected in a sample of 30 individuals. Five of these haplotypes were detected more than once. The most common haplotype was ht 115 (20%). Neither ht 64 nor ht 128 were sampled frequently in this population. Haplotype 10b was found on a hg 2 chromosome.

Connaught

In Connaught, 28 haplotypes were detected in a sample of 57 individuals. Eleven of these haplotypes were detected more than once. The most common was ht 128 (21.1%). Haplotype 64, ht 128 and ht 115 were each sampled more than 4 times in the population.

England

In the English sample 27 haplotypes were found in 40 individuals. Nineteen haplotypes were sampled only once. The most common haplotype was ht 114 (12.5%), which differs from the common Irish haplotype, ht 115, by one mutational step at DYS390. Only ht 22 was shared in both the English and the Turkish samples. All other shared haplotypes were shared with Northwest European populations only.

Scotland

In the Scottish sample 13 haplotypes were found in 17 individuals. Ten were sampled only once. The most common haplotype was ht 128 (17.6%). No Scottish

haplotype was found in the Danish or the Turkish sample. Haplotypes ht 128, ht 22 and ht 57 were shared with individuals sampled within Ireland only.

Norman and Norse

In the Norman and Norse sample 13 haplotypes were found in 18 individuals. Nine were sampled once. The most common haplotype was ht 115 (20%). No haplotype was shared with the Turkish sample.

Denmark

Eleven haplotypes were found in 12 individuals from Denmark. The most common haplotype which was found twice in this population was ht 115 (16.7%). Seven haplotypes were unique to Denmark whereas the remaining 4 haplotypes were shared with Northwest European populations. No Danish haplotypes were found in the Turkish population.

Asia

The six Asian samples had no haplotypes in common and did not share any haplotypes with any other populations although ht 68a was found in the Asian sample on hg 26. Haplotype 68b was found on a hg 1 chromosome in the Munster sample.

4.3.5. Microsatellite haplotypes in haplogroups

The following describes the partitioning of microsatellite haplotypes into discrete chromosome lineages as defined by slowly evolving unique event polymorphisms (see **chapter 3**).

Haplogroup 1 chromosomes

In total, 55 haplotypes were detected on the 189 hg 1 chromosomes. Of the 55, 36 were sampled only once. Only one of the STR haplotypes, ht 68b, was shared with

another haplogroup, hg 26. The majority (33.9%) of hg 1 chromosomes contained the microsatellite haplotypes ht 115, ht 64 and ht 128. Additionally, ht 63, ht 114 and ht 116 were common. Haplotype 63 differs from ht 64 by one mutational step at DYS390. Haplotype 114 and ht 116 differ from the common ht 115 by one mutational step each at DYS390 and ht 64 and ht 115 differ by one mutational step at DYS391. Therefore the loci delimiting the common haplotypes within hg 1 chromosomes are DYS390 and DYS391. The most common haplotypes share allele 190 bp at DYS19, allele 251 bp at DYS389-1, allele 254 bp at DYS392 and allele 124 bp at DYS393.

Haplgroup 2 chromosomes

On the 68 hg 2 chromosomes 43 haplotypes were detected. Of the 43, 32 were sampled only once. Two of the STR haplotypes, ht 10b and ht 41b, were shared with other haplogroups, namely hg 21 and hg 8 respectively. The most frequent haplotype was ht 12, which was found only in the Turkish population. Other frequent haplotypes were ht 22 and ht 85. Haplotype 22 was found in a number of diverse populations, but ht 85 was found on Turkish chromosomes only. Between ht 12, ht 22 and ht 85 only two loci demonstrate common alleles, these being DYS19 allele 190 bp and DYS392 allele 248 bp. Haplotypes 12 and 85, both found solely in the Turkish population, differ by one mutational step at each of DYS390 and DYS391.

Haplogroup 3 chromosomes

Five haplotypes were detected on the 5 hg 3 chromosomes. Alleles 248 bp at DYS392 and 124 bp at DYS393 were common among all haplotypes. All other loci demonstrated a number of variable alleles on the hg 3 chromosomes.

Haplogroup 8 chromosomes

Twenty-eight haplotypes were detected on the 48 hg 8 chromosomes. Of the 28, 15 were detected only once in the population. The STR haplotype ht 41a was also found on a hg 2 chromosome. The most common haplotypes were ht 19 and ht 28. Between these haplotypes alleles differ at DYS19 and DYS390 by 2 and 3 mutational steps

respectively. Other common haplotypes were ht 43, ht 87 and ht 88. Haplotype 43 is two mutational steps from ht 28 differing at DYS19 and DYS393. Haplotype 88 is one mutational step from ht 19 differing at DYS391, and ht 87 is one mutational step from ht 87 differing at DYS390. Therefore no loci are preferentially involved in the differentiation between common haplotypes on hg 8 chromosomes.

Haplogroup 21 chromosomes

Five haplotypes were detected on the 6 hg 21 chromosomes. Only ht 14 was detected more than once. Haplogroup 21 haplotypes have a predominance of allele 207 bp at DYS390, which is an infrequent allele in other haplogroups, found in only 22 haplotypes in all populations.

Haplogroup 26 chromosomes

Eight haplotypes were detected on the 8 hg 26 chromosomes. Allele 283 bp at DYS391 is the only allele predominant on these haplotypes. At the remaining loci no allele is preferentially associated with hg 26 chromosomes. Haplotype 68a was present on a hg 26 chromosome and also on a hg 1 chromosome.

4.3.6. Haplotypic diversity measures

Gene diversity is given in **table 4.6** for each population. It is defined as the probability that two randomly chosen haplotypes are different in a population and is estimated by the following equation:

$$H = n / n\text{-}1 \ (1 - \Sigma p^2)$$

Haplotypic diversity measures were calculated from STR haplotypes only and did not include SNP haplotypic data.

Population	n	Average gene diversity (H)	Standard Error
Togo	38	0.478	0.286
Connaught	61	0.454	0.271
Denmark	12	0.591	0.364
England	42	0.529	0.310
Leinster	40	0.505	0.298
Munster	50	0.363	0.229
Norman	15	0.476	0.298
Scotland	16	0.546	0.332
Turkey	50	0.555	0.322
Ulster	24	0.489	0.296

Table 4.6: Average gene diversity measures *(Nei 1987)* for six-locus Y STR haplotypes calculated in diverse population groups

Unlike the haplotypic diversity measures for X chromosome microsatellite haplotypes, Y chromosome STR haplotypes did not illustrate a greater variation in the African population compared to the European populations. Instead, the greatest haplotypic diversity was observed in the Danish population, although this may be an artifact of small sample size. In agreement with the allele length variance measures, the Turkish population was found to have a high haplotypic diversity with Y STR loci. The haplotypic diversity measure is also relatively high in the Scottish population, although again this may be a sample size artifact. Within the Gaelic populations, Leinster demonstrated the highest level of variability and Munster the lowest. The mean gene diversity value for global Y chromosomes was estimated as 0.489.

4.3.7. Phylogenetic analysis : populations as OTUs

Phylogenetic trees based on allele frequencies at Y STR loci in geographically diverse population groups have been successful in reconstructing patterns of evolutionary population history *(de Knijff et al 1997; Deka et al. 1996; Lucotte and Hazout 1996).*

Distance based phylogenetic approaches require that the raw allelic data is transformed into a single measure of evolutionary distance as a definition of the similarity between OTUs. The distance measure used for the analysis of X chromosome STR loci, D_{dm} was used to assess genetic relationships between diverse geographical groups from Y STR data. A distance matrix was calculated for population x population comparisons and a neighbor-joining tree was created from these distances.

The resulting tree **(figure 4.7)** demonstrated the following features. First, the clear separation of African and all non-African populations is evident. Long branch lengths at Y STR loci leading to African nodes have rarely been noted *(Deka et al 1996),* however, here the unusually long branch to the African node is evident. Second, the Turkish population is situated at a midpoint between the African and European populations. Third, an unusually long branch separates the Munster node from the main branch.

Ddm	TOGO	DENMARK	ENGLAND	TURKEY	SCOTLAND	NORMAN	ULSTER	MUNSTER	LEINSTER
DENMARK	1.325								
ENGLAND	1.336	0.089							
TURKEY	1.067	0.205	0.265						
SCOTLAND	1.324	0.072	0.005	0.254					
NORMAN	1.703	0.271	0.090	0.552	0.102				
ULSTER	2.335	0.326	0.178	0.739	0.195	0.080			
MUNSTER	2.080	0.345	0.323	0.910	0.335	0.242	0.161		
LEINSTER	1.835	0.246	0.079	0.530	0.095	0.039	0.045	0.197	
CONNAUGH	2.382	0.416	0.245	0.843	0.259	0.098	0.023	0.149	0.055

Table 4.7: Ddm genetic distances between all populations

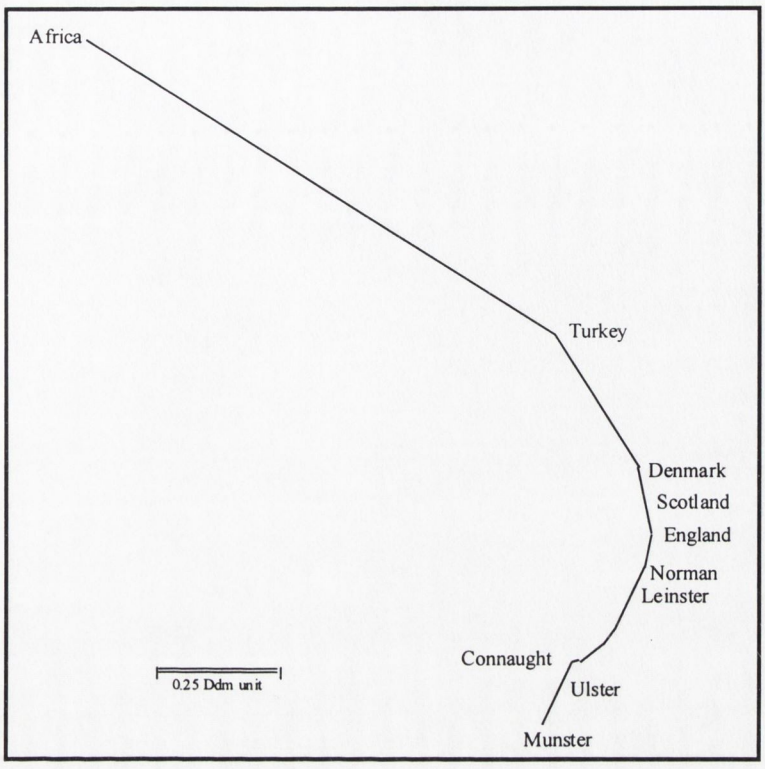


Figure 4.7: Neighbour-joining tree relating diverse population groups using the Ddm genetic distance measure.

4.3.8. Population structure: Analysis of Molecular Variance

A number of tests of population structure were performed using the AMOVA approach constituted in the program Arlequin *(Schneider et al 1997).* In order to assess the variance components within and between populations, the following groupings were made: [Ulster, Munster, Leinster, Connaught] V [Norman, Scotland, England] V [Africa] V [Turkey]. An AMOVA test calculated that 54.35% of human Y chromosome genetic diversity is distributed as differences between individuals within populations and that 45.13% of the total genetic variance accounts for differences between populations.

Significant variance components were achieved in an AMOVA test when population cohorts, separated by surname origins, were compared. Gaelic names ([Ulster, Munster, Leinster and Connaught]) were tested against non-Gaelic names sampled within Ireland ([England, Scotland and Norman]). The results indicated an among group partitioning of molecular variance which, in a permutation test, showed a significant genetic difference between Gaelic and non-Gaelic surname samples ($p<0.05$). Full results are given in **table 4.8.** Significant partitioning was detected using UEP data (outlined in **section 3.3.10**). Additionally, in accordance with previous studies, the majority of genetic variation was detected between individuals within population groups (86.56%) and very little of the total diversity was explained by the difference between haplotypes in diverse populations (12.31%).

Population groups	Ulster; Munster; Leinster; Connaught Scotland; England; Norman Africa Turkey		Ulster; Munster; Leinster; Connaught Scotland; England; Norman	
Source of variation	**% variation**	**p value =**	**% variation**	**p value =**
Among groups	45.13	0.0010	12.31	0.0302
Among populations within groups	0.51	0.1210	1.13	0.1522
Within populations	54.35	0.0000	86.56	0.0000

Table 4.8: Results of AMOVA tests of population structure from 6 Y STR haplotypes. Two separate tests were performed with two different groupings. The first grouping concerned the separation of all the global populations, the samples from Ireland being divided into surnames of Gaelic and non-Gaelic origin. The second grouping concerned the assessment of population substructure within and between populations of Gaelic Irish origin and non-Gaelic Irish origin.

4.3.9. Multidimensional Scaling (MDS)

Multidimensional scaling is a technique that constructs a graphical representation showing the relationship between a number of objects (in this case, population groups), from a matrix of distances calculated between them *(Manly, 1986)*. The objects may be represented in one dimension as points on a line, in two dimensions as points on a plane, or indeed in three dimensions as points in space. Further dimensions can be achieved if the data allows and if the level of analysis requires it.

All populations were examined in one and two dimensions. **Figure 4.8** shows the two dimensional graphical representation of population interrelationships calculated from pairwise Fst genetic distances. **Table 4.9** shows the numerical outputs from MDS analysis.

The pairwise Fst distance matrix is shown in **table 4.10**. The greatest genetic distance was found between the Togolese and Connaught populations (Fst = 0.968). The population found to be most closely related to the African population was Turkey (Fst = 0.642). Within the Gaelic Irish populations Connaught and Leinster were the most distantly related (Fst = 0.215), Connaught being most closely paired with Munster (Fst = -0.021) and Ulster and Leinster also showing small Fst values. Genetic distances increased roughly with geographic distance for all population comparisons.

	Dimension	
Population	**1**	**2**
Ulster	0.8887	-0.5980
Munster	1.6937	0.3454
Leinster	0.5225	-0.4973
Connaught	1.6051	0.3040
Africa	-3.2874	-0.2117
Denmark	-0.4260	0.2789
England	-0.1985	0.2542
Norman	0.2985	-0.4908
Scotland	-0.2964	-0.0105
Turkey	-0.8102	0.6257

Table 4.9: Eigenvalues for MDS plot from individual difference Euclidean distance based on Fst measures

When eigenvalues, obtained from MDS analysis, were examined, dimension 1 showed the Turkish population midway between the Togolese population and the most

distant Irish population. When dimension 1 and 2 were plotted on a plane the African population was clearly separated from the other geographic population groups. Interestingly, the Norman, Ulster and Leinster populations clustered and the Munster and Connaught populations clustered. The English, Danish and Scottish populations also grouped together with little evident association with other populations.

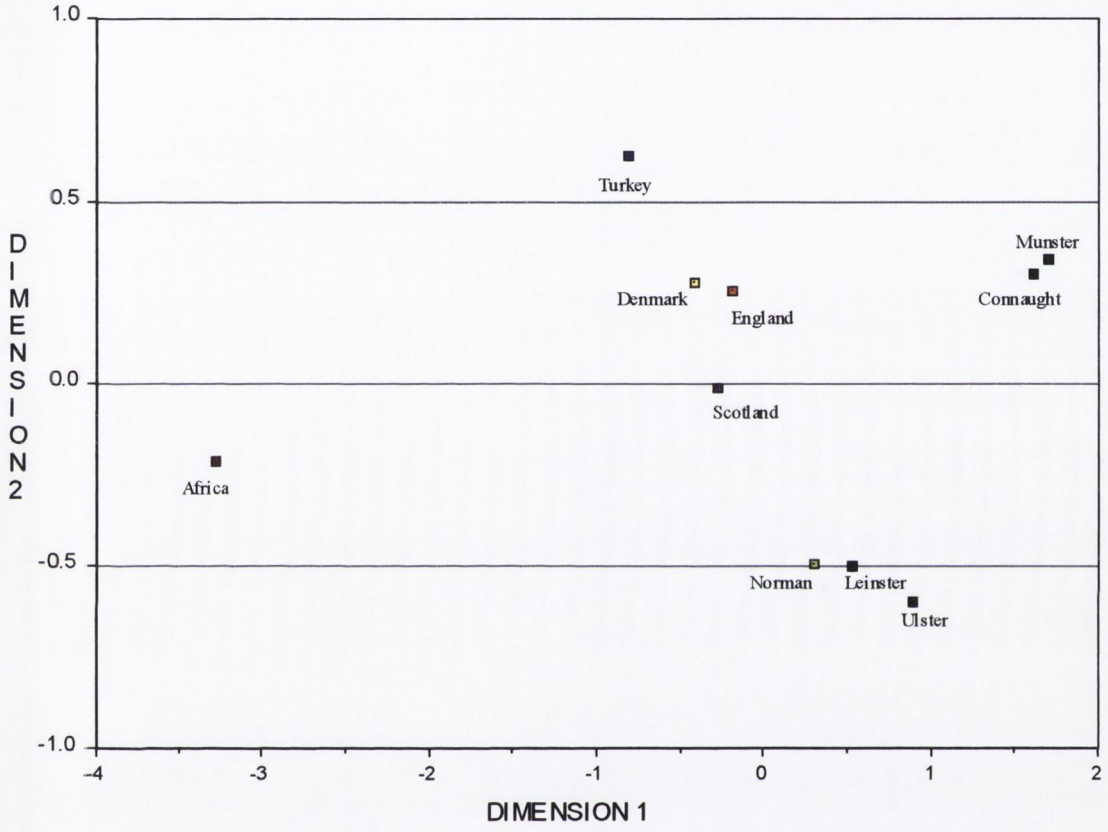


Figure 4.8: Multidimensional scaling plot of Fst distances between populations. The African population is the furthest removed from all other populations. Interestingly the Gaelic Irish populations cluster to the exclusion of the exogenous Irish and other European populations with the exception of the Norman population.

MDS, also carried out with the X STR data, achieved a separation of the African population from non-African populations. X STR loci, however, could not resolve the relationships between closely related populations. It is evident, however, that Y STR loci are powerful tools for resolving recent human population history.

	Ulster	Munster	Leinster	Connaught	England	Scotland	Norman	Togo	Asia	Denmark	Ireland
Munster	0.103										
Leinster	-0.035	0.143									
Connaught	0.175	-0.021	0.215								
England	0.026	0.253	0.004	0.330							
Scotland	0.039	0.346	0.010	0.458	-0.040						
Norman	-0.034	0.026	-0.008	0.087	0.086	0.109					
Togo	0.908	0.958	0.892	0.968	0.854	0.868	0.932				
Asia	0.508	0.865	0.429	0.912	0.273	0.205	0.633	0.907			
Denmark	0.045	0.373	0.019	0.495	-0.026	-0.069	0.110	0.873	0.195		
Ireland	...	...	...	...	0.231	0.298	-0.029	0.935	0.740	0.318	
Turkey	0.441	0.610	0.417	0.672	0.326	0.267	0.478	0.642	0.082	0.267	0.692

Table 4.10: Pairwise Fst genetic distances for STR data between population groups. Samples collected in Ireland were separated into population cohorts by surname origins.

Following population growth, less common nodes should arise from the common ancestral haplotype creating a star-like pattern of nodes and branches. In all of the phylogenies the area of the node is proportional to the number of times that haplotype was sampled in the population and the branch lengths are proportional to the number of mutational events separating the nodes.

Irish phylogenies

The most common haplotype in Connaught, ht 128, is found with a number of closely related branches radiating from it, and it lies centrally on the network. The hg 1 haplotypes cluster at one side of the network with the sole hg 2 chromosome found at least 8 mutational steps from the most closely related hg 1 chromosome.

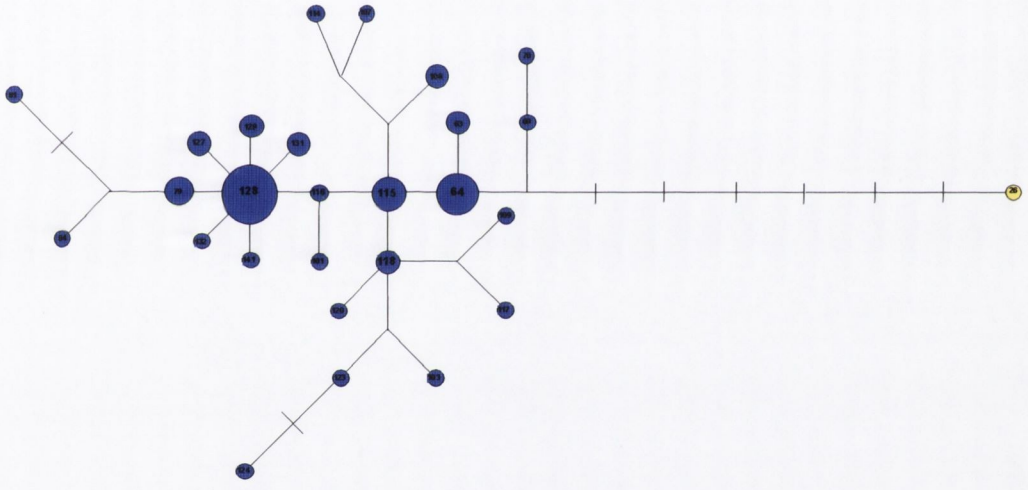


Figure 4.9: Maximum parsimony network relating haplotypes sampled in the Connaught population (n = 57). The most frequent haplotype sampled was ht 128 which lies centrally in the network and shows closely related haplotypes radiating from it. The sole hg 2 haplotype sampled in the Connaught population is distal to all hg 1 chromosomes.

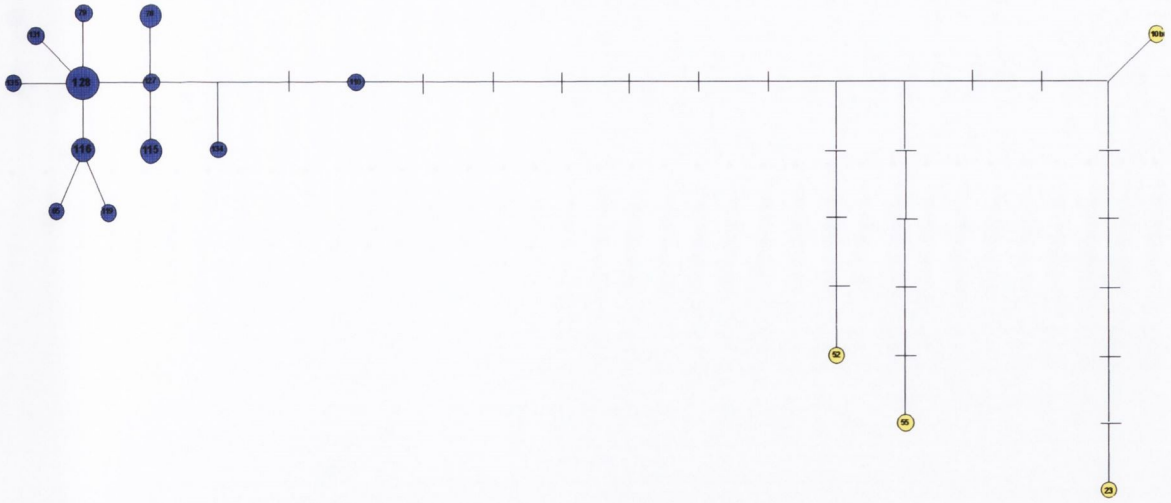


Figure 4.10: Maximum parsimony network relating all haplotypes sampled in the Ulster population (n = 22). The most common haplotype, ht 128, lies central in the core of the network surrounded by closely related hg 1 chromosomes. The hg 2 chromosomes sampled in the Ulster population lie at the far extreme of the network from the hg 1 chromosomes. The hg 2 chromosomes are more diverse than the hg 1 chromosomes and do not cluster to any significant extent.

In Ulster, the same most common haplotype, ht 128, is also most likely to be the ancestral haplotype. It lies centrally on the network and has a number of closely related branches diverging from it. Four hg 2 chromosomes are present in the Ulster population. The hg 2 chromosomes are found on long branches and show little relationship with each other.

The most common and putatively ancestral haplotype, in the Munster population is ht 64. This and ht 115, also common in the Munster population, are found centrally on the network. The two hg 2 chromosomes are distal to all other chromosomes.

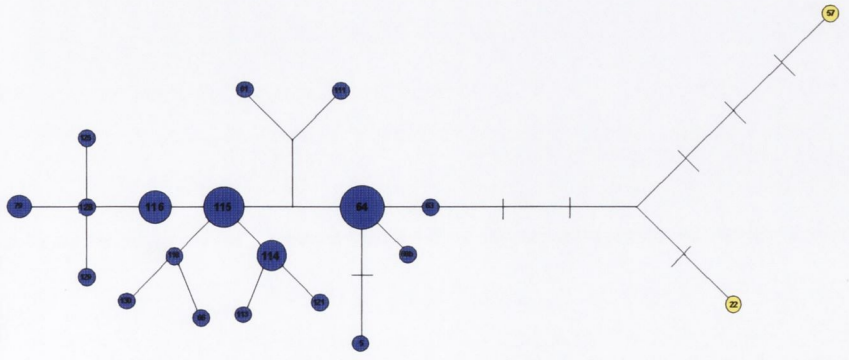


Figure 4.11: Maximum parsimony network relating haplotypes sampled in the Munster population (n = 37). The two common haplotypes, ht 115 and ht 64, lie centrally in the network.

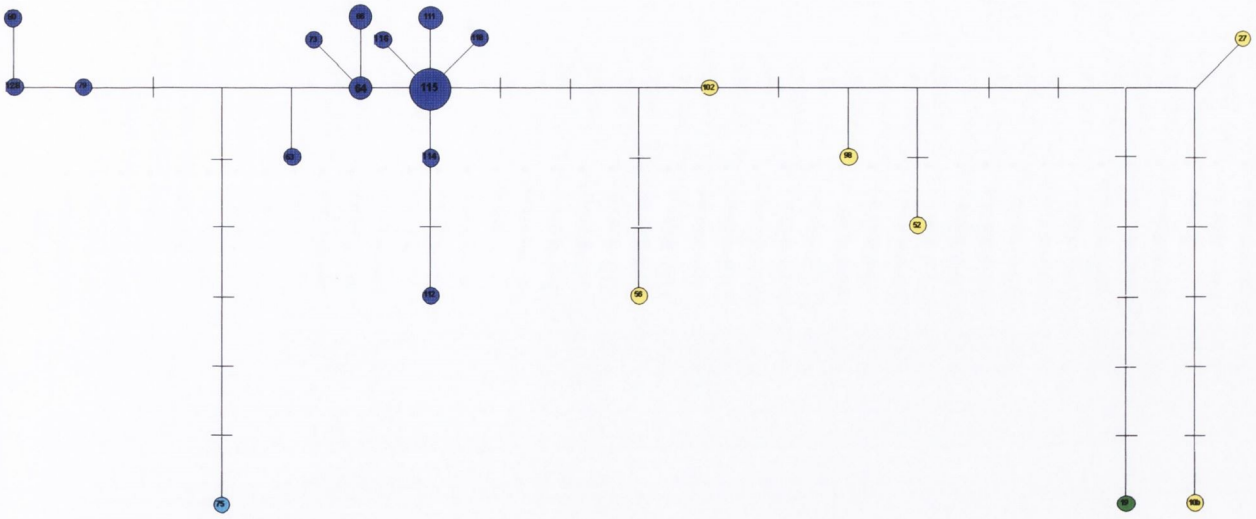


Figure 4.12: Maximum parsimony network relating haplotypes sampled in the Leinster population (n = 30). Four different haplogroups were sampled and they are separated in the network. The hg 1 chromosomes cluster together, the most common hg 1 haplotype, ht 115, lying in a central position in the network surrounded by closely related hg 1 chromosomes. The hg 2 chromosomes are more diverse and there is no evident clustering of haplotypes.

The Leinster phylogeny is more diverse than the other Gaelic networks. The most common haplotype is ht 115 and is located centrally on the network. A number of hg 1 chromosomes, however, do not cluster with ht 115 and are found distal on the network. The presence of seven non-hg 1 chromosomes in the Leinster population adds to its diversity.

When all of the Gaelic haplotypes were related in a phylogeny, the common haplotypes in each subpopulation were all located centrally on the network, specifically, ht 128, common in Ulster and Connaught, ht 64 common in Munster, and ht 115, common in Leinster. Additionally, the hg 2 chromosomes, obvious outliers in the networks of the subpopulations, show more intrinsic structure.

All of the Gaelic networks share a number of obvious features. First, common ancestral haplotypes are located centrally on the phylogeny. Second, less common closely related haplotypes are found radiating in star-like patterns from the ancestral node. Third, non-hg 1 haplotypes are not associated on the networks with hg 1 chromosomes and are found on long, distal branches. Fourth, non- hg 1 chromosomes do not cluster with the closeness of the hg 1 chromosomes, which are topologically coherent.

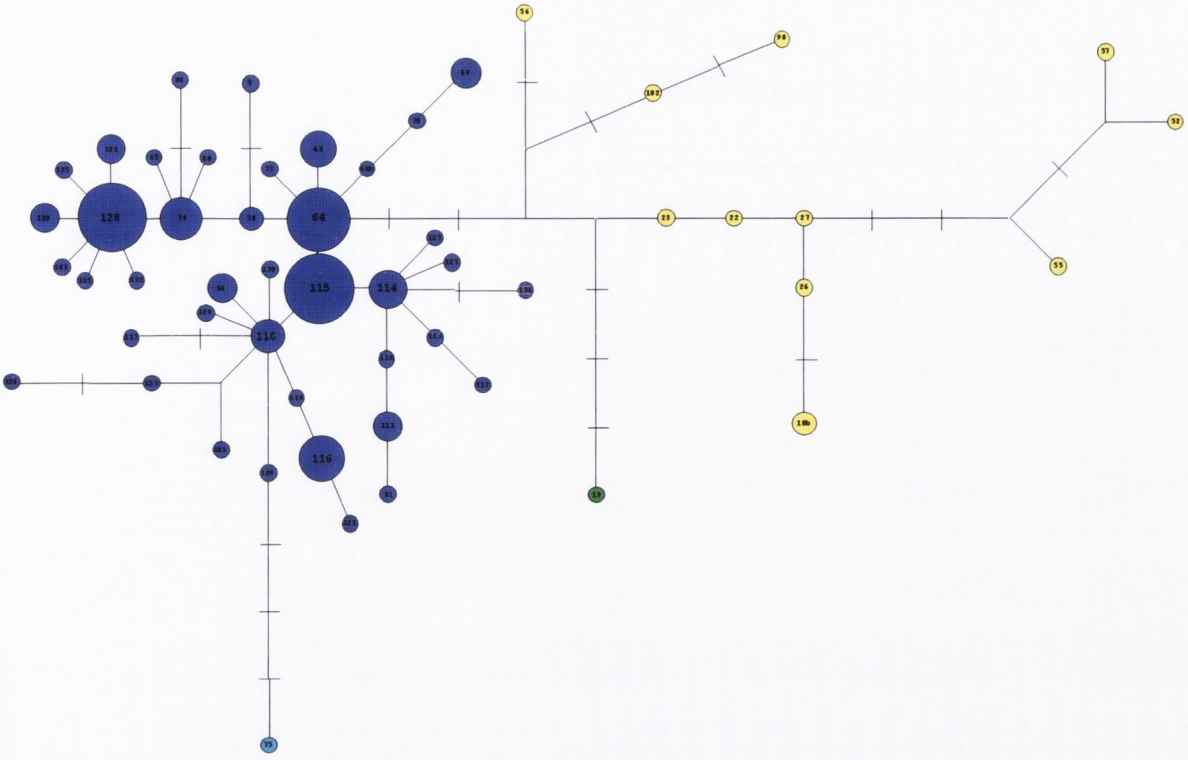


Figure 4.13: Maximum parsimony network relating all halotypes sampled in the Gaelic Irish population (n = 146). Notably, the common haplotypes, ht 64, ht 115 and ht 128, are centrally located in the network. The hg 1 chromosomes demonstrate phylogenetic coherence and star-like patterns suggestive of recent population growth. The hg 2 chromosomes are more diverse and may be intrusive.

Non-Irish Northwest European phylogenies

The maximum parsimony networks for the Scottish, Danish and Norman/Norse networks do not impart much information. The small sample sizes do not facilitate the

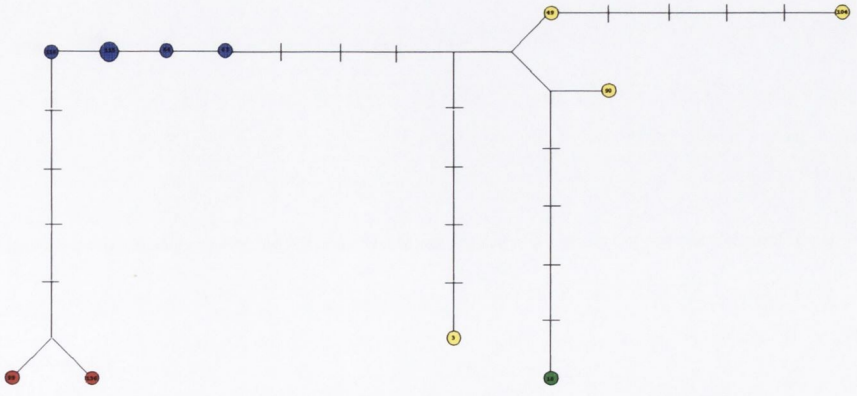


Figure 4.14: Maximum parsimony netork relating haplotypes sampled in the Danish population (n = 12). Four hapolgroups were sampled in the Danish population and haplotypes within those haplogroups tend to cluster.

identification of an ancestral haplotype from phylogenetic information. The diversity within these populations, as demonstrated by the presence of more numerous non-hg 1 chromosomes, is evident in the topology of the non-Gaelic Northwest European phylogeny. The English phylogeny, with a comparable sample size to the Gaelic subpopulations, demonstrates greater diversity than the Gaelic populations.

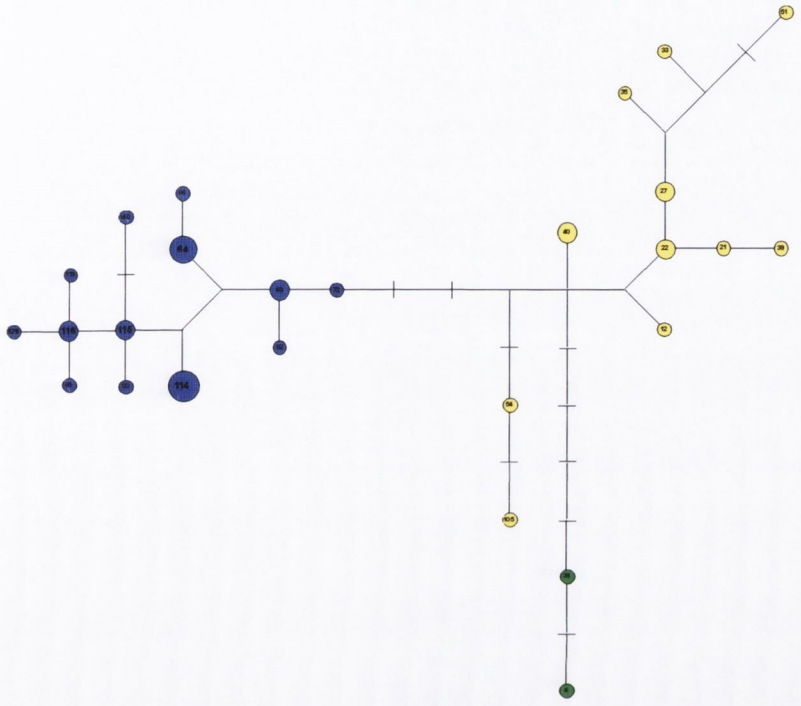


Figure 4.15: Maximum parsimony network relating haplotypes sampled in the population with names of English origin (n = 40).

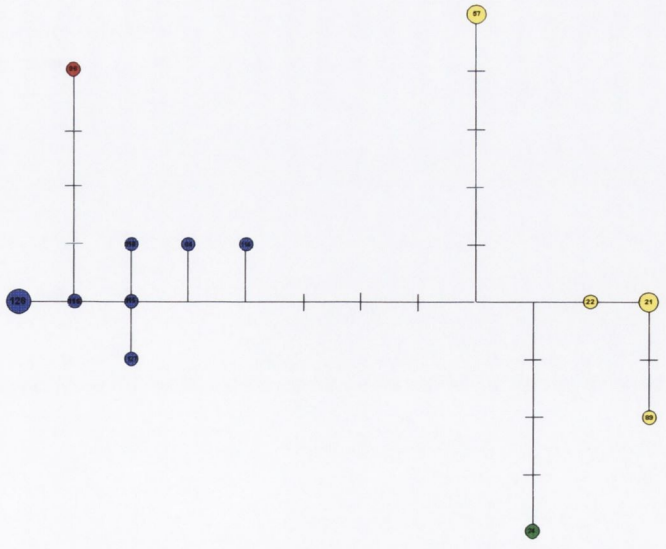


Figure 4.16: Maximum parsimony network relating haplotypes sampled in the population with names of Scottish origin (n = 17).

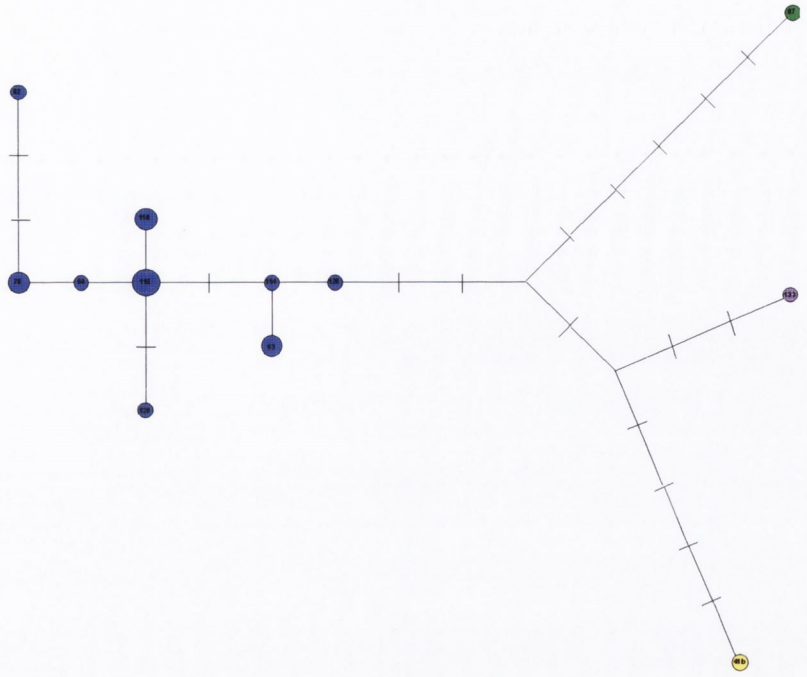


Figure 4.17: Maximum parsimony network relating haplotypes sampled in the population with names of Norman and Norse origin (n = 18).

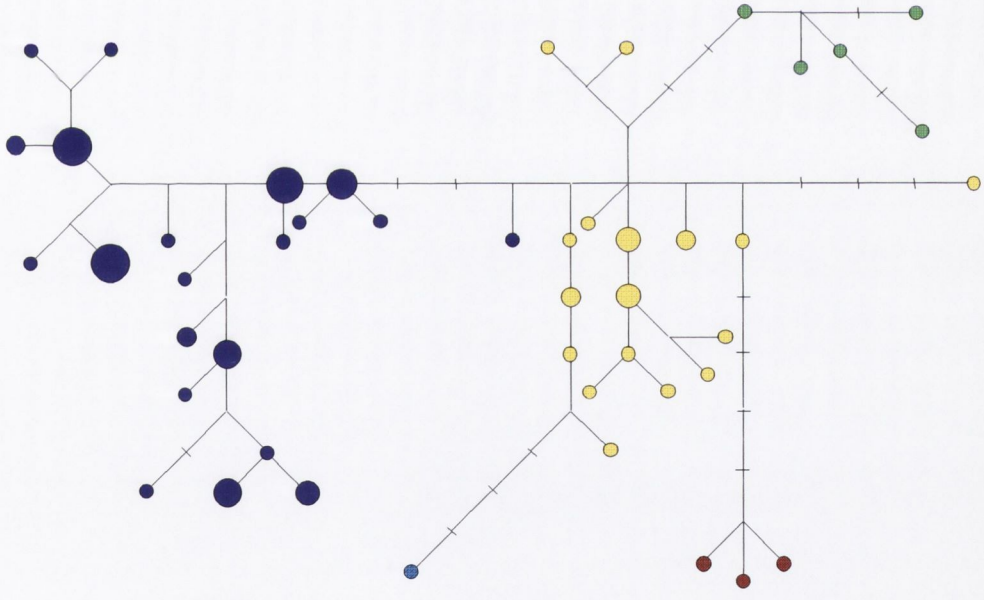


Figure 4.18: Maximum parsimony network relating all non-Irish haplotypes sampled in Northwest European populations (n = 88). The most obvious feature of the network is the separate clustering of diverse haplogroups to the exclusion of others. No putatively ancestral haplotype can be identified from this phylogeny.

When a phylogeny was generated for all non-Gaelic Northwest European chromosomes, a number of features were evident. All haplotypes of similar haplogroup origin grouped on the network. However, common ancestral haplotypes in each

chromosomal lineage could not be easily identified. No star-like clusters of closely related haplotypes were evident within any of the diverse chromosomal lineages.

Turkish phylogeny

The Turkish phylogeny is topologically the most diverse of all of the networks. In total a minimum number of 76 independent evolutionary events are required to relate all of the 50 haplotypes within this population. Taking a similar sized population, for example Connaught, the minimum number of mutational steps required in the maximum parsimony phylogeny to relate 57 individual haplotypes is only 40. In the Turkish sample this is a function of the presence of numerous distantly related Y chromosomal lineages. Unlike the other phylogenies, however, the hg 1 chromosomes do not cluster and are more closely associated with hg 2, hg 26 and hg 12 chromosomes than they are with each other. Unlike the Gaelic phylogenies, the most common haplotype, ht 12, assumed to be the ancestral chromosomal type, is not surrounded by closely related nodes, and in fact is found two mutational steps away on the tree from the most closely related haplotype to it.

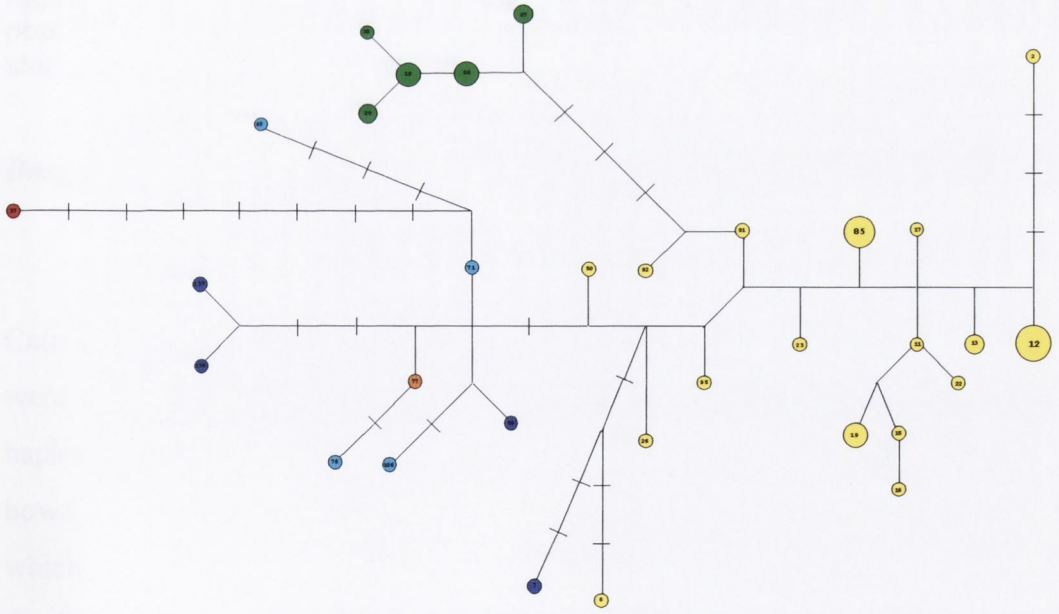


Figure 4.19: Maximum parsimony network relating haplotypes sampled in the Turkish population (n = 50). The Turkish population is more phylogenetically diverse than any of the other populations.

Togolese phylogeny

The phylogeny created from haplotypes present in the African population exhibits the star-like pattern present in the Gaelic Irish networks, with the most common haplotype, ht 28, lying central on the network, with numerous closely related haplotypes radiating from it. The hg 8 haplotypes cluster to the exclusion of the hg 21 chromosomes.

Figure 4.20: Maximum parsimony network relating haplotypes sampled in the Togolese population (n = 37). The hg 8 chromosomes demonstrate coherence and an obvious star-like pattern. The hg 21 chromosomes are more diverse and show no clustering.

Basque and Catalan phylogenies

Haplotype data was drawn from Perez-Lezan *et al.* (1997) for the Basque and Catalan populations. In the five-locus haplotypes (DYS19/388/390/391/392), four loci were common in this study. Maximum parsimony networks were created from five-locus haplotype data in the same manner as for the populations in this study. The networks, however, are not directly comparable as a lesser number of loci were used, only four of which were common in both studies. Additionally, further SNP information is unknown for the Basque and Catalan populations, although it has been estimated that 89% of Basque chromosomes are hg 1 (see **chapter 3**). Additionally, the most common Catalan STR haplotype is most similar to a hg 2 haplotype identified in this study.

Despite the non-uniformity of the data, a crude comparison of phylogenetic patterns can be made assuming that allele calling in the two studies is correct. Allele

calling in this study was determined from comparison with M13 sequence (see **chapter 2**). Furthermore, the most common alleles identified in populations in this study have been found to be the most common in other studies of the same populations *(deKnijff et al 1997)*. Equivalent four-locus haplotypes between the two studies are given in table X.

4 locus haplotype *Perez-Lezaun*	**4 locus haplotype** *This study*
1, 25	108, 111, 115, 118, 120
2	61, 64, 66, 73
3, 18	13, 24
4	63, 72
5, 6	11, 22, 23, 142
7	27
8, 9	26, 29, 32, 41
10	-
11	45
12	-
13	65
14	-
15	-
16	-
17	68, 70
19	103
20	92
21	107, 110, 114
22	126
23	-
24	139, 140
26	-
27	113, 123
28	106, 121, 124
29	-

Table 4.12: Comparable four-locus Y STR haplotypes between this study and Pérez-Lezaun *et al.* (1997)

The most common haplotype in the Basque population is *ht 1 (Pérez-Lezaun et al. 1997)* which is comprised of alleles 190/129/215/287/254 at loci DYS19/388/390/391/392 respectively. If DYS388 was removed from the Basque and Catalan haplotypes then *ht 1* and *ht 25* are identical. If DYS389 and DYS393 were removed from the data in this study then four locus haplotypes can be compared. *Haplotypes 1* and *25 (Pérez-Lezaun et al 1997)* are identical then to ht 108, ht 111, ht 115, ht118 and ht 120 *(this study)*. All of these haplotypes are present on hg 1 chromosomes. It is worth noting that ht 115 is the

most common haplotype in the Leinster, Norman and Danish populations in this study, and is also sampled in the Ulster, Munster, Connaught, English and Scottish populations. Haplotype 108 was sampled in Connaught only, ht 111 was sampled in Munster and Leinster and ht 120 was sampled in Connaught alone. Haplotype 118 was more common and was found in the Munster, Leinster, Connaught, Scottish and Norse populations. Apart from the common ht 115, none of the other Basque haplotypes were sampled outside Ireland.

The Basque phylogeny shows *ht 1* central on the network with a number of related haplotypes joined to it. This is consistent with patterns exhibited in the Gaelic networks.

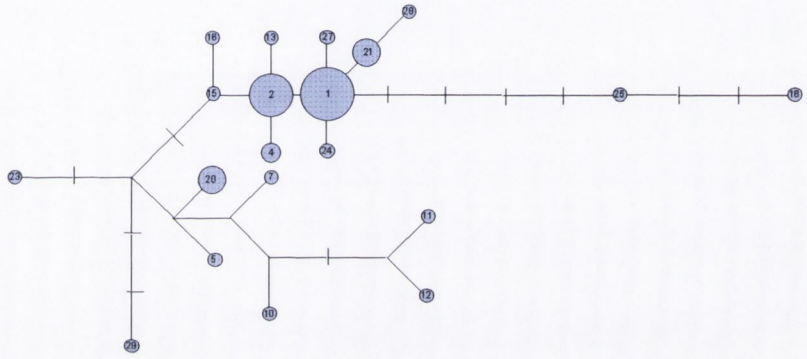


Figure 4.21: Maximum parsimony network relating haplotypes in the Basque population (n = 49). Haplogroup information is unavailable, but the majority (~90%) of Basque Y chromosomes are hg 1. Haplotypes are numbered according to Pérez-Lezaun *et al. (1997),* equivalent haplotypes in this study are given in **table 4.12**.

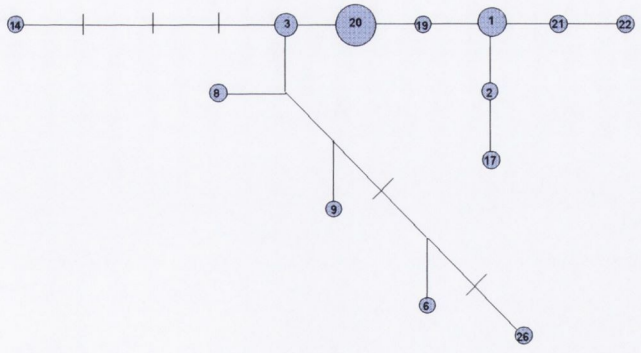


Figure 4.22: Maximum parsimony network relating haplotypes in the Catalan population (n = 22). Haplogroup information is unavailable, but the most common Catalan haplotype, *ht 20*, is most similar to a hg 2 chromosome (ht 92) in this study. Haplotypes are numbered according to Pérez-Lezaun *et al. (1997),* equivalent haplotypes in this study are given in **table 4.12**.

In contrast to the Basque populations, the most common haplotype in the Catalan population is *ht 20.* A comparable four-locus haplotype in this study is ht 92.

Interestingly ht 92 is found on a hg 2 chromosome only and was sampled only once in the Turkish population and no other population. There is little structure in the phylogeny generated from Catalan microsatellite diversity.

Haplogroup phylogenies

Maximum parsimony networks were also constructed to assess the relationships of haplotypes within discrete haplogroup lineages. The networks are shown in **figures 4.23 and 4.24** and **appendix B**. Most notably, the hg 1 chromosomes are coherent and cluster tightly. The common hg 1 haplotypes, ht 64, ht 114, ht 115, ht 116 and ht 128, are found at central core positions on the network. Star-like patterns are most evident from the ht 128, ht 114 and ht 115 nodes. There are a number of distal hg 1 haplotypes.

Figure 4.23: Maximum parsimony network relating haplotypes on hg 1 chromosomes (n = 189).

The haplotypes shared between populations within hg 1 are illustrated in **appendix C.** Notably, the more distal hg 1 haplotypes were sampled in the Turkish population. None of the Turkish hg 1 chromosomes are sampled frequently and neither are they more related to each other than they are to other non-Turkish hg 1 chromosomes. The

peripheral hg 1 haplotypes are distantly related to the core ancestral haplotypes in Northwest Europe. The hg 1 chromosomes sampled in the English, Scottish and Danish populations include those that are most common in hg 1.

	n	Minimum number of mutational steps
hg 1	189	72
hg 2	68	67
hg 3	2	10
hg 8	31	27
hg 21	23	20
hg 26	9	24

Table 4.14: Minimum number of mutational steps required in the maximum consensus maximum parsimony trees to relate all haplotypes.

Haplogroup 2 chromosomes are non-coherent and many outlier haplotypes exist. The most frequent haplotype, ht 12, shows no phylogenetic evidence of being an ancestral chromosomal type. Only two branches arise from ht 12. Haplotype 22 on the other hand, although less common than ht 12 is located at a central position and has a number of common haplotypes radiating from it. Many nodes are separated by more than one mutational step, which is in contrast to the hg 1 phylogeny.

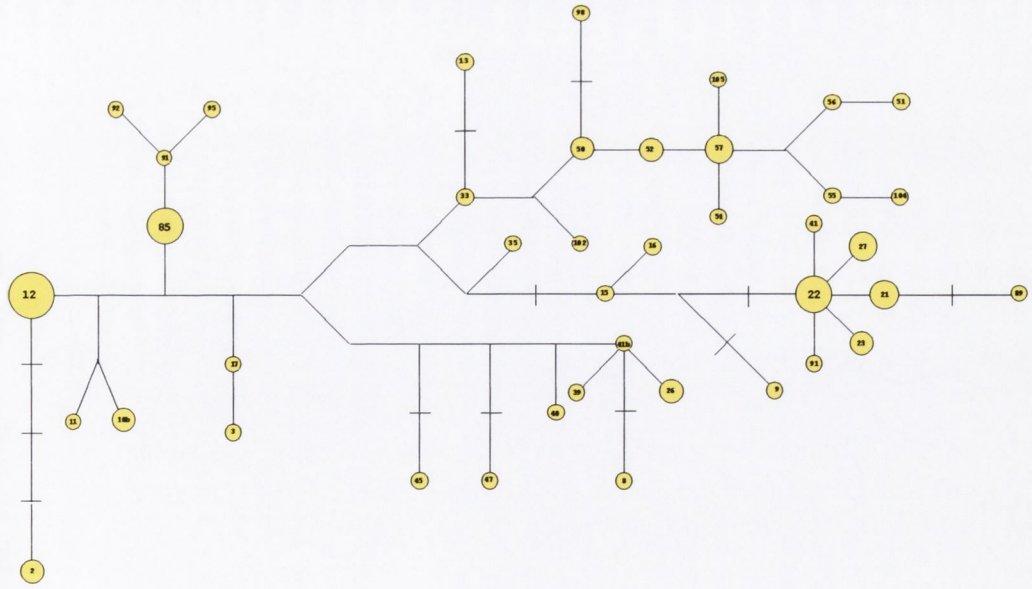


Figure 4.24: Maximum parsimony network relating haplotypes on hg 2 chromosomes (n = 68).

The maximum parsimony phylogenies of hg 3, hg 8, hg 21 and hg 26 (includes hg 12 chromosome) demonstrate little core structure and are shown in **appendix B**.

4.3.11. Mismatch distributions

Pairwise mismatch distributions, using only STR data, were created in the same manner as outlined in **chapter 2.** The distance measure D_{swasm} was used to produce pairwise difference matrices. However, rather than the examination of populations only, the dynamics of discrete chromosomal lineages were also investigated.

The mismatch distributions of Y STRs in defined haplogroups are illustrated in **figure 4.25.** When all of the haplotypes from all haplogroups were pooled, a smooth unimodal distribution resulted. The distributions of the STR diversity in hg 1, hg 2 and hg 8 result in smooth unimodal distributions. The modal number of differences between hg 1 Y chromosomes was 2. The modal number of differences between hg 2 and hg 8 chromosomes was 4 and 2 respectively. The distributions for hg 21 and hg 26 are in contrast with those for hg 1, hg 2 and hg 3. The pairwise difference distributions for hg 21 and hg 26 are uneven by comparison. Interestingly, the hg 21 distribution is bimodal. The means of the distributions are given in **table 4.13.**

Mismatch distribution	Mean no. differences
All haplogroups	4.826
hg 1	2.548
hg 2	4.206
hg 3	3.800
hg 8	2.452
hg 21	2.474
hg 26	4.821

Table 4.14: Mean number of differences between pairs of haplotypes within each chromosome lineage. The mean values for hg 2 and hg 26 are most similar to the mean number of differences found between all pairs of haplotypes in the entire sample set.

The diversity of hg 1 chromosomes sampled in Ireland was summarised in the mismatch curve shown in **figure 4.25.** The bulk of lineage coalescences involve 2-3 mutations and the mean of the distribution is 2.55.

Reasoning, from previously published data, that the majority of Basque chromosomes were of hg 1 type (see **chapter 3**), with similar SNP frequencies to Irish populations, comparable pairwise difference distributions were created from 4 locus (DYS19/390/391/392) haplotypes for the Irish and Basque populations. For comparison, pairwise mismatch curves were also created for 4 locus haplotype information from the

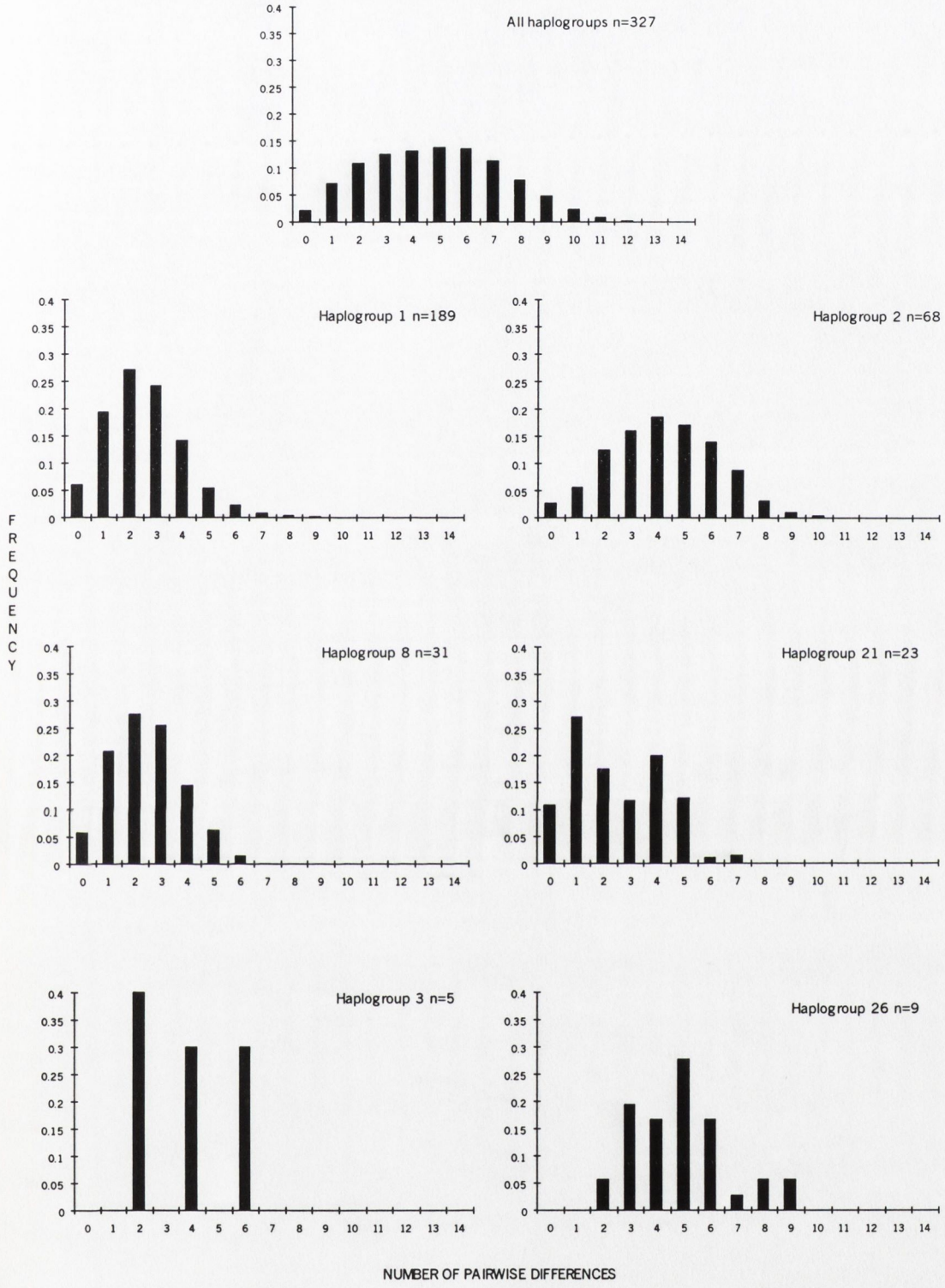


Figure 4.25: Pairwise difference distributions in defined haplogroups. Haplogroup 2 chromosomes demonstrate the greatest antiquity, the mean, mode and median number of differences being the most distal from the y-axis. Haplogroup 2 chromosomes also demonstrate a distribution most similar to the normal distribution created by pooling all haplotypes in all populations.

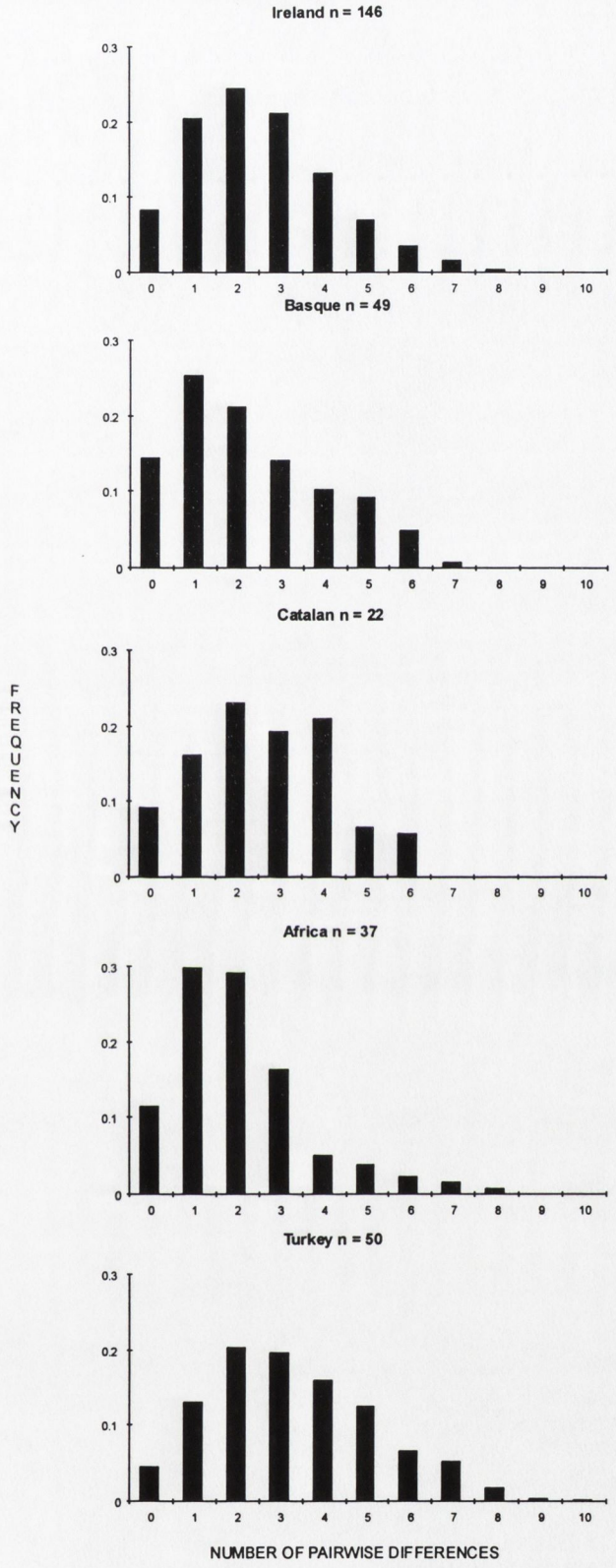


Figure 4.26: Pairwise difference distributions for four-locus comparable haplotypes (DYS19 / 390 / 391 / 392) in

Catalan, Togolese and Turkish populations. These mismatch curves are illustrated in **figure 4.26.** The distributions in the most part are unimodal in nature and demonstrate smooth, even curves. The Catalan distribution, however, is somewhat uneven in nature. The majority of the Irish chromosomes demonstrated 2 differences between them. In the Basque population most of the chromosomal coalescences involved only 1 mutation. In the African population the modal number of pairwise differences between chromosomes was 1-2, and in the Turkish population the modal number of differences was 2. The means of the distributions varied. Generally, the greater the mean number of differences, the older and more diverse the population, coinciding with a shift in the curve away from the vertical axis. Interestingly the African population demonstrates the lowest mean number of differences with a value of 2.06. The Turkish population had the highest mean number of differences with a value of 3.34. The Irish, Basque and Catalan populations had similar mean values (2.55, 2.32 and 2.68 respectively).

4.3.12. Coalescence dates : Time to Most Recent Common Ancestor (TMRCA)

Three factors have the ability to influence the calculation of the coalescence time to the most recent common ancestor (TMRCA); the genetic distance measure, the mutation rate and the generation time. Three genetic distance measures, and therefore three different dating methods, the average squared distance (ASD) method *(Goldstein et al 1995; Slatkin 1995*), Goldstein's variance method *(Goldstein et al 1996)*, and the Bertranpetit and Calafell (B&C) method *(Bertranpetit and Calafell 1996)* were used to calculate the TMRCA in a number of haplogroups and populations. All of these methods rely on the diversity of microsatellite haplotypes that has arisen from one ancestral Y chromosome and calculate the time taken to reach that level of diversity before reaching equilibrium.

Generally, a mutation rate of 0.21% *(Heyer et al 1997)* was used, however, for comparison, mutation rates described by Cooper *et al. (1997)* were also employed.

A generation time of 25 years was used as standard *(Thomas et al 1998)* although this may underestimate the coalescence times if, realistically, a larger generation time should be used. A generation time of 27 years has been calculated from contemporary populations believed to have breeding practices similar to Paleolithic peoples *(Weiss 1973).*

The ASD and the B&C methods both rely on the initial determination of an ancestral or root haplotype. As described, the ancestral haplotype is generally the most common haplotype. However, especially in populations with small sample size or with older coalescence, the ancestral common haplotype may no longer be present. Therefore, the ancestral haplotype was determined for a population of chromosomes within a discrete haplogroup by taking the modal allele at each locus. In most cases, the modal haplotype was also the ancestral haplotype determined from haplotype frequency and phylogenetic information. The time for the contemporary diversity in a population of chromosomes to have arisen from that root haplotype was calculated in each given haplogroup.

Average squared distance method

The expected value of ASD is a linear function over time. It is independent of sample size and therefore direct estimates of divergence times can be made if an accurate mutation rate is known *(Goldstein et al 1995)*. The average squared distance was calculated between a population of chromosomes and their root haplotype using the Microsat program *(Minch)*. The formulation of ASD calculation has been described in **chapter 2**, although without the necessity to correct for interpopulational variability, the ASD has an expectation of μt (not $2\mu t$), where μ is the mutation rate and t the coalescence time in generations. The TMRCA of a haplogroup is therefore related to the ASD by the relationship

$$ASD = \mu t$$

and therefore $$\text{Age (in generations)} = ASD / \mu$$

Goldstein's variance method

The variance method of dating *(Goldstein et al 1996)* is based on the time to accumulate the observed variance at each microsatellite locus assuming that all of the observed diversity has arisen from one chromosome at time zero and the variance at that time was zero. This method of dating is useful for the application to populations of

chromosomes defined by unique event polymorphisms that have arisen from single chromosomes.

Importantly, variance at a microsatellite locus fluctuates simply due to random processes of mutation and drift. When mutation-drift equilibrium is reached then the expected variance in microsatellite repeat number, assuming a stepwise model of mutation, is

$$V\hat{} = (Ne-1)\,\mu$$

where Ne is the effective population size and $\mu/2$ is the probability of mutation changing a microsatellite by +1 and -1 repeats *(Moran 1975),* i.e. μ is the mutation rate.

If mutation-drift equilibrium has not been reached then the observed variance in the present population can be used to estimate the time elapsed since time zero.

The dynamic for the variance in repeat number is

$$Vr(t) = Vo(1-1/Ne)t+(Ne-1)[1-(1-1/Ne)t]\mu$$

and therefore

$$t = \ln(1-Vr(t)/V\hat{})/\ln(1-1/Ne)$$

when Vo = 0 and Vo is the variance in the population at time zero under the assumption of an initial expansion from one Y chromosome when there is no variation. The above equation can be used to calculate the time passed since Vo = 0 only if the observed variance is less than the expected variance at mutation-drift equilibrium i.e. that mutation-drift equilibrium has not yet been reached.

Therefore it is necessary to firstly discount the possibility that the diversity is at mutation-drift equilibrium (V^) by determining whether the observed diversity (Vr(t)) is significantly different from that expected at mutation-drift equilibrium by using the equations for 95% confidence limits V1 and V2.

$$V1=\exp\{\ln Vr(t)-Vg(t)/(2lVr(t)2)-(2/Vr(t))*(Vg(t)/l)^{0.5}\}$$

$$V2=\exp\{\ln Vr(t)-Vg(t)/(2lVr(t)2)+(2/Vr(t))*(Vg(t)/l)^{0.5}\}$$

where Vg(t) is the variance about Vr, as a function of time since Vo = 0, scaled by the population size (Ne = 4900 *(Hammer 1995)*). Vg(t) is calculated by the following equation:

Vg(t*)=1/5[(V^+12V^2)(1-e-t*)-1/6(V^+2V^2)(1-e-6t*)+2/5V^2e-t*(1-e-5t*)-12V^2t*e-t*]-V^2(1-e-t*)2

where t* = t/N

To calculate the 95% confidence limits around V^ a TMRCA of 200,000 years for all Y chromosomes was used *(Hammer 1995, Underhill 1997).*

The observed diversity in the population (Vr(t)) is calculated as

$$V_r(t) = (\Sigma\ Vi)/l$$

where Vi is the variance at each locus and l is the number of loci in the haplotype.

If Vr(t) is outside the 95% confidence limits for V^ (i.e. V1 and V2) then mutation-drift equilibrium has not yet been reached and this dating method is appropriate for that population of chromosomes.

The derived equation

$$t = \ln(1-Vr(t)/V^)/\ln(1-1/Ne)$$

can therefore be used to calculate TMRCA for a population of chromosomes within a haplogroup and the upper and lower confidence limits for this date are calculated by substituting V1 and V2 for V^ in the equation above.

Bertranpetit and Calafell method

The B&C method for dating *(Bertranpetit and Calafell 1996)* is based on the average number of mutations accumulated in a population of chromosomes from the ancestral or root haplotype. The ancestral haplotype was determined as the modal haplotype in a population of present chromosomes as described above. Alternatively, it can be determined as the haplotype with the least number of mutational steps from all other chromosomes ascertained by creating a pairwise difference matrix of all possible combinations of haplotypes.

The TMRCA is calculated from the following relationship

$$\lambda / \mu = t \text{ (in generations)}$$

where λ is the sum of the number of mutations accumulated between the ancestral haplotype and the haplotype on all other chromosomes in the population divided by the number of loci in the haplotype, which equals the number of loci per chromosome multiplied by the number of chromosomes.

4.3.13. TMRCA estimates for haplogroup lineages

The TMRCA was calculated for each haplogroup using all of the above dating methods. Dates were calculated using an Excel™ spreadsheet designed and kindly supplied by Matthew Hurles *(Department of Genetics, University of Leicester, Leicester)*.

Dating method	hg 1	hg 2	hg 3	hg 8	hg 21	hg 26
ASD	3798 *1600-13300*	8143 *3500-28500*	17464 *7500-61100*	4762 *2000-16700*	5869 *2500-20500*	8381 *3600-29300*
B&C	3338 *1400-11700*	5806 *2500-20300*	4762 *2000-16700*	3712 *1600-13000*	3882 *1600-13600*	6834 *2900-23900*
Goldstein	2001 *750-10200*	4222 *1500-23000*	7150 *2500-42900*	3585 *1300-19200*	5281 *1900-29800*	8265 *2900-51600*

Table 4.15: TMRCA calculated for all haplogroups using 3 different methods of dating. All of the dates are based on a microsatellite mutation rate of 0.21% *(Heyer et al. 1997)* and a generation time of 25 years *(Thomas et al. 1998)*. The upper and lower 95% confidence intervals are shown in italics below the dates and are calculated from the 95% confidence limits of the mutation rate.

Haplogroup 1

The ancestral haplotype, determined from the modal allele at each locus, in the hg 1 population was identical to ht 115. Haplotype 115 was the most common hg 1 haplotype and was found on 13.6% of all hg 1 chromosomes. Using the ASD method a date for hg 1 coalescence of 3,798 *YBP* was estimated. Factoring in 95% confidence limits for the microsatellite mutation rate, 0.06-0.49% *(Heyer et al 1997),* this resulted in an estimated 1,628-13,292 *YBP* confidence interval. Goldstein's variance method and the B&C method of dating calculated younger estimates of coalescence dates for hg 1 chromosomes. Using the strategy above, the B&C method estimated a TMRCA for hg 1 chromosomes as 3,338 *YBP* (CI 1,431-11,684 *YBP*). Goldstein's variance method estimated a coalescence time of 2,001 *YBP* (CI 748-10,219 *YBP*).

Because a number of hg 1 chromosomes were peripheral on the networks (mainly those from Turkey) the age of hg 1 chromosomes may be distorted. This age of hg 1 chromosomes may not even be indicative of European hg 1 coalescence as all of the diversity in Europe has not been sampled. From phylogenies it is clear that most of the hg 1 diversity in Ireland may have arisen within Ireland itself. It made sense, therefore, to calculate the coalescence for hg 1 chromosomes in the Irish population considering that most hg 1 chromosomes were detected in Ireland and that the majority (89.7%) of Irish chromosomes were hg 1. Using the ASD method, a date of origin for Irish hg 1 chromosomes was estimated as 3,905 *YBP* (CI 1,673-13,667 *YBP*). Using a 27 year generation time *(Weiss 1973)* an alternative date of origin was estimated as 4,217 *YBP* (CI 1,807-14,760 *YBP*).

Haplogroup 1 chromosomes are older in Turkey than elsewhere in Europe. In the Northwest of Europe, Irish hg 1 chromosomes have the most ancient origin and Danish hg 1 chromosomes are the youngest.

In order to assess whether hg 1 chromosomes were older in the Irish population than in other populations a number of calculations were performed. Notably, hg 1 chromosomes in Northwestern Europe were found to have an older origin in the Connaught population than in all other populations, the TMRCA estimated as 4,643 *YBP*. However, only in Turkey are the dates significantly different.

	ASD	TMRCA	lower 95% CI	upper 95% CI
Ireland	0.328	3,900	*1,700*	*13,700*
All non-Irish	0.285	3,400	*1,500*	*11,900*
England	0.267	3,200	*1,400*	*11,100*
Scotland	0.204	2,400	*100*	*8,500*
England and Scotland	0.250	3,000	*1,300*	*10,400*
Norman and Norse	0.267	3,200	*1,400*	*11,100*
Denmark	0.133	1,600	*700*	*5,500*
Turkey	1.250	14,800	*6,400*	*52,000*

Table 4.16: TMRCA for hg 1 chromosomes sampled in a number of populations.

The Turkish hg 1 chromosomes were the oldest with a coalescence date of 14,881 *YBP*. The least diversity had accumulated in the Danish hg 1 chromosomes, the date of origin being the youngest at 1,583 *YBP*. The estimated coalescence dates of hg 1 chromosomes in different populations are given in **table 4.16.**

Haplogroup 2

The ancestral haplotype, determined from the modal allele at each locus, in hg 2 chromosomes was identical to ht 22. Although ht 22 was not the most common haplotype in the hg 2 population, it was found at a reasonable frequency (7.4%). Haplotype 22 was sampled in a number of diverse populations whereas the most common hg 2 haplotype, ht 12, was sampled only in the Turkish population. The ASD method estimated a coalescence date for all hg 2 chromosomes as 8,143 *YBP* (CI 3,490-28,500 *YBP*). As for hg 1 chromosomes the other dating methods calculated younger coalescence times. Goldstein's method estimated an age of hg 2 chromosomes as 4,222 *YBP* (CI 1,549-23,029 *YBP*) and the B&C method estimated an age of 5,806 *YBP* (CI 2,488-20,323 *YBP*). In all cases hg 2 chromosomes were estimated to be older than hg 1 chromosomes.

Haplogroup 3

Only a small number of hg 3 chromosomes were detected and therefore the best estimate of coalescence in this population should come from ASD dating as this method is independent of population size. The ancestral haplotype, determined from the modal allele at each locus, was not sampled but was determined to be 194/251/219/283/248/124 (locus order as previously shown). The TMRCA for hg 3 chromosomes was estimated as 17,464

YBP (CI 7,485-61,125 *YBP*). The other methods estimated much younger dates. Goldstein's variance method calculated a coalescence time for hg 3 chromosomes as 7,150 *YBP* (CI 2,566-42,917 *YBP*) and the B&C method estimated a date of 4,762 BP (CI 2,041-16,667 *YBP*).

Haplogroup 8

Haplogroup 8 chromosomes were detected in the African population only. The ancestral haplotype, determined from the modal allele at each locus, was identical to ht 28, one of the most common haplotypes in this population found at a frequency of 8.3%. All three methods of dating were used. The ASD method estimated the TMRCA for hg 8 chromosomes as 4,762 *YBP* (CI 2,041-16,667 *YBP*), the B&C method estimated this time as 3,712 *YBP* (CI 1,591-12,993 *YBP*) and Goldstein's variance method estimated the date as 3,585 *YBP* (CI 1,323-19,184 *YBP*).

Haplogroup 21

Haplogroup 21 chromosomes were the only Y chromosome types to be detected in both African and non-African populations. The ancestral modal haplotype was identical to ht 14 which, in fact, was only detected twice in the hg 21 population. Incidentally ht 14 was sampled in the African population only. Using all three dating methods a range of 3,882-5,869 *YBP* was estimated. The ASD method calculated the TMRCA as 5,869 *YBP* (CI 2,515-20,542 *YBP*), the B&C method estimated this time as 3,882 *YBP* (CI 1,664-13,587 *YBP*) and Goldstein's variance method estimated the date as 5,281 *YBP* (CI 1,922-29,781 *YBP*).

Haplogroup 26

The ancestral modal haplotype for hg 26 chromosomes was identical to ht 71. The ASD method calculated the TMRCA for all hg 26 chromosomes as 8,381 *YBP* (CI 3,592-29,333 *YBP*). Coalescence dates calculated using the B&C method and Goldstein's variance method were younger at 6,834 *YBP* (CI 2,929-23,920 *YBP*) and 8,265 *YBP* (CI 2,943-51,634 *YBP*) respectively.

4.3.14. Coalescence times between haplogroups

In order to determine the divergence times between haplogroups, estimates of coalescence dates were calculated using the ASD dating method. However, because two chromosomal lineages were involved the relationship of ASD to the time of origin was

$$ASD = 2\mu t$$

In an examination of hg 1 coalescence, all of the dates, with the exception of the hg 1 - hg 12 comparison, predated the time of origin of hg 1 chromosomes. The oldest split was detected between hg 1 and the African-only haplogroup, hg 8, with a coalescence time for these two haplogroup lineages of 15,720 *YBP* (CI 6,737-55,021 *YBP*). In descending order the coalescence times between hg 1 and other haplogroups were younger between hg 1 - hg 3, hg 1 - hg 21, hg 1 - hg 2, hg 1 - hg 26, hg 1 - hg 16, the youngest split being that between hg 1 and hg 12 (3,845 *YBP*; CI 1,648-13,458 *YBP*). All of the coalescence dates determined between haplogroups are given in **table 4.17.**

In an examination of hg 2 coalescence, the oldest split was determined as that between hg 2 and hg 3 chromosomes, with a coalescence time of 14,994 *YBP* (CI 6,426-52,479 *YBP*). Haplogroup 2 chromosomes were most closely related to hg 21 chromosomes, these haplogroups having a coalescence time of 6,851 *YBP* (CI 2,936-23,979 *YBP*).

Haplogroup 3 chromosomes in general had a much older coalescence with all other haplogroups with an average coalescence date of 18,672 *YBP*. The oldest split was between hg 3 and hg 12 chromosomes with a coalescence time calculated as 21,625 *YBP* (CI 9,268-75,688 *YBP*). This coalescence time was the oldest detected between all haplogroup comparisons. The youngest coalescence time with hg 3 was that between hg 3 and hg 1 with an estimated split occurring approximately 14,554 *YBP* (CI 6,237-50,938 *YBP*).

The most recent split involving hg 8 chromosomes was that with hg 2 approximately 8,536 *YBP* (CI 3,658-29,875 *YBP*). The next youngest split was between hg 8 and hg 21 chromosomes about 9,583 *YBP* (CI 4,107-33,542 *YBP*). The oldest coalescence was that between hg 8 and hg 3, estimated at 20,500 *YBP* (CI 8,786-71,750 *YBP*).

Haplogroup 12 and hg 16 diverged more recently than any other two haplogroups. The TMRCA for hg 12 and hg 16 was estimated at 2,976 *YBP* (CI 1,276-10,417 *YBP*). Haplogroup 12 chromosomes had their most recent common ancestor approximately 21,625 *YBP* (CI 9,268-75,688 *YBP*) with hg 3, the oldest split between hg 12 chromosomes and any other haplogroup.

Haplogroup 16 chromosomes had their most recent ancestor common with hg 12 chromosomes (as described above). The oldest split between hg 16 chromosomes, like hg 12 chromosomes, was between hg 16 and hg 3 approximately 19,446 *YBP* (CI 8,334-68,063 *YBP*).

hg	1	2	3	8	12	16	21	26
1	**3800** ***1600-13300***							
2	11200 *4800-3900*	**8100** ***3500-28500***						
3	14600 *6200-50900*	15000 *6400-52500*	**17500** ***7500-61100***					
8	15700 *6800-55000*	8500 *3600-29900*	20500 *8800-71800*	**4800** ***2000-16700***				
12	3800 *1600-13400*	11300 *4800-39500*	21600 *9300-75700*	14500 *6200-50800*	- -			
16	5300 *2300-18500*	12900 *5500-45000*	19400 *8300-68000*	15000 *6400-52500*	3000 *1300-10400*	- -		
21	13700 *5800-47800*	6900 *2900-24000*	18600 *8000-65000*	9600 *4100-33500*	13000 *5600-45700*	16700 *7200-58500*	**5900** ***2500-20500***	
26	7700 *3300-26800*	13400 *5700-46700*	21000 *9000-73600*	17200 *7400-60300*	4700 *2000-16600*	7300 *3100-25500*	14125 *6000-49400*	**8400** ***3600-29300***

Table 4.17: Time to most recent common ancestor between different haplogroups estimated from the ASD measure and assuming a 0.21% mutation rate and a 25 year generation time. Below the dates are the upper and lower confidence intervals calculated from 95% confidence limits of the mutation rate *(Heyer et al 1997)*. Dates in bold type are the coalescence times of chromosomes within each discrete haplogroup.

The most recent split between hg 21 chromosomes was that with hg 2 estimated at 6,851 *YBP* (CI 2,936-23,979 *YBP*) and slightly later then with hg 8 about 9,583 *YBP* (CI 4,107-33,542 *YBP*). Again, the most ancient split with hg 21 chromosomes was that between hg 21 and hg 3 estimated at 18,554 *YBP* (CI 7,952-64,938 *YBP*).

Haplogroup 26 and hg 12 diverged approximately 4,738 *YBP* (CI 2,031-16,583 *YBP*). The estimated oldest coalescence with hg 26 chromosomes was between hg 26 and, again, hg 3 approximately 21,030 *YBP* (CI 9,013-73,604 *YBP*).

4.3.15. TMRCA within populations: A comparison of diversity

The population history of the Basques has been investigated in a number of genetic systems and is well known (see **section 1.1.11**). That the Basque population has comparable hg 1 frequencies with Irish populations is interesting and may suggest an ancient relationship. The phylogenetic and the population dynamics of the Basques have been studied here alongside the Irish populations. In order to investigate this relationship further, attempts were made to date the Basque and Catalan populations using the available data and comparable dating methods to those above.

When the ASD method was applied the relationship

$$ASD = 2\mu t$$

was employed in order to account for intrapopulational variance.

For direct comparison purposes, the 4 loci in common with the Basque, Catalan and Irish data were used. In all cases the Basque and Catalan coalescence times were older than the Irish coalescence times. The TMRCA estimated by the ASD method for the Basque population was 6,256 *YBP* (CI 2,681-21,896 *YBP*) and for the Irish population the TMRCA was 3,101 *YBP* (1,329-10,854 *YBP*). Generally, all ASD dates calculated using different combinations of loci determined the Basque population to be approximately twice the age of the Irish population. Dating was performed using the full data sets for each population, in the case of the Basque and Catalan populations this consisted of 5 locus haplotypes, and for the Irish, 6 locus haplotypes. As aforementioned, only 4 of these loci were common between all data sets.

The ASD method was employed to estimate directly comparable dates for the three populations, Basque, Catalan and Irish, and also for the hg 1 population of chromosomes, using the 4 comparable loci. The Basque population was the oldest population with a coalescence of 6,256 *YBP* (CI 2,656-21,896) and the Catalan was the next oldest with a date of origin for all its chromosomes estimated at 5,548 *YBP* (CI 2,378-19,417 *YBP*).

Haplogroup 1 chromosomes were not as old (4,786 *YBP* CI 2,051-16,750) as either of the Basque or Catalan populations but were older than the Irish population (3,101 *YBP* CI 1,329-10,854). It is worth noting that hg1 chromosomes (the majority found within Ireland) are older than all the diversity found in the Irish population.

Dating method	Basque (4 loci)	Ireland (4 loci)	Basque (5 loci)	Ireland (6 loci)
ASD	6256	3101	5298	2536
	2700-21900	*1300-10900*	*7500-61100*	*1100-8900*
B&C	4616	3567	4373	3003
	2000-16200	*1500-12500*	*1900-15000*	*1300-10500*
Goldstein	5407	2665	7335	2001-
	1900-33500	*1000-15000*	*2600-46300*	*700-10200*

Table 4.18: Directly comparable dates for the coalescence of Y chromosomes within the Basque and Irish populations. All dates were calculated using the mutation rate of Heyer *et al.* (1997) and a generation time of 25 years *(Thomas et al. 1998).* The Basque population is approximately twice as old as the Irish population.

In order to draw comparisons with dating methods and coalescence estimates for the Basque and Catalan populations by Pérez-Lezaun *et al. (1997),* two alternative mutation rates *(Cooper et al 1996)* were used. Both of the alternative mutation rates estimate older dates and are therefore not as conservative as the rate used above.

	Catalan (5 loci)	Basque (5 loci)	Ireland (6 loci)	hg 1 (6 loci)
ASD	0.756	0.890	0.420	0.319
TMRCA (2.10×10^{-3})	4,494	5,298	2,536	3,798
TMRCA (1.03×10^{-3})	9,174	10,805	5,097	7,747
TMRCA (1.20×10^{-4})	78,750	92,708	43,750	66,500

Table 4.19: TMRCA estimated in the Catalan, Basque and Irish populations and in global hg 1 chromosomes using a variety of mutation rates: 2.10 x 10-3 (Heyer *et al.* 1997), 1.03×10^{-3} and 1.2×10^{-4} (Cooper *et al.* 1997).

4.3.16. Haplotype and surname sharing

Thirty different surnames were sampled more than once in Ireland. Of these surname groups, 10 shared the same Y chromosome 6-locus STR haplotype. For example, both Bagnall individuals had ht 27 chromosomes, both Doyle individuals were ht 118 and

both Rafter individuals were ht 63. Overall, therefore, 33.3% of shared names shared identical Y chromosomes.

4.4 Discussion

4.4.1. Distributions of alleles at Y STR loci

Y chromosome STRs are powerful tools for determining recent population histories. The Y chromosome microsatellite locus utilised most widely in population genetic studies is DYS19, for which there is an abundance of information regarding the geographic distribution of its alleles *(Hammer and Horai 1995; Ciminelli et al 1995; Zerjal et al. 1997; Roewer et al 1996; Hammer et al 1997; Pérez-Lezaun et al 1997; Santos et al 1996; de Knijff et al 1997, Sajantila et al 1996).* The DYS19 locus contains between 10 - 18 GATA repeats, the frequency of each allele being unevenly distributed among global populations. The most common allele at DYS19 here is the 190 bp allele which is found at frequencies of over 75% in the English, Scottish and Irish populations and at 46% in the Turkish population. Other studies have also shown that the 190 bp allele is modal in European populations *(de Knijff et al 1997, Sajantila et al 1996, Hammer et al 1997, Underhill et al 1996).* The 194 bp allele is the most frequent in the Togolese population (56.8%) and has been found to be the most common in Asian and Oceanasian populations also *(Underhill et al 1996, Hammer et al 1997).* According to Sajantila *et al. (1996)* and Underhill *et al. (1996)*, the most frequent allele in Sub-Saharan Africans is the 198 bp allele (37% and 38.5% respectively).

Variable allele distributions at the DYS390 locus indicate an allele specific to the African continent. The 203 bp allele is predominant in the Togolese population (66.7%) but is rare and almost absent in European populations. For example, in the Irish population the 203 bp allele is found at a frequency of 0.8%. Kayser *et al. (1997)* also detect a predominance of allele 203 bp in the African Ovambo (76%) and Pygmy (32%) populations. The prevalence of the 203 bp allele in African populations and its virtual absence in European populations suggests that this allele is African specific.

None of the other Y STR loci demonstrated population specific alleles. In order, therefore, to make evolutionary and historical inferences about the populations Y STR loci must be used as units of composite haplotypes. The following discussion pertains to the use of Y STR loci in complex 6-locus haplotypes. Additionally, the haplotypes have been partitioned according to their occurrence in discrete lineages as discerned by SNPs.

	186 bp	190 bp	194 bp	198 bp
North Europe[H]		58.1		
South Europe[H]		36.8		
East Asia[H]			46.3	
South Asia[H]			53.5	
Australasia[H]			38.2	
Inuit[deK]	77.5			
Samoa[deK]				80.0
Europe[deK]		49.9		
India[deK]		35.4		
Finland[S]		96.0		
Estonia[S]		35.0	35.0	
Saami[S]		68.0		
Sweden[S]		62.0		
Switzerland[S]		53.0		
Basque[S]		76.0		
Sub-Saharan Africa[S]				37.0
Africa[U]				38.5
Asia[U]			47.5	
Oceania[U]			61.5	
Europe[U]		77.7		
America[U]	71.0			
Ireland*		83.9		
England*		75.0		
Scotland*		82.4		
Norman/Norse*		83.3		
Turkey*		46.0		
Togo*			56.8	

Table 4.20: Modal allele frequencies at DYS19 in a number of ethnically diverse human population groups. Data from Hammer *et al. (1997)* [H], de Knijff *et al. (1997)* [deK], Underhill *et al. (1996)* [U], Sajantila *et al. (1997)* [S], and this study *. In all of the European populations the 190 bp allele is the most common. In the African and Asian populations alleles 194 bp and 198 bp are the more common.

4.4.2. Is the stepwise mutation model for microsatellites appropriate for Y STRs?

A stepwise mutation model has been used to best describe the mutational mechanism of STR loci *(Weber and Wong 1993, Brinkmann et al 1995, Brinkman et al 1998).* Most commonly the number of repeats increases incrementally in single repeat units *(Weber and Wong 1993, Di Rienzo et al 1994),* although there may be restraints on expansion. Bimodal distributions of alleles at STR loci may suggest a departure from the stepwise model for mutation at these loci and might suggest a possible, uncommon occurrence of larger, irregular, deletions and insertions. Bimodal distributions at Y STR

loci have been detected at a number of loci including DYS19 *(Hammer et al. 1997),* DYS388 *(Thomas et al. 1998, Perez Lezaun 1997),* DYS390 *(Forster et al 1998)* and DYS392 *(this study and de Knijff 1997).*

Here, the distribution of alleles at the DYS392 locus might tentatively suggest this departure. In all populations the 248 bp and 254 bp alleles are present but the intermediate 251 bp allele is uncommon and, in some populations, absent. This is most obvious in the Scottish, Norman and English populations (see **figure 4.5**) and is illustrated to a lesser extent in the other populations. At all of the other loci (except DYS390) the African population either shares the same MFA or differs by one repeat variant from the European populations. At DYS392, however, the MFA in the African population, is either similar (in the case of the Asian, Danish, Scottish and Turkish populations) or differs by two or more repeat units (in the case of the English, Norman/Norse and Irish populations). De Knijff *et al. (1997)* also detect this bimodality in all regional populations, the common alleles being 11, 13 and 14 (equivalent to alleles 248 bp, 254 bp and 257 bp) with a less common occurrence of allele 12 (equivalent to allele 251 bp).

Similarly, Forster *et al. (1998)* detected a non stepwise signature in the distribution of DYS390, which was most evident in Australasian populations. Small allelic variants, not detected in this study, were found in Australians at low frequencies. Alleles 18, 19 and 22-27 were detected in Australians, but intermediate alleles 20 and 21, which were found in other Asian populations, were absent. On sequencing the DYS390 locus in a number of populations, Forster *et al. (1998)* identified a complex internal structure of the repeats. They suggest that allele calling by virtue of the number of repeats at the locus may be incorrect. The internal variability at DYS390 may therefore render it unsuitable for population studies attempting to identify differences between diverse ethnic populations by employing the stepwise model of mutation for STR loci. In another recent population genetics study, Thomas *et al. (1998)* discounted the DYS388 locus from a set of Y STRs on the basis of its bimodality.

4.4.3. Genetic diversity measures at Y STR loci

The Y chromosome has often demonstrated lower nucleotide diversity than both autosomes and the X chromosome *(Jakabuczka et al 1989, Ellis et al 1990, Malaspina et al 1990, Spurdle and Jenkins 1992).* The lack of recombination on the Y chromosome, except at the small pseudoautosomal region at the end of the chromosome, the smaller effective population size of the Y chromosome, and the reproductive success of only a few males may lead to lower diversity levels. However, the Y chromosome is subject to a higher mutation rate due to its residence in male lineages and the known mutagenic nature of spermatogenesis.

The reduced diversity on the Y chromosome has also been attributed to a recent selective sweep *(Dorit et al 1995, Whitfield et al 1995).* Dorit *et al. (1995)* sequenced a 729 bp intron upstream of the ZFY zinc-finger exon and a short region of the exon sequence and detected no variation in 38 geographically diverse males. Whitfield *et al (1995)* have also maintained that the detection of only 3 substitutions in 18.3 kb of sequence from a non-coding region of Yp could be the result of a recent selective sweep.

Underhill *et al. (1996)* observed a 7.9 fold reduction in predicted polymorphism on the Y chromosome compared to autosomes, but found no evidence for this to have resulted from a selective sweep. More recently, Underhill *et al. (1997)* detected much greater levels of polymorphism on the Y chromosome using the technique of denaturing high-performance liquid chromatography. In ~ 9,500 bp of sequence in each of ~ 50 males, they detected 22 polymorphisms. They estimated that one nucleotide difference between two randomly chosen individuals is expected every 3.2×10^3 - 3.8×10^3 nucleotides. This is comparable to polymorphism frequencies on autosomes.

Other authors, *(Poloni et al 1997, Hammer 1995, Paabo 1995)* report results that are inconsistent with selection for a favourable mutation leading to a selective chromosomal sweep on the Y chromosome. For example, Hammer *(1995)* sequenced a 2.6 kb region on the Y chromosome, incorporating the Y Alu insertion element, in 16 human males and 4 chimpanzee males. By comparing the diversity at mtDNA loci and at Y chromosome loci using a π : D ratio (where π is the nucleotide difference among humans and D is the divergence between humans and chimpanzees), and applying the HKA test for selection *(Hudson et al 1987),* Hammer *(1995)* rejected the hypothesis of the fixation of a favourable mutation on the human Y chromosome. The π : D ratio at the

YAP locus was 1 : 51, at the mtDNA COII locus it was 1 : 54 and at the mtDNA ND4-5 locus it was 1 : 40. Comparisons of diversity were also made with autosomal loci, the effective size of the chromosomal populations being accounted for. Once more the hypothesis that lower sequence variation on the Y chromosome has resulted from a selective sweep was rejected.

This study, together with Sajantila *et al. (1996)* and Pérez-Lezaun *et al. (1997),* also argues against the likelihood of a recent selective sweep on the Y chromosome by demonstrating similar diversity at Y STRs and autosomal loci.

The effective number of Y chromosomes in the population is one quarter that of autosomes. At equilibrium the diversity at an autosomal locus can be estimated as

$$D = 4N\mu / (4N\mu + 1)$$

where N is the number of individuals and μ is the mutation rate *(Hartl and Clark 1989).*

Assuming the same population size and a similar mutation rate the diversity at Y loci can be equated with the diversity at autosomes by applying the following equation:

$$D_{au} = 4D_y / (3D_y + 1)$$

where D_{au} is the equivalent diversity at autosomal loci and D_y is the diversity at Y chromosome loci.

Averaged across all populations tested in this study, the diversity at Y STR loci was estimated as 0.498. When corrected for effective size, the mean D value for Y STRs becomes 0.799. This value is in agreement with the diversity value estimated by Pérez-Lezaun *et al. (1997)* for 20 unlinked autosomal STRs (0.730). In the same populations Pérez-Lezaun *et al. (1997)* estimated Y STR diversity (corrected for effective number) as 0.755. Additionally, projecting figures from Sajantila *et al. (1996*), the diversity for one Y STR and two Y UEPs (corrected for effective number) is estimated as 0.836. These figures are all congruous and do not suggest lower diversity at Y chromosomes compared to autosomes. If mutation rates at Y STRs were much faster than at autosomal STRs then a selective sweep may indeed be possible, the diversity having been regenerated. However, the best estimates for mutation rates at Y STR and autosomal STR loci are similar *(Weber and Wong, 1996, Heyer et al 1997).*

	Y	Autosomes	Y (corrected)	Autosomes
This study	0.498	-	0.799	-
Sajantila *et al. (1996)*	0.561	-	0.836	-
Perez-Lezaun *et al. (1997)*	0.435	0.750	0.755	0.750

Table 4.21: Relative genetic diversities on Y chromosomes and autosomes. The first two columns show the raw diversity on each chromosomal type (where data is available) in each of three studies. The second two columns show the estimated genetic diversity following correction for effective number of chromosomes. The estimates of diversity in each study were derived from the following populations and loci: Sajantila *et al. (1996),* Finns, Estonians, Saami, Swedes, Swiss, Basques and Sub-Saharan Africans, 1 Y STR /2 Y UEPs; Pérez-Lezaun *et al. (1997),* Basques and Catalans, 8 Y STRs and 20 autosomal STRs; this study, Togolese, Danes, English, Scottish, Normans/Norse, Turkish and Irish, 6 Y STRs.

4.4.4. The usefulness of Y STRs in the discrimination of closely related populations.

The higher level of genetic diversity in the Togolese population at X STR loci concords with a greater antiquity for African populations relative to Europeans and may suggest an African origin for modern humans (see **table 2.11**). Higher diversity levels in African populations have been detected before using X STR loci *(Scozzari et al. 1997),* mtDNA *(Vigilant et al 1991)* and autosomal minisatellites and microsatellites *(Bowcock et al. 1994, Armour et al 1996, Tishkoff et al. 1996).* In contrast, a greater diversity in African populations than non-Africans is not detected using Y STR loci *(this study, Scozzari et al 1997).* It has been suggested that this may be a result of a lower effective population size of Y chromosomes in Africa *(Torroni et al 1990, Spurdle et al 1994, Jobling and Tyler-Smith 1995).*

Although diversity values are not as high in the African population as expected, Africans are still separated on a phylogeny from all non-Africans and the greatest genetic distances, and therefore the most ancient divergences, are those between Africans and non-Africans.

Y STRs have been of utility in the determination of more recent, historical population relationships *(this study, Santos et al 1999, de Knijff et al 1997, Poloni et al 1997).* Y STR haplotypes are strongly associated with population groups and closely related populations may be distinguished by Y STR loci *(Roewer et al. 1996).* In this study, the genetic structure of Y chromosomes was examined by AMOVA, in which a

hierarchical level of variation within and among population groups was assessed. In an initial analysis it was shown that when the subdivision of the total diversity in diverse global populations is considered an estimate of 54.4% of the total variation between human Y chromosomes is found within populations. This is in contrast to similar tests using X STR loci, where the variability within populations was 88.5%, and using autosomal loci, where the fraction of variance within populations was 85% *(Barbujani et al 1997)*. In a previous comparison of the proportioning of variance within and between populations at homologous X and Y STR loci, a greater fraction of within population variance was detected at X STRs than Y STRs in widely sampled populations *(Scozzari et al. 1997)*. For example, at the Y STR locus DYS413, the fraction of the variance described within populations was 79.4%. At the homologous X STR loci, DXS8175 and DXS8174, the proportion of the variance found within populations was somewhat greater with values of 92.0% and 96.0% respectively. Similarly, Santos *et al. (1999)* found only 59% of the total variability in global Y chromosomes among individuals within populations.

By removing the African population from the Y STR data in this study, the variance is partitioned in a similar manner to X STRs and autosomal DNA polymorphisms, where 86.6% of the total variability is that between individuals within populations. This suggests that Y STR haplotype analysis is more efficient for determining the subdivision of variance between more closely related populations than between worldwide populations. For example, populations were grouped in an AMOVA test according to whether surname cohorts were of indigenous Irish or exogenous Irish origin. A significant amount of the total variability (12.3%) was found among groups, suggesting that Irish individuals of Gaelic origin, and Irish individuals of non-Gaelic origin can be distinguished by Y STR loci. De Knijff *et al. (1997)* have also illustrated the power of Y STRs to discriminate between closely related populations and found that a significant proportion (8.1%) of the total genetic variability in Dutch and German populations is found between the two populations. In a study of partitioning of variance akin to this study, Poloni *et al. (1997)* determined that 25% of total human genetic diversity is due to differences among language groups and that only 5% of the total variability is the result of differences among populations within language families.

4.4.5. The origin of Turkish Y chromosomes

Turkey lies at a geographic and historical crossroads between Africa, Asia and Europe. The genetic history of the Turkish population provides evidence for a diverse admixed population, shaped by population migrations and expansions during the Paleolithic, Neolithic and in the more recent, historical past. The Turkish sample, comprised of 6 discrete Y chromosome haplogroups, is the most diverse population group in this study. The diversity within these haplogroups, discerned by assaying rapidly evolving microsatellite loci, is higher than in either the African or other European populations.

Although levels of Y STR diversity are highest in Turkey, other genetic analyses place the Turkish population at a genetic intersection between African and European populations. For example, the variance of allele lengths at X STRs is highest in Africans, lowest in Irish and intermediate in Turks. Additionally, the average gene diversity at X STRs in the three populations illustrates the intermediate position of the Turkish population between the more diverse Africans and the less diverse Europeans. The calculation of genetic distances from Y STR haplotypes again places the Turkish population midway between the African and European populations on phylogenetic trees. Similarly MDS illustrates the intermediate genetic position of the Turkish population in both X chromosome and Y chromosome systems.

Mitochondrial DNA control region sequence analysis also suggests that the Turkish population is intermediate between Africans and Europeans *(Comas et al. 1997).* Neighbour-joining trees, created from distance matrices of Kimura's two-parameter model, place Turkish mtDNA sequences midway between African and European sequences. Comas *et al. (1997)* produced a phylogenetic tree illustrating four main clusters, labelled A-D. In cluster C, the C allele is found in the Caucasoid reference sequence. The polymorphic T allele is found at low frequencies in Europeans (8%, 7% and 4% in British, Sardinians and Basques, respectively), at high frequencies in Africans (90%) and at intermediate frequencies in Turks (29%). Frequencies of polymorphic alleles in the B and D clusters also illustrate the intermediate position of Turkish sequences. Additionally, in a phylogenetic tree relating populations, the Turkish population is found to lie between a Middle Eastern population and Basque, Sardinian and British populations. These authors suggest that Turkey was in a pathway for the colonisation of Europe, and

perhaps also Asia, by modern humans emanating from Africa via the Near East some time between 35,000 and 100,000 *YBP*. The Turkish population, therefore, provides an example of an admixed Asian/European population, located at a geographic crossroads between Africa, Asia and Europe.

A Paleolithic Turkish population

Most models for the emergence of modern humans in Europe posit a migration of anatomically modern man out of Africa initially into the Near East. The earliest modern human remains detected in the Near East are found in Qafzeh, Israel dating to ~100,000 *YBP (Vallades et al 1988),* although the majority of archeological remains of anatomically modern humans in the Near East and Europe date to a later time of ~35,000 - 40,000 *YBP*. This later date coincides with the abrupt disappearance in the archaeological record of archaic hominids in Europe. Most genetic data is consistent with a complete replacement of *Homo sapiens neanderthalensis* by anatomically modern humans at this time. Coinciding with their demise there is genetic evidence for a global modern human population expansion *(Rogers and Harpending 1992, Rogers and Harpending 1996, Rogers et al. 1996, Sherry et al. 1994).*

A population expansion of modern humans in Turkey has been dated to 35,000 - 100,000 *YBP (Comas et al. 1996).* This Paleolithic expansion most likely arose as a result of the success of modern humans in a novel environment to which they had the ability to readily adapt. The detection of expanding Paleolithic mtDNA lineages in Turkey suggests that Y chromosomal lineages might also be traced, in this population, to the same period.

The most likely candidate for a Paleolithic Y chromosome type in Turkey is hg 2. The majority of the Turkish Y chromosome population are comprised of hg 2 Y chromosomes (58%) which are found in most global populations and which are ancient by comparison to other Y chromosome types *(Hammer et al. 1997,1998; Hurles et al 1998; Jobling et al 1994; Jobling et al 1998).* Haplogroup 2 chromosomes are equivalent to haplotype 1B in Hammer *et al. (1998).* Haplotype 1B chromosomes are geographically widespread and are detected in 55.2% of global Y chromosomes. The time to coalescence for global haplotype 1B chromosomes has been estimated as 110,000 *YBP* (+/- 36,000 years). The global distribution and the time to coalescence of 1B chromosomes suggests that they evolved in Africa and were spread throughout the rest of the world during the

initial migration of modern humans out of Africa coinciding with the replacement of archaic hominids.

In the entire European sample hg 2 chromosomes are estimated to have coalesced ~8,000 *YBP* although this is a conservative estimate and the date may be much older (95% CI 3,500 - 28,500 *YBP*). Projections of figures for hg 1 chromosomes may also be used for dating the diversity of hg 2 chromosomes in Europe. For example, when estimating diversity from highly polymorphic internal modular structures of the minisatellite, MSY1, Jobling *et al. (1998)* detected a diversity in hg 2 chromosomes 7.8 times greater than in European hg 1 chromosomes. If European hg 1 chromosomes coalesced ~4,000 *YBP*, then drawing on the data from Jobling *et al. (1998)* the time to coalescence of hg 2 chromosomes lies some time around ~30,000 *YBP*. This date corresponds with the first archaeological evidence for modern humans outside Africa and agrees with the dates for a Turkish population expansion. The relatively low diversity of hg 2 chromosomes compared to global 1B chromosomes may suggest a bottleneck during the founding of European populations from Africa. Alternatively, because diversity here was calculated by the assessment of heterozygosity values using microsatellite loci, the high mutation rate and the recurrent nature of mutation of Y STRs may have erased the traces of such an ancient migration. It is reasonable to suggest that on emergence from Africa, hg 2 Y chromosomes expanded rapidly in a small founding Paleolithic population and that extant hg 2 chromosomes are relics of the ancestral Near Eastern Y chromosome type.

This hypothesis must be consolidated with further genetic evidence. There is little evidence in the phylogenetic network of hg 2 chromosomes to suggest a recent population expansion. By comparison to hg 1 chromosomes, hg 2 chromosomes are diverse and incoherent and exhibit no properties relating to such population growth. For example, following the rapid growth of a population star-like patterns are expected in the networks which reflect lineage survivorship of the ancestral haplotype and those haplotypes most closely related to it. The most common hg 2 haplotype in the Turkish population, ht 12, is isolated on the periphery of the phylogeny and three mutational events are required to relate ht 12 to its closest derivatives. The next most common Turkish hg 2 chromosome, ht 85, is also isolated in the network, and similarly its position does not suggest an expansion of related chromosomes. In contrast, the most common northwest European hg 2 haplotype, ht 22, lies central in the network and common closely related haplotypes lie adjacent to it separated by only one mutational event. This suggests that a more recent

expansion of hg 2 chromosomes may have occurred in extreme regions of the continent but there is no direct evidence to suggest a Paleolithic hg 2 expansion in Turkey.

Pairwise difference distributions are visual summaries of diversity. Visual inspection of the hg 2 pairwise difference distribution chart clearly suggests an expanding population of chromosomes if, as previously discussed, smooth unimodal distributions reflect population growth and uneven, bimodal distributions reflect populations at equilibrium. By comparison to the other distributions, the hg 2 distribution illustrates the smoothest curve of all Y chromosome lineages. Additionally, it is the distribution most similar to the curve generated from the total sample and it exhibits the most ancient pattern; the modal value (4 differences) of the number of differences being more distal from the vertical axis of the chart than any of the other distributions. Although no conclusive predictions may be made, it is likely that hg 2 chromosomes underwent rapid growth sometime after the emergence of modern humans in the Near East ~30,000 *YBP*, and that the signature of this expansion is difficult to discern in the limited phylogenetic sample used here.

The Neolithic in Turkey

Haplogroup 21 chromosomes are found in the Turkish population at a moderate frequency (22%) and are the only Y chromosomes detected here in both African and non-African populations. Additionally, hg 21 is the only YAP+ lineage found in European populations. Two alternative hypotheses may account for the presence of hg 21 in Africa and Europe. If a postulated Asian origin for YAP+ chromosomes is correct then either hg 21 chromosomes evolved in Africa and spread into Europe and the rest of the world following the reintroduction of Y chromosomes into Africa from Asia, or the hg 21 lineage evolved prior to the reintroduction of Y chromosomes into Africa (see discussion **chapter 3.4**). The origin of the ancestral YAP+ haplotype (3G in *Hammer et al. 1998*) has been identified in Asia on the basis of frequency data and a recent nested cladistic analysis *(Hammer et al. 1998).* The coalescence date for the most ancestral YAP+ chromosomes, 3G, has been estimated as 55,000 *YBP* (+/- 19,000 years). The origin for the derived YAP+ lineage, hg 21, has been identified in North Africa ~20,000 *YBP* from where it spread into the Middle East ~10,000 *YBP*. Here, the divergence of hg 21 from hg 2 chromosomes is estimated to be younger, ~7,000 *YBP*.

The years spanning 7,000 - 10,000 *YBP* were probably the most important time in the history of modern humans in the Near East. The innovation of agricultural technologies has been described as *"a revolution whereby man ceased to be purely a parasite and ... became a creator emancipated from the whims of his environment" (Childe 1928).* This may not be entirely true, as human settlements, including those associated with agriculture, have constantly been challenged and influenced by the environment.

At the end of the last Ice Age, ~12,500 BC, there was an abundance of human settlements throughout Anatolia. One of the best known settlements is Abu Hureya on the Euphrates river which dates to 11,500 *YBP (Fagan 1980).* The inhabitants of Abu Hureya were Natufian peoples who lived by means of hunting and gathering. Natufian societies were also present at other sites throughout the Fertile Crescent such as Jericho, Tell Aswad, Ain Mallaha, Beidha and Mureybet *(Ryan and Pitman 1998).* Between 10,500 - 9,400 BC there was a change in global temperature associated with the Younger Dryas, a time when the glacial ice sheets advanced southwards into Europe once more. During this time there is evidence for an abandonment of Natufian settlements due, most likely, to depleting natural resources associated with cooling temperatures and increased aridity *(Moore and Hillman 1992).* At Abu Hureya evidence for human occupation disappears before 9,500 BC.

It has been suggested that the inhospitable climate of the Younger Dryas may have encouraged the cultivation of local crops *(Bar-Josef and Belfer-Cohen 1989)* and the development of agricultural technologies. Many theories exist as to the exact location of the emergence of agriculture in the Near East, but broadly speaking most evidence indicates that the initial domestication of most plant and animal species took place in the Fertile Crescent region.

The most widely known Neolithic site in Anatolia is Çatal Hüyük located in the southern Konya plain. Çatal Hüyük was discovered in 1958 by James Mellaart and immediately identified as a Neolithic town dating to later than 9,000 *YBP (Mellaart 1967).* However, there is evidence for a desertion of the Neolithic site ~6,200 BC, without evidence for violence or destruction, which coincides with a second mini Ice Age in Europe *(Dansgaard et al. 1993).* As with Abu Hureya, environmental conditions forced the abandonment of settlements in the search for a more sustainable environment. The lowland shores of the Black Sea, then a freshwater lake, might have been a refuge for

agriculturalists during unfavourable environmental conditions. DNA sequence evidence from the earliest domesticated einkorn wheat suggests that early farming villages on the shores of the Black Sea were inhabited by agriculturalists who originated further south in Anatolia *(Heun et al. 1997).*

Renfrew *(1989)* argues that Indo-European languages spread, from a center of origin, northwards into Europe, Dravidian languages spread eastwards towards India and Pakistan and Afro-Asiatic languages spread southwards towards Arabia and North Africa. Recent evidence for Anatolia being central in the dispersal of languages comes from a recent genetic study of Y chromosomes. Poloni *et al. (1997)* used linguistic affinities to classify population groups in a genetic analysis of the p49a,f/TaqI polymorphic system. In a MDS analysis the Turkish population was found at an intersection between Indo-Europeans and Afro-Asiatics and clustered with Sephardim and Ashkenazim Jews and Lebanese.

The clinal distribution of hg 21 chromosomes in Europe may suggest the northward migration into extreme regions of the continent. Haplogroup 21 chromosomes share a recent common ancestor ~6,000 *YBP*. Significantly this date is in agreement with the postulated date for the dispersal of agriculturalists throughout Europe. It may be that proto-Indo-European speaking peoples, concentrated in Anatolia at this time, contained Paleolithic hg 2 Y chromosomes and the putative Anatolian Neolithic Y chromosome type, hg 21. This is suggested by the gradient of hg 21 frequencies (**figure 3.5**) which decrease in a cline from Turkey in the southeast to populations in northwest Europe. For example, compared to 22% hg 21 chromosomes in Turkey, only 8.3% of the Danish population have hg 21 Y chromosomes and the English population has even less (5%). Drawing on data from Hammer *et al. (1997; 1998),* YAP+ haplotypes 3A and 4 are equivalent to hg 21. Haplotype 3A and 4 frequencies are high in southern Europe (33% in Greeks, 31% in Italians) and low in northwest Europe (7% in British, 10% in German). This gradient has been attributed to the demic expansion of Neolithics from the southeast into Europe ~10,000 *YBP*.

It might be feasible to envision bands of agriculturalists migrating northwards admixing with indigenous Mesolithic peoples, consigning fewer hg 21 chromosomes to indigenous peoples as further regions are reached and at the same time assisting in the establishment of the gradient of hg 1 chromosomes. This hypothesis requires further testing with populations neighbouring Turkey and requires frequency data for hg 21

chromosomes in Dravidian and Afro-Asiatic speaking populations. If the peoples inhabiting Anatolia during the Neolithic period contained a large proportion of hg 21 chromosomes, it is predicted that a southern and eastern cline of phylogenetically similar hg 21 chromosomes would be evident, hg 21 frequencies decreasing with increased distance.

Recent Asian influences in Turkey

Haplogroup 1 chromosomes are uncommon in the Turkish sample in whom they comprise only 8% of all Y chromosomes. Haplogroup 1 is the most widespread Y chromosome type in Europe and it follows that Turkish hg 1 chromosomes most likely originated in Europe. However, in a phylogenetic analysis it has been shown that Turkish hg 1 chromosomes do not cluster with other European hg 1 chromosomes. Additionally, no Turkish hg 1 chromosomes are shared in other European populations. Asian hg 1 chromosomes, which make up 22% of Indian and Sri Lankan populations, are not phylogenetically similar to European chromosomes *(Jobling et al 1996)* and it is possible that Turkish hg 1 chromosomes may have an Asian origin. Additionally, Asian hg 1 chromosomes are considerably more diverse (0.818) than European hg 1 chromosomes (0.083) *(Jobling et al 1998)* and are the oldest hg 1 chromosomes in this study, dating to 14,800 *YBP*. This scenario is consistent with a minor Asian influence in Turkish peoples.

Candidate immigrants carrying hg 1 chromosomes from Asia include the Seljuk and Ottoman Turks who entered Anatolia between 1081 AD and 1453 AD. These bands of Turkish nomads originated in Central Asia near Mongolia. The Turkic and Mongolian language subfamilies, together with the Tungus subfamily, are closely related and are both members of the Altaic family of languages. Prior to the Turkic invasion, the Indo-European language, Greek was spoken in the Byzantine enclave of Turkey. Therefore, the language spoken in modern day Turkey is a recent adoption of a language of exogenous origin and it has been suggested that it was incorporated through a mechanism of elite dominance associated with little genetic influence *(Cavalli-Sforza et al. 1994)*. It seems likely that hg 1 chromosomes in Turkey were incorporated with the bands of Turkoman nomads who entered Anatolia in the second millennium AD replacing the indigenous Byzantine language with an Altaic Turkic tongue. The genetic data suggests a cultural

diffusion of the new language, rather than a demic replacement, which was accompanied by a practically complete dilution of the incoming Turkic Mongoloid genes.

Few hg 26 chromosomes are found in the populations studied here, the greatest frequency being found in Turkey (8%). It is therefore difficult to make predictions about the origins of this haplogroup. Haplogroup 26 chromosomes are abundant in East Asia *(Hurles et al. 1998)* but are uncommon in the European and African data. Three of the six Asian chromosomes in this study had hg 26 Y chromosomes. The hg 26 chromosomes were from China and Malaysia.

It has been estimated that hg 26 and hg 2 chromosomes shared a common ancestor ~13,400 *YBP*. All hg 26 chromosomes coalesce ~8,400 *YBP* but the geographic origin of the hg 26 mutation is unknown. It is possible that hg 26 chromosomes diverged from hg 2 chromosomes in the Near East during the late Paleolithic and were spread from there into Asia. A more plausible hypothesis is that hg 26 chromosomes evolved from hg 2 chromosomes in Asia and the presence of hg 26 in Turkey and Europe reflects later migrations back towards Europe after the initial colonisation of the Asian continent, perhaps with the expansion of the Kurgan peoples.

It has been suggested that hg 12 and hg 16 chromosomes evolved in Mongolia from hg 26 chromosomes in the last 2,400 - 4,000 years *(Zerjal et al. 1997)*. In perfect agreement with these dates, here the divergence of hg 12 from hg 26 is estimated at 4,700 *YBP* and the divergence of hg 16 from hg 12 is ~3,000 *YBP*, although the geographic location for the original mutations cannot be identified. The only hg 12 chromosome in this study is found in the Turkish population. The microsatellite ht 77, on the Turkish hg 12 chromosome, is identical to haplotype T1 in the study by Zerjal *et al. (1997)* which was found in two Khalkh Y chromosomes of Mongolian origin. Notably, the Khalkh population is quite diverse, 27 different microsatellite haplotypes being detected in 38 hg 12 Mongolians. Additionally, a high frequency (86%) of hg 12 chromosomes has been detected in Turkic speaking Yakut peoples *(Zerjal et al. 1997)*. By analogy to this data, it is conceivable that the one hg 12 chromosome sampled in the Turkish population is of Mongolian origin, entering Turkey with Turkic speaking peoples at the same time as the introgression of hg 1 Y chromosomes.

4.4.6. The origins of Togolese Y chromosomes

Greater genetic diversities in African populations have been attributed to a number of demographic scenarios. High diversity generally suggests a greater antiquity for a population because it has had more time to accumulate mutations *(Stoneking and Cann 1989)* and African populations are generally regarded to be the oldest modern human populations. Alternatively, the greater diversity in Africans has suggested an older population which gave rise to smaller founder populations within Africa which migrated into Europe and Asia carrying only subsets of the total African diversity *(Relethford et al 1995)*. Another explanation suggests that a bottleneck in the population emerging from Africa into Asia and Europe accounts for the decreased diversity in non-African populations and a relatively higher diversity in Africans *(Relethford and Jorde 1995)*.

West African X chromosomes, sampled in Togo, are more phylogenetically diverse than European X chromosomes but, conversely, Togolese Y chromosomes are not as diverse as those in some European populations. Similarly, allele length variance is less in the African Y chromosomes than some European Y chromosomes but greater variance is found in Togolese X chromosomes than corresponding European X chromosomes. Although differences in diversity may be accounted for by differential mutation rates at STR loci on the sex chromosomes, without testing, further explanation is required.

The conflicting X and Y chromosome data for the Togolese population may be resolved by examining the origins of the specific chromosomal lineages within the population. The highest frequency (20%) for the most ancestral Y chromosome has been detected in Khoisan peoples *(Hammer et al 1998)* suggesting that Africans are the oldest human populations. Only 3% of West Africans have the ancestral YAP- chromosome *(Hammer et al. 1998)*. No YAP- chromosomes were detected in the Togolese population, including the deepest root of the Y chromosome tree, hg 2. The absence of ancient lineages suggests that Y chromosomes in this population are relatively young and that the origin of YAP+ chromosomes in Asia (55,000 *YBP*) predates the origin of this population. The presence of the recent YAP+ derivatives, hg 21 and hg 8, in the Togolese population suggests it as a young population *(Hammer et al. 1998)*.

It has been estimated that the hg 21 lineage arose in North Africa from the ancestral YAP+ lineage ~20,000 *YBP (Hammer et al. 1998)*. Here, a common ancestor for YAP- and hg 21 chromosomes has been estimated to be much younger ~7,000 *YBP*,

although this is likely to be an underestimate. If, as suggested, hg 21 chromosomes evolved in North Africa, the presence of this lineage in Europe and Africa suggests a recent migration of hg 21 both northwards into Europe and south and west within Africa from its center of origin.

A number of analyses suggest the presence of two possible hg 21 sublineages, one specific to Europe and another present only in Africa. For example, in a phylogenetic analysis of hg 21 chromosomes, Togolese hg 21 chromosomes do not cluster with European hg 21 chromosomes, except for one occurrence of an African haplotype, ht 93, in the Turkish cluster. Haplotype 1, ht 10a, ht 14 and ht 86 are African-specific and the remainder are found in Europe only. By visual inspection of the phylogeny, the Togolese and European haplotypes are distantly related. Additionally, on inspection of the pairwise difference distribution for hg 21 chromosomes a bimodal distribution is clear. Although, as discussed, uneven bimodal distributions are consistent with stable non-fluctuating population sizes, bimodal mismatch distributions may also result from the presence of more than one chromosomal lineage. It is likely that the bimodal distribution illustrated with hg 21 chromosomes reflects the presence of two sublineages, one of which contains a subset of the total hg 21 diversity.

Diversity levels in Sub-Saharan African, North African and European hg 21 chromosomes have been estimated before *(Hammer et al 1997).* North Africans have the highest diversity (0.614), Europeans the lowest (0.275) and Sub-Saharan Africans intermediate levels (0.416). It is plausible that a number of small population bottlenecks have shaped the uneven distribution of diversity of hg 21 chromosomes. For example, the diversity present in North African hg 21 chromosomes has arisen over ~20,000 years. The migration of hg 21 chromosomes into the Near East ~10,000 *YBP*, as suggested by Hammer *et al. (1997),* may conceivably have preserved much of the diversity present in North African hg 21 chromosomes. During the Neolithic expansion of agriculturalists from the Near East into the rest of Europe, a population bottleneck reducing the diversity of European hg 21 chromosomes compared to that in the Near East might have occurred. Additionally, extensive admixture with indigenous Mesolithics might have decreased not only the frequency of hg 21 chromosomes (46.2% in North Africans, 4.3% in Northwest Europeans; *Hammer et al. 1997*) but also the diversity inherent within that lineage. A second potential population bottleneck is suggested by the decreased diversity of hg 21 in Sub-Saharan Africans compared to North Africans and it is this proposed bottleneck that

may have contributed to the decreased diversity in Togolese Y chromosomes. The lesser diversity in the Togolese population may therefore represent a subset of the diversity that is present in the Turkish population. Comparable data from North African populations is required to substantiate this hypothesis.

However, the coalescence dates and the occurrence of more ancient Y chromosome lineages in the Turkish population also suggest this. For example, ancient hg 2 Y chromosomes are absent in Togo but are predominant in Turkey. It has been shown that the majority of Turkish Y chromosomes originated ~30,000 *YBP*, whereas Y chromosomes in Togo shared a common ancestor less than 10,000 *YBP*.

The predominance of the most recently evolved YAP+ lineage, hg 8, in the Togolese population (83.8%) also suggests that this African population has a relatively recent origin. The TMRCA of hg 8 chromosomes has been estimated as 4,800 *YBP* which, apart from hg 1 (3,800 *YBP*), is the youngest lineage sampled in this study. A higher frequency and greater diversity of hg 8 chromosomes in Sub-Saharan Africans suggests that the origin of the hg 8 lineage was in this region *(Hammer et. al. 1997).* Although no other African data is available, it is possible that hg 8 and hg 21 chromosomes entered Togo with the recent Bantu expansions ~2,500 *YBP.*

The largest Togolese tribe, the Ewe, speaks a south-central Niger-Congo language, a subfamily of the Niger-Kordofanian language family that includes the widespread Bantu language group *(Cavalli-Sforza 1994).* The original homeland of Bantu languages may have been in northern Zaire *(Fagan 1980)* from where they spread successfully into many regions of Africa displacing the indigenous San hunter-gatherers and introducing new ironworking and agricultural technologies. It has been suggested that West Africa may have been one of the first regions to benefit from novel agricultural technologies *(Cavalli-Sforza et al 1994)* which may have encouraged a rapid population expansion in these regions. The high frequency of hg 8 in West Africans supports this view *(Hammer et al. 1998).* The Ewes, who were hunters and farmers moved to the region that is now Togo from the Niger river valley between 1100 - 1300 AD and may have adopted the Bantu iron and agricultural technologies prior to migration to the present location. Cavalli-Sforza *et al. (1994)* suggest two independent expansions in West Africa resulting from the introduction of agriculture, one in Senegal and the other in the Niger – Mali – Burkino Faso region. They suggest that the latter expansion gave rise to the Ewe tribe.

Visual inspection of the Togolese Y chromosome phylogeny suggests a recent expansion of hg 8 chromosomes in this population possibly linked to the recent adoption of agricultural techniques. The most frequent haplotype, ht 28, is central in the network and many closely related haplotypes radiate from the ancestral node. This suggests a recent rapid population growth of hg 8 chromosomes in this population supported by the even unimodal distribution of the pairwise difference distribution for the hg 8 lineage.

The lesser diversity at Y STRs in the Togolese population might therefore be explained by the recent origin of Y lineages in this population. The oldest conceivable date for the Togolese population, on the basis of the high frequency of hg 8 chromosomes, does not exceed 10,000 *YBP*. In contrast, Turkish Y chromosomes date to possibly over ~30,000 *YBP* with the first emergence of modern humans in the Near East. Additionally, the Turkish population has been influenced by many diverse Y chromosomes in a multitude of waves of immigration both from Africa and Asia whereas the Togolese population is a comparatively homogenous recent population. Additionally, the sampling of the African population was carried out in one area, Sokodé, Togo, whereas Turkish Y chromosomes were sampled in three regions of Anatolia (Izmir, Erzurum and Ankara).

This may explain the uncharacteristic lack of diversity in African Y chromosomes compared to non-African chromosomes but does not entirely explain the greater diversity at X STRs. A more likely explanation is the different effect of reproductive success between males and females. Few highly successful male lineages will be retained in a population, particularly in a society with polygamous mating practices. This might result in local homogenisation of successful Y chromosomes which would show a decrease in diversity within a population and an increase in the genetic differentiation between separated populations

It has been suggested that X chromosomes in Europe and Africa may be subdivided into discrete lineages by using slowly evolving polymorphisms. That there are two Togolese Y chromosome lineages, hg 8 and hg 21, and more than two Irish Y chromosome lineages supports, by analogy, the hypothesis, proposed in **chapter 2,** that there are two African X chromosome lineages and at least two, and possibly three, separate European X chromosome lineages. The assessment of diversity within each X chromosome lineage may reveal a lesser diversity within the Togolese as indicated by examination of individual Y chromosomes. If European diversity is partitioned into three lineages and Togolese diversity into two lineages, a greater insight into the history of X

chromosomes may be uncovered. Additionally, differential processes of mutation and selection acting on the sex chromosomes may result in ambiguous findings of population histories, and until more is known of the specific mutational processes it will be difficult to make clear comparisons between the two systems.

4.4.7. The origins of Irish Y chromosomes.

The origins of the people of Ireland are enigmatic and controversial *(Waddel 1998, Mallory 1991)* and there has been little previous investigation of the Irish population from a genetic perspective. Inhabiting an island on the western edge of Europe, the Irish population may possess a genetic heritage relatively undisturbed by the major demographic movements which have shaped mainland Europe *(Cavalli-Sforza et al. 1994).*

The most striking feature of the genetics of the Irish population is the high frequency of hg 1 chromosomes in Gaelic Irish populations (89.7%) and the near fixation of this haplogroup in the western Connaught population (98.3%). These are the highest frequencies for hg 1 chromosomes in Europe and are close to the estimated frequency for this haplogroup in the putatively Mesolithic Basque population (89%). The European distribution of hg 1 chromosomes has been discussed in **chapter 3**.

That hg 1 chromosomes are widespread in Europe and are also present in Asia *(Jobling et al. 1998)* suggests this haplogroup as the putative European Mesolithic haplogroup; the Y chromosome lineage inherent in the pre-Neolithic hunter-gatherer populations inhabiting the continent prior to ~10,000 *YBP*. The greater diversity of Asian hg 1 chromosomes (0.818) than European hg 1 chromosomes (0.083) as determined by diversity of internal modular structures at the minisatellite MSY1 locus (*Jobling et al. 1998)* and the differential clustering of Asian and European hg 1 chromosomes in phylogenetic analyses *(this study, Jobling et al. 1997)* suggests an ancient divergence of Asian and European hg 1 chromosomes.

The prevalence of hg 1 in Ireland and its close phylogenetic clustering suggests that the Irish population has largely retained its genetic Mesolithic heritage. This might be surprising considering the numerous exogenous influences that have been suggested and documented in this country. This may be comparable to the genetically isolated Basque population. The high frequency of the hg 1 lineage in these geographically peripheral populations suggests a substantial influence of isolation and resistance to introgression.

A recent expansion of Irish hg 1 chromosomes.

Haplogroup 1 chromosomes in Europe are closely related and phylogenetic analysis suggests a recent expansion of this chromosome lineage in Europe, and in particular in Ireland. Haplogroup 1 chromosomes are coherent and cluster together in a network. The common hg 1 chromosomes lie central in the network and star-like patterns are most evident from the two most common Irish haplotypes, ht 128 and ht 115. This phylogenetic pattern suggests a rapid population growth with the survivorship of the ancestral lineages and their closely related derivatives. The peripheral chromosomes in the hg 1 network are Turkish chromosomes and, as suggested, they are not expected to cluster with European hg 1 chromosomes due to their very different origins.

A recent population expansion of hg 1 chromosomes is also suggested by the smooth, even distribution of hg 1 pairwise differences. By comparison with the pairwise difference distribution of hg 2 chromosomes, hg 1 chromosomes have a recent origin. Using the ASD distance, a mutation rate of 0.21% and a generation time of 25 years the date of the hg 1 expansion has been estimated as 3,789 *YBP*. By removing non-Irish hg 1 chromosomes the date for the expansion of hg 1 chromosomes in Ireland is estimated as ~3,900 *YBP*. Non-Irish hg 1 chromosomes are younger by comparison, the bulk of hg 1 chromosomes in the sample outside Ireland coalescing 3,400 *YBP*.

Problems in the estimation of ancestral ages

Estimates of dates pertaining to ancestral population expansions are subject to considerable error margins because of the use of alternative methods of dating and uncertainty in mutation rate estimates and generation time estimates. For example, the variance method of dating estimates a coalescence for hg 1 chromosomes almost twice as young as the ASD date (2,001 *YBP*), and likewise, the B&C method estimates a younger date (3,338 *YBP*). Similarly, the choice of mutation rate may alter the time to coalescence. Most dramatically, a much slower mutation rate of 0.012%, as suggested by Cooper *et al. (1996),* estimates the TMRCA for hg 1 chromosomes as ~66,500 *YBP*. Furthermore, a generation time of 27 years *(Weiss 1973)* increases the date by 1.08 times that estimated using a standard generation time of 25 years. The estimate for coalescence of Irish hg 1 chromosomes using a generation time of 27 years becomes ~4,200 *YBP*. It is not

inconceivable, therefore, that Irish hg 1 chromosomes are much younger or indeed much older than 3,789 *YBP* and, by relying solely on dating methods, the TMRCA may exceed 10,000 *YBP* (upper 95% CI 13,700 *YBP*).

These dates can only be taken as expansion times given an initial small population with little initial variation. Corroborative evidence from archeology is required to substantiate population expansion times as estimated from genetic data. For example, it is extremely unlikely that the population expansion in Ireland dates to over 10,000 *YBP* as this date exceeds that, determined by archeology, for the first evidence for human settlement on the island. The first evidence for seasonal occupation of Ireland is found in the northern regions of the country and indicates a date of ~9,000 *YBP (Woodman 1985)* However, evidence from a number of archaeological sites suggests that the first large scale settlements of hunter-gatherer peoples in Ireland was perhaps not until later in the Mesolithic period ~7,500 *YBP*. One hypothesis might suggest that the population growth occurred prior to the settlement of the island, but this requires a large scale intrusion to the island by very closely related peoples. This seems unlikely and is not suggested by archeological evidence. It is more likely that the rapid population growth did not occur until after the establishment of a permanent, settled population on the island some time after ~7,500 *YBP*.

Late Neolithic / Early Bronze Age Irish origin

The best estimate for a date of expansion of hg 1 chromosomes in Ireland is that calculated using the ASD dating method, the most conservative mutation rate (0.21%) and a generation time of 27 years. The expansion of hg 1 chromosomes in Ireland is estimated as 4,217 *YBP* (95% CI 1,807-14,760 *YBP*). This date suggests a Late Neolithic / Early Bronze Age origin of hg 1 chromosomes in Ireland.

Were this a valid date, there should be archeological evidence for a prosperous expanding population in Ireland at this time. It has been discussed that the introduction of agricultural technologies may be central in the increase of population capacity on any given land. Indications of agricultural communities in Ireland are unambiguously evident from shortly after 6,000 *YBP (Waddel 1998)* and it is suggested, therefore, that the introduction of farming technologies to the island facilitated a rapid population expansion early in the Bronze Age in Ireland dating to 4,200 *YBP*.

The results presented here are therefore consistent with an origin of the bulk of Irish Y chromosome variation in an agriculturally facilitated expansion of variants which themselves were of the pre-Neolithic type. This tentatively suggests that agriculture may have been adopted by the already substantial insular Mesolithic population, rather than resulting from significant population replacement at the fringe of Europe. The cline of hg 21 chromosomes across Europe suggests the movement of Neolithic agriculturalists (see **section 3.4**). The low frequency of hg 21 chromosomes in Ireland gives some suggestion of a cultural diffusion of agriculture rather than a demic diffusion and it may be that there was no substantial replacement of the indigenous Mesolithics.

Post-Neolithic influences

There is little evidence to suggest a large scale population replacement of indigenous Irish peoples at any time in the demographic history of the island. However, the presence of non-hg 1 Y chromosomes in eastern Gaelic samples yields the first genetic evidence for a substantial secondary prehistoric influence in the Irish genepool.

Most evident is the presence of hg 2 chromosomes in the Leinster (27%) and Ulster samples (18%). Haplogroup 2 chromosomes may have been introduced in a number of intrusions along with other non-hg 1 lineages. The persistence of hg 2 chromosomes in eastern population cohorts and the near absence in the west of the country has been discussed in **sections 3.3 and 3.4.**

Genetic isolation of peripheral European populations

The low frequency of non-hg 1 chromosomes in western Gaelic cohorts suggests little introgression of exogenous genes into the western regions of Ireland from the eastern regions and Britain and Europe. The extreme frequencies of hg 1 chromosomes in Connaught and Munster suggest genetic isolation. Fst genetic distances were used to assess the extent of gene flow between the population cohorts. The relative lack of gene flow between western and eastern cohorts and the comparable, and perhaps even greater, gene flow between eastern populations and England and Scotland is striking. For example, between Connaught and Munster and between Connaught and Ulster the Fst distances were zero and 0.125 respectively. By comparison the Fst distance between

Connaught and Leinster was much greater, 0.215. This is a greater value to the genetic distance between Leinster and England and Leinster and Scotland (0.004 and 0.010 respectively). This might suggest greater levels of admixture between Leinster and Connaught and Leinster and Britain, due to the proximity of Leinster to Britain and Europe. The greatest genetic distance between populations of the British Isles was that between the Connaught population and England (0.330). The Fst genetic distances suggest little continental admixture between geographically extreme, isolated populations.

The comparable four-locus modal European haplotype as determined by Jobling *et al. (1997)* is 190 bp /251 bp /254 bp /124 bp (DYS19/389-1/392/393). This is identical to haplotypes ht 62, ht 63, ht 64, ht 65, ht 112, ht 114, ht 115 and ht 116 *(this study).* It is worth noting that ht 128 is not among these. Haplotype 128 is the most common haplotype in the Connaught population (21.1%) but is uncommon in Leinster (3.3%), Munster (2.7%) and England (2.5%) and is absent in the Danish, Norman and Norse populations. Intermediate frequencies of ht 128 are detected in the Scottish (17.6%) and Ulster (18.2%) populations. The presence of intermediate frequencies of ht 128 in Ulster and Scotland suggests an ancient admixture between these two populations most likely due to their geographic proximity. The low frequency of ht 128 in the other populations and the high frequency in Connaught suggests the genetic isolation of Connaught and the retention of an ancient Y chromosome type. The high frequency of ht 128 in the Connaught population, which is almost fixed for the putative ancestral hg 1 chromosome type, suggests that ht 128 is the most ancestral Y chromosome type in Ireland. This haplotype is not detected in the Basque population *(Pérez-Lezaun et al 1997).*

The high frequency of hg 1 chromosomes in Ireland, the result of its geographic isolation on the periphery of Europe, is mirrored in only one other European population, the putative Mesolithic Basque population.

The maximum parsimony network relating Basque Y chromosomes, on visual inspection, is more diverse than the Irish networks. The ancestral haplotype, *ht 1*, lies central on the network with radiating branches leading to closely related haplotypes. Although the Irish and Basque phylogenies are not directly comparable some basic suggestions may be made about the relationship of the Basque and Irish populations. All of the estimated coalescence dates for the Basque population are almost exactly twice as old as the estimated coalescence times for the Irish population and may even be older. MtDNA D-loop sequence analysis dates the origin of the Basque population to some time

in the pre-Neolithic or even the Upper Paleolithic *(Bertranpetit et al 1995, Comas et al 1997).* Pérez-Lezaun *et al. (1997)* also estimate a much older date for the origin of the Basques by assuming a much slower mutation rate *(Cooper et al 1996)* and the B&C method of dating. The estimates using these parameters calculate a TMRCA for the Basque population between 7,100 – 61,020 *YBP*. In this study, in a comparable four-locus analysis using a comparable methodology (i.e. using the relationship ASD = 2μt) the TMRCA for Basque Y chromosomes is estimated as 6,256 *YBP* whereas the TMRCA for the Irish population is 3,101 *YBP*. Using all of the available data the TMRCA for the Basque population is estimated as 5,298 *YBP* and the TMRCA for the Irish population is estimated as 2,536 *YBP*. This suggests that the origin of Basque Y chromosomes is much older than the origin of Irish Y chromosomes.

However, in an analysis of pairwise differences the distributions suggest otherwise (see **figure 4.26**). For example, the bulk of Basque Y chromosome haplotypes share one difference and the bulk of Irish Y chromosomes share two differences. Additionally, the mean of the Irish distribution is 2.55 whereas the mean of the Basque distribution is 2.32. By comparison, the mean of the Turkish distribution is 3.34, the mean of the Catalan distribution is 2.68 and the mean of the Togolese distribution is 2.06. A low diversity in the Basque population has been detected in the internal modular structures of the MSY1 locus. Of all populations examined (Basque, British, Chinese, Cook Islander, Indian, Indonesian and Surui) the Basques had the least diversity (0.111). The Oceanic population were considerably more diverse (0.853) and the British population had intermediate levels of diversity (0.513) *(Jobling et al. 1998).* Additionally, in an analysis of mtDNA sequence divergence, intrapopulation diversity was found to be lowest in the Basques compared to other European populations *(Richards et al 1996).* It may be that geographic isolation and resistance to admixture from the continent has limited the diversity within this population.

Despite this anomaly, phylogenetic data and estimates of coalescence times suggest the Basque population as the more ancient. The sharing of common haplotypes in both populations suggests an ancient relationship which might be related to the migration of peoples northwards following the end of the Ice Age in Europe ~18,000 YBP (see **section 3.4** for further discussion). For example, the most common Basque haplotype is *ht 1 (Pérez-Lezaun 1997)* which, in a comparable four-locus haplotype, is identical to some of the more common Irish haplotypes (ht 108, ht 111, ht 115, ht 118 and ht 120). According to Kayser *et al. (1998)* the most common, completely comparable, six-locus Basque

haplotype is ht 118. Haplotype 118 is found in all of the Irish populations as well as the Scottish and Norse sample, but is not found in the Turkish, Danish, English or Norman populations.

It is likely that the geographic isolation of the Basque and Irish populations, and the paucity of evident admixture from the continent, has preserved a signature of ancient European Mesolithic Y chromosomes.

Literature cited

Literature cited

Abbate, E. A., A. Albianelli, A. Azzaroli, M. Benvenuti, B. Tesfamariam, P. Bruni, N. Cipriani, R.J. Clarke, G. Ficcarelli, R. Macchiarelli, G. Napoleone, M. Papini, L. Rook, M. Sagri, T.M. Tecle, D. Torre and I. Villa (1998) "A one-million-year-old Homo cranium from the Danakil (Afar) Depression of Ethiopia". Nature **393**: 458-60

Aldridge, J., L. Kunkel, G. Bruns, U. Tantravahi, M. Lalande, T. Brewster, E. Moreau, M. Wilson, W. Bromley, T. Roderick, *et al.* (1984). "A strategy to reveal high-frequency RFLPs along the human X chromosome." Am J Hum Genet **36**(3): 546-64.

al-Maghtheh, M., V. Ray, S.S. Mastana, M.D. Garralda, S.S. Bhattacharya, S.S. Papiha (1993). "Variation in DNA polymorphisms of the short arm of the human X chromosome: genetic affinity of Parsi from western India." Hum Hered **43**(4): 239-43.

Alonso, S., and J. A. Armour (1998) "MS205 minisatellite diversity in Basques: evidence for a pre-Neolithic component." Genome Res. **8**: 1289:98

Alonso, S., I. Fernandez-Fernandez, A. Castro and M.M. De Pancorbo (1998). "Genetic characterization of APOB and D17S5 AFLP loci in a sample from the Basque Country (northern Spain)." Hum Biol **70**(3): 491-505.

Ammerman, A.J. and L. .L. Cavalli-Sforza (1973) "A population model for the diffusion of early farming in Europe" In The explanation of cultural change. C. Renfrew, ed. London, Duckworth

Anagnostopoulos, T., P. M. Green, G. Rowley, C.M. Lewis and F. Giannelli (1999). "DNA Variation in a 5-Mb Region of the X Chromosome and Estimates of Sex-Specific/Type-Specific Mutation Rates." Am J Hum Genet **64**(2): 508-517.

Anton, S. C. and K.J. Weinstein (1999) "Artificial cranial deformation and fossil Australians revisited" J. Hum. Evol. **36**: 195-209

Aris-Brosou, S. and L. Excoffier (1996). "The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism." Mol Biol Evol **13**(3): 494-504.

Armour, J.A., T. Anttinen, C.A. May, E.E. Vega, A. Sajantila, J.R. Kidd, K.K. Kidd, J. Bertranpetit, S. Paabo and A.J. Jeffreys (1996). "Minisatellite diversity supports a recent African origin for modern humans." Nat Genet **13**(2): 154-60.

Asfaw, B., T. White, O. Lovejoy, B. Latimer, S. Simpson and G. Suwa (1999) "Australopithecus garhi: a new species of early hominid from Ethiopia" Science **284**: 629-35

Avise, J. C. (1994) Molecular markers, Natural History and Evolution. New York, Chapman and Hall.

Bandelt, H. J., and P. Forster (1997) "The myth of bumpy hunter-gatherer mismatch distributions." Am J Hum Genet;**61**(4):980-3

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

Bar-Yosef, O. and A. Belfer-Cohen (1989). "The Origins of Sedentism and Farming Communities in the Levant." Journal of World Prehistory **3**(4): 447-98.

Batzer, M. A., M. Stoneking, M. Alegria-Hartman, H. Bazan, D.H. Kass, T.H. Shaikh, G.E. Novick, P.A. Ioannou, W.D. Scheer, R.J. Herrera, *et al.* (1994). "African origin of human-specific polymorphic Alu insertions." Proc Natl Acad Sci U S A **91**(25): 12288-92.

Bell, G.I., and J. Jurka (1997) "The length distribution of perfect dimer repetitive DNA is consistent with its evolution by an unbiased single-step mutation process." J Mol Evol **44**:414-21

Bernstein HS, D.F. Bishop, K.H. Astrin, R. Kornreich, C.M. Eng, H. Sakuraba and R.J. Desnick (1989) "Fabry disease: six gene rearrangements and an exonic point mutation in the alpha-galactosidase gene." J Clin Invest **83**(4):1390-9

Bertorelle, G., and M. Slatkin (1995) "The number of segregating sites in expanding human populations, with implications for estimates of demographic parameters." Mol Biol Evol **12**(5):887-92

Bertranpetit, J. and F. Calafell (1996). Genetic and geographic variability in cystic fibrosis: evolutionary considerations. Variation in the human genome. D. Chadwick and G. Cardew. Chichester, John Wiley & Sons: 97-118.

Bertranpetit, J. and L. L. Cavalli-Sforza (1991). "A genetic reconstruction of the history of the population of the Iberian Peninsula." Ann Hum Genet **55**(Pt 1): 51-67.

Bertranpetit, J., J. Sala, F. Calafell, P.A. Underhill, P. Moral and D. Comas (1995). "Human mitochondrial DNA variation and the origin of Basques." Ann Hum Genet **59**(Pt 1): 63-81.

Bowcock, A. M., A. Ruiz-Linares, J. Tomfohrde, E. Minch, J.R. Kidd JR and L.L. Cavalli-Sforza (1994). "High resolution of human evolutionary trees with polymorphic microsatellites." Nature **368**(6470): 455-7.

Brauer, G. (1992) "Africa's place in the evolution of modern humans" In Continuity or replacement. Controversies in Homo sapiens evolution. G. Brauer and F. H. Smith, eds. Rotterdam, A.A. Balkener

Brinkmann, B., A. Moller, and P. Wiegand (1995) "Structure of new mutations in 2 STR systems." Int J Legal Med **107**(4):201-3

Brinkmann, B., A. Junge, E. Meyer E and P. Wiegand (1998). "Population genetic diversity in relation to microsatellite heterogeneity." Hum Mutat **11**(2): 135-44.

Brown, W. M. (1985) "The mitochondrial genome of animals". In Molecular and Evolutionary Genetics, R. McIntyre, ed. New York: Plenum Press

Brownstein, B.H., G.A. Silverman, R.D. Little, D.T. Burke, S.J. Korsmeyer, D. Schlessinger and M.V. Olson (1989) "Isolation of single-copy human genes from a library of yeast artificial chromosome clones." Science **244**(4910):1348-51

Buchanan, F.C., L.J. Adams, R.P. Littlejohn, J.F. Maddox JF and A.M. Crawford (1994) " Determination of evolutionary relationships among sheep breeds using microsatellites" Genomics **22**: 397-403

Burke, D.T., G.F.Carle, and M.V. Olson (1987) "Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors." Science **236**(4803):806-12

Calderon, R., C. Vidales, J.A. Pena, A. Perez-Miranda and J.M. Dugoujon (1998). "Immunoglobulin allotypes (GM and KM) in Basques from Spain: approach to the origin of the Basque population." Hum Biol **70**(4): 667-98.

Cann, R., M. Stoneking and A.C. Wilson (1987). "Mitochondrial DNA and human evolution." Nature **325**: 31-36.

Casanova, M., P. Leroy, C. Boucekkine, J. Weissenbach, C. Bishop, M. Fellous, M. Purrello, G. Fiori and M. Siniscalco (1985). "A human Y-linked DNA polymorphism and its potential for estimating genetic and evolutionary distance." Science **230**(4732): 1403-6.

Castelloe, J. and A.R. Templeton(1994). "Root probabilities for intraspecific gene trees under neutral coalescent theory." Mol. Phylogenet. Evol. **3**: 102-113.

Cavalli-Sforza, L. (1986). "Population structure" In Evolutionary perspectives and the New Genetics. H. Gershowitz, D. L. Rucknagel and R.E. Tashian, eds. New York, Alan R.Liss

Cavalli-Sforza, L. and A. W. F. Edwards (1967) "Phylogenetic analysis: models and estimation procedures" Am. J. Hum. Genet **19**: 223-57

Cavalli-Sforza, L., P. Menozzi and A. Piazza. (1994). The history and geography of human genes. Princeton, Princeton University Press.

Chakraborty, R., M. Kimmel, D.N. Stivers, L.J. Davison and R. Deka (1997). "Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci." Proceedings of the National Academy of Sciences, USA **94**: 1041-1046.

Chikhi, L., G. Destro-Bisol, V. Pascali, V. Baravelli, M. Dobosz and G. Barbujani (1998). "Clinal variation in the nuclear DNA of Europeans." Hum Biol **70**(4): 643-57.

Childe, V. G. (1928). The Most Ancient East: the Oriental Prelude to European Prehistory. London, Kegan, aul, Trench and Trubner.

Ciminelli, B.M., F. Pompei, P. Malaspina, M. Hammer, F. Persichetti, P.F. Pignatti, A. Palena, N. Anagnou, G. Guanti, C. Jodice *et al.* (1995) "Recurrent simple tandem repeat mutations during human Y-chromosome radiation in Caucasian subpopulations." J Mol Evol **41**(6):966-73

Cohen, M. N. (1977) The food crisis in prehistory: Over population and origins of agriculture. New Haven, Conn., Yale University Press.

Comas, D., F. Calafell, E. Mateu, A. Pérez-Lezaun and J. Bertranpetit (1996) "Geographic variation in human mitochondrial DNA control region sequence: the population history of Turkey and its relationship to the European populations." Mol Biol Evol **13**(8):1067-77

Comas, D., F. Calafell, E. Mateu, A. Pérez-Lezaun, E. Bosch and J. Bertranpetit (1997). "Mitochondrial DNA variation and the origin of the Europeans." Hum Genet **99**(4): 443-9.

Comas, D., E. Mateu, F. Calafell, A. Pérez-Lezaun, E. Bosch, R. Martinez-Arias and J. Bertranpetit (1998). "HLA class I and class II DNA typing and the origin of Basques." Tissue Antigens **51**(1): 30-40.

Cooper, G., W. Amos, D. Hoffman and D.C. Rubinsztein (1996). "Network analysis of human Y microsatellite haplotypes." Hum Mol Genet **5**(11): 1759-66.

Crouau-Roy, B. and J. Clayton (1995). "Haplotypes without children: PCR applied to close loci on individual human sperm." Hum Biol **67**(1): 171-8.

Cruciani, F., D. Sellitto, P. Santolamazza, T. Vespertilli, M. Lerone, G. Spedini and R. Scozzari (1994). "An MspI polymorphism in the X-specific region proximal to the pseudoautosomal boundary. A new example of a unique "African" marker?" Hum Genet **94**(2): 215-6.

Dansgaard, W. (1993). "Evidence for General Instability of Past Climate from a 250-kyr Ice-Core Record." Nature **15**(218-220).

Davies K.E., N.H. Thomas, R.J. Daniels and V. Dubowitz (1991) "Molecular studies of spinal muscular atrophy" Neuromuscul Disord **1**(2):83-5

de Knijff P., M. Kayser, A. Caglia, D. Corach, N. Fretwell, C. Gehrig, G. Graziosi, F. Heidorn, S. Herrmann, B. Herzog, M. Hidding, K. Honda, M. Jobling, M. Krawczak, K. Leim, S. Meuser, E. Meyer, W. Oesterreich, A. Pandya, W. Parson, G. Penacino, A. Perez-Lezaun, A. Piccinini, M. Prinz, L. Roewer *et al.* (1997). "Chromosome Y microsatellites: population genetic and evolutionary aspects." Int J Legal Med **110**(3): 134-49.

Deka, R., L. Jin, M.D. Shriver, L.M. Yu, S. DeCroo, J. Hundrieser, C.H. Bunker, R.E. Ferrell and R. Chakraborty (1995). "Population genetics of dinucleotide (dC-dA)n/(dG-dT)n polymorphisms in world populations." American Journal of Human Genetics **56**: 461-474.

Deka, R., L. Jin, M.D. Shriver, L.M. Yu, N. Saha, R. Barrantes, R. Chakraborty and R.E. Ferrell (1996). "Dispersion of human Y chromosome haplotypes based on five microsatellites in global populations." Genome Res **6**(12): 1177-84.

Deka, R., M.D. Shriver, L.M. Yu, R.E. Ferrell and R. Chakraborty (1995). "Intra- and inter-population diversity at short tandem repeat loci in diverse populations of the world." Electrophoresis **16**: 1659-1664.

Di Rienzo, A., A.C. Peterson, J.C. Garza, A.M. Valdes, M. Slatkin and N.B. Freimer (1994). "Mutational processes of simple-sequence repeat loci in human populations." Proceedings of the National Academy of Sciences, USA **91**: 3166-3170.

Dorit, R. L., H. Akashi and W. Gilbert (1995). "Absence of polymorphism at the ZFY locus on the human Y chromosome." Science **268**(5214): 1183-5.

Drayna, D., K. Davies, D. Hartley, J.L. Mandel, G. Camerino, R. Williamson and R. White (1984). "Genetic mapping of the human X chromosome by using restriction fragment length polymorphisms." Proc Natl Acad Sci U S A **81**(9): 2836-9.

Ellis, N., Taylor A, Bengtsson BO, Kidd J, Rogers J and P. Goodfellow (1990). "Population structure of the human pseudoautosomal boundary." Nature **344**(6267): 663-5.

Excoffier, L. and M. Slatkin (1995). "Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population." Mol Biol Evol **12**(5): 921-7.

Excoffier, L. and M. Slatkin (1998). "Incorporating genotypes of relatives into a test of linkage disequilibrium." Am J Hum Genet **62**(1): 171-80.

Excoffier, L., P. E. Smouse and J.M. Quattro (1992). "Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data." Genetics **131**(2): 479-91.

Fagan, B. M. (1980). People of the Earth: an introduction to World Prehistory. Canada, Little, Brown and Company (Inc.).

Felsenstein, J. (1989). "PHYLIP: phylogeny inference package (version 3.2)." Cladistics **5**: 164-166.

Foley, R. (1998) "The context of human genetic evolution" Genome Research. **8**: 339-347

Forster P, , Kayser M, Meyer E, Roewer L, Pfeiffer H, Benkmann H and B. Brinkmann (1998) "Phylogenetic resolution of complex mutational features at Y-STR DYS390 in aboriginal Australians and Papuans." Mol Biol Evol **15**(9):1108-14

Fullerton, M., R. M. Harding *et al.* (1997) "The genetic ancestry of modern humans: Inferences from the analysis of DNA sequence diversity at the beta-globin locus." Am. J. Hum. Biol. **9**: 118

Gabunia, L. and A. Veuka (1995). "A Plio-Pleistocene hominid from Dmanisi, East Georgia, Caucasus." Nature **373**: 509-512.

Gamble, C. (1986). The Paleolithic settlement of Europe. Cambridge, Cambridge University Press.

Gamkrelidze, T. V. and V. V. Ivanov (1990). "The early history of the Indo-European languages." Scientific American **262**(3): 110-116.

Garza, J.C., M. Slatkin and N.B. Freimer (1995) "Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size." Mol Biol Evol **12**:594-603

Gibbons, A. (1997). "Y chromosome shows that Adam was African." Science **278**(5339): 804-5.

Gill, M., G. Daly, S. Heron S, Z. Hawi and M. Fitzgerald (1997). "Confirmation of association between attention deficit hyperactivity disorder and a dopamine transporter polymorphism." Mol Psychiatry **2**(4): 311-3.

Gimbutas, M. (1970). Proto-Indo-European culture: The Kurgan culture during the fifth, fourth and third millennia B.C. Philadelphia, Philadelphia University Press.

Goldman, A., M. Ramsay and T. Jenkins (1995). "New founder haplotypes at the myotonic dystrophy locus in southern Africa." Am J Hum Genet **56**(6): 1373-8.

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

Goldstein, D. B., A. R. Linares, L.L. Cavalli-Sforza and M.W. Feldman (1995). "Genetic absolute dating based on microsatellites and the origin of modern humans." Proc Natl Acad Sci U S A **92**(15): 6723-7.

Goldstein, D. B, Zhivotovsky LA, Nayar K, Linares AR, Cavalli-Sforza LL and M.W. Feldman (1996). "Statistical properties of the variation at linked microsatellite loci: implications for the history of human Y chromosomes." Mol Biol Evol **13**(9): 1213-8.

Gort, L., A. Chabas and M.J. Coll (1998). "Hunter disease in the Spanish population: molecular analysis in 31 families." J Inherit Metab Dis **21**(6): 655-61.

Greenland Ice-core Project (GRIP) members (1993) "Climate instability during the last interglacial period recorded in the GRIP ice-core." Nature **364**: 203-207

Grewal, R. P., R. Cantor, G. Turner, R.K. Grewal and S.D. Detera-Wadleigh (1998). "Genetic mapping and haplotype analysis of oculopharyngeal muscular dystrophy." Neuroreport **9**(6): 961-5.

Guglielmino, C. R., A. De Silvestri, A. Rossi, V. De Micheli (1998) "Surnames, HLA genes and ancient migrations in the Po Valley (Italy)." Ann Hum Genet **62** (Pt 3):261-9

Guldberg, P., Zschocke J, Dagbjartsson A, Henriksen KF and F. Guttler (1997). "A molecular survey of phenylketonuria in Iceland: identification of a founding mutation and evidence of predominant Norse settlement." Eur J Hum Genet **5**(6): 376-81.

Guo, S.W. and E.A. Thompson (1992) "Performing the exact test of Hardy-Weinberg proportions for multiple alleles" Biometrics. **48**: 361-72

Hackett, W. E. R. and G. W. P. Dawson (1958). "The distribution of the ABO and simple Rhesus (D) blood groups in the Republic of Ireland from a sample of 1 in 37 of the adult population." Irish Journal of Medical Science **387**: 99-109.

Haldane, J. B. S. (1935) Biology in everyday life. London. Allen & Unwin,

Hamada, H., M.G. Petrino and T. Takunaga (1982) "A novel repeated element with Z-DNA forming potential is widely found in evolutionarily diverse eukaryotic genomes. Proc. Natl. Acad. Sci, USA **79**: 6465-6469

Hamada, H. M. Seidman, B.H. Howard and C.M. Gorman (1984) "Enhanced gene expression by the poly(dT-dG)n•(dC-dA) sequence." Mol. Cell. Biol. **4**: 2622-2630

Hammer, M. F. (1994). "A recent insertion of an alu element on the Y chromosome is a useful marker for human population studies." Mol Biol Evol **11**(5): 749-61.

Hammer, M. F. (1995). "A recent common ancestry for human Y chromosomes." Nature **378**(6555): 376-8.

Hammer, M.F., S. Horai (1995) "Y chromosomal DNA variation and the peopling of Japan." Am J Hum Genet **56**(4):951-62

Hammer, M. F., Spurdle AB, Karafet T, Bonner MR, Wood ET, Novelletto A, Malaspina P, Mitchell RJ, Horai S, Jenkins T and S.L. Zegura (1997). "The geographic distribution of human Y chromosome variation." Genetics **145**(3): 787-805.

Hammer, M. F., Karafet T, Rasanayagam A, Wood ET, Altheide TK, Jenkins T, Griffiths RC, Templeton AR and S.L. Zegura (1998). "Out of Africa and back again: nested cladistic analysis of human Y chromosome variation." Mol Biol Evol **15**(4): 427-41.

Harbison, P. (1988). Pre-Christian Ireland: From the first settlers to the Early Celts. London, Thames and Hudson.

Harding, R. M., Fullerton SM, Griffiths RC, Bond J, Cox MJ, Schneider JA, Moulin DS and J.B. Clegg (1997). "Archaic African and Asian lineages in the genetic ancestry of modern humans." Am J Hum Genet **60**(4): 772-89.

Harley, H. G., J. D. Brook, et al. (1991). "Detection of linkage disequilibrium between the myotonic dystrophy locus and a new polymorphic DNA marker." Am J Hum Genet **49**(1): 68-75.

Harpending, H., S. Sherry, *et al.* (1993). "The genetic structure of ancient human populations." Curr. Anthropol. **34**: 483-496.

Harpending, H. C., Batzer MA, Gurven M, Jorde LB, Rogers AR and S.T. Sherry (1998). "Genetic traces of ancient demography." Proc Natl Acad Sci U S A **95**(4): 1961-7.

Harris, H. (1966) "Enzyme polymorphisms in man" Proc R Soc Lon Ser B. **164**: 298-310

Harris EE, J. Hey (1999) "X chromosome evidence for ancient human histories." Proc Natl Acad Sci U S A **96**(6):3320-4

Hassan, F. A. (1975) "Determination of the size, density and growth rate of hunting-gathering populations." In Population, Ecology and Social Evolution. S. Poloyor, ed. The Hague, Mouton

Hawley, M. E. and K. K. Kidd (1995). "HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes." J Hered **86**(5): 409-11.

Heun, M. (1997). "Site of einkorn wheat domestication identified by DNA fingerprinting." Science **278**: 1312-1314.

Heyer, E., Puymirat J, Dieltjes P, Bakker E and P. de Knijff (1997). "Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees." Hum Mol Genet **6**(5): 799-803.

Hooton, E. A. (1940). "Stature, head form and pigmentation of adult male Irish." American Journal of Physical Anthropology **26**: 229-249.

Hooton, E. A., C. W. Dupertuis, *et al.* (1955). "The Physical Anthropology of Ireland." Papers of the Peabody Museum **30**(1-2).

Horai, S. (1995). "Evolution and the origins of man: clues from complete sequences of hominoid mitochondrial DNA." Southeast Asian J Trop Med Public Health **26 Suppl 1**: 146-54.

Huang, W., Ciochon R, Gu Y, Larick R, Qiren F, Schwarcz H, Yonge C, de Vos J and W. Rink (1995). "Early Homo and associated artifacts from Asia." Nature **378**: 275-278.

Hudson, R.R., A. G. Saez, and F. J. Ayala (1997) "DNA variation at the Sod locus of Drosophila melanogaster: an unfolding story of natural selection." Proc Natl Acad Sci U S A **94**(15):7725-9

Hurles, M.E., et al. (1998) "European Y-chromosomal lineages in Polynesians: a contrast to the population structure revealed by mtDNA." Am. J. Hum. Genet. **63**, 1793-1806

Jakubiczka, S., Arnemann J, Cooke HJ, Krawczak M and J. Schmidtke (1989) "A search for restriction fragment length polymorphism on the human Y chromosome". Hum Genet **84**(1):86-8

Jeffreys, A.J., V. Wilson and S.L. Thein (1985) "Hypervariable "minisatellite" regions in human DNA" Nature **314**: 67-73

Jin, L., Macaubas C, Hallmayer J, Kimura A and E. Mignot (1996). "Mutation rate varies among alleles at a microsatellite locus: phylogenetic evidence." Proc Natl Acad Sci U S A **93**(26): 15285-8.

Joachim, M. A. (1983). Paleolithic cave art in ecological perspective. Cambridge, Cambridge University Press.

Jobling, M. A. (1994) "A survey of long-range DNA polymorphisms on the human Y chromosome." Hum Mol Genet 3(1):107-14

Jobling, M. A. and C. Tyler-Smith (1995). "Fathers and sons: the Y chromosome and human evolution." Trends Genet **11**(11): 449-56.

Jobling, M. A., Samara V, Pandya A, Fretwell N, Bernasconi B, Mitchell RJ, Gerelsaikhan T, Dashnyam B, Sajantila A, Salo PJ, Nakahori Y, Disteche CM, Thangaraj K, Singh L, Crawford MH and C. Tyler-Smith (1996). "Recurrent

duplication and deletion polymorphisms on the long arm of the Y chromosome in normal males." Hum Mol Genet 5(11): 1767-75.

Jobling, M.A., A. Pandya, and C. Tyler-Smith (1997) "The Y chromosome in forensic analysis and paternity testing." Int J Legal Med **110**(3):118-24

Jobling, M. A., Williams GA, Schiebel GA, Pandya GA, McElreavey GA, Salas GA, Rappold GA, Affara NA and C. Tyler-Smith (1998). "A selective difference between human Y-chromosomal DNA haplotypes." Curr Biol **8**(25): 1391-4.

Jobling, M. A., N. Bouzekri and P.G. Taylor (1998). "Hypervariable digital DNA codes for human paternal lineages: MVR-PCR at the Y-specific minisatellite, MSY1 (DYF155S1). Hum Mol Genet. **7**(4):643-53.

Johanson, D. C. (1989). The current status of Australopithecus. Milan, Jaca Books.

Jones, C. (1997). Final Neolithic/Early Bronze Age occupation of the Burren: a brief review of the evidence., The Shannon Archaeological and Historical Society.

Jorde, L. B., Bamshad MJ, Watkins WS, Zenger R, Fraley AE, Krakowiak PA, Carpenter KD, Soodyall H, Jenkins T and A.R. Rogers (1995). "Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data." American Journal of Human Genetics **57**: 523-538.

Jorde, L. B., Rogers AR, Bamshad M, Watkins WS, Krakowiak P, Sung S, Kere J and H.C. Harpending (1997). "Microsatellite diversity and the demographic history of modern humans." Proc Natl Acad Sci U S A **94**(7): 3100-3.

Kaessmann H, Heissig F, von Haeseler A and S. Paabo (1999) "DNA sequence variation in a non-coding region of low recombination on the human X chromosome." Nat Genet **22**(1):78-81

Karafet, T., Zegura SL, Vuturo-Brady J, Posukh O, Osipova L, Wiebe V, Romero F, Long JC, Harihara S, Jin F, Dashnyam B, Gerelsaikhan T, Omoto K and M.F. Hammer (1997). "Y chromosome markers and Trans-Bering Strait dispersals." Am J Phys Anthropol **102**(3): 301-14.

Karafet, T. M., Zegura SL, Posukh O, Osipova L, Bergen A, Long J, Goldman D, Klitz W, Harihara S, de Knijff P, Wiebe V, Griffiths RC, Templeton AR and M.F. Hammer (1999) "Ancestral Asian source(s) of new world Y-chromosome founder haplotypes." Am J Hum Genet **64**(3):817-31

Kayser, M., de Knijff P, Dieltjes P, Krawczak M, Nagy M, Zerjal T, Pandya A, Tyler-Smith C and L. Roewer (1997). "Applications of microsatellite-based Y chromosome haplotyping." Electrophoresis **18**(9): 1602-7.

Kimmel, M., Chakraborty R, King JP, Bamshad M, Watkins WS and L.B. Jorde (1998). "Signatures of population expansion in microsatellite repeat data." Genetics **148**(4): 1921-30.

Klein, R. G. (1989). Biological and behavioural perspectives on modern human origins in southern Africa. Princeton, Princeton University Press.

Knight, A., Batzer MA, Stoneking M, Tiwari HK, Scheer WD, Herrera RJ and D.L. Deininger (1996). "DNA sequences of Alu elements indicate a recent replacement of the human autosomal genetic complement." Proc Natl Acad Sci U S A **93**(9): 4360-4.

Krings, M., Stone A, Schmitz RW, Krainitzki H, Stoneking M and S. Paabo (1997). "Neandertal DNA sequences and the origin of modern humans." Cell **90**(1): 19-30.

Kwok, C. (1996). "Mutational analysis of the 2kb 5' to SRY in XY females and XY intersex subjects." J. Med. Genet. **33**(6): 465-468.

Laan, M. and S. Paabo (1997). "Demographic history and linkage disequilibrium in human populations." Nat Genet **17**(4): 435-8.

Lahr, M. M. (1996) The evolution of human diversity. Cambride, Cambridge Universoty Press.

Lahr, M. M. and R. A. Foley (1998). "Towards a theory of modern human origins: geography, demography, and diversity in recent human evolution." Am J Phys Anthropol **Suppl 27**(2): 137-76.

Larin Z, A.P. Monaco and H Lehrach (1991) "Yeast artificial chromosome libraries containing large inserts from mouse and human DNA. Proc Natl Acad Sci U S A **88**(10):4123-7

Leakey, R. E. and A. C. Walker (1976). "Australopithecus, Homo erectus and the single species hypothesis." Nature **261**: 572.

Leakey, M. G. and R. E. Leakey (1978). The Koobi Fora research project: The fossil hominids and an introduction to their context, 1968-1974. Oxford, Oxford University Press.

Levinson, G. and G.A. Gutman (1987) "Slipped strand mispairing: a major mechanism for DNA sequence evolution" Mol. Biol. Evol. **4**: 203-221

Lewontin, R. C. (1972). The apportionment of human diversity. New York, Appleton-Century-Crofts.

Lewontin, R. C. and J. L. Hubby (1966) "A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura." Genetics **54**: 595-609

Litt, M. and J.A. Lutty (1989) "A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene" Am J Hum Genet. **44**: 397-401

Lucotte, G., and N. Y. Ngo (1985) "p49f, A highly polymorphic probe, that detects Taq1 RFLPs on the human Y chromosome" Nucleic Acids Res **13**(22):8285.

Lucotte, G. and S. Hazout (1996). "Y-chromosome DNA haplotypes in Basques." J Mol Evol **42**(4): 472-5.

Lucotte, G., Sriniva KR, Loirat F, Hazout S and J. Ruffie (1990) "The p49/TaqI Y-specific polymorphisms in three groups of Indians." Gene Geogr **4**(1):21-7

Macaulay V., Richards M, Hickey E, Vega E, Cruciani F, Guida V, Scozzari R, Bonne-Tamir B, Sykes B and A. Torroni (1999) "The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs." Am J Hum Genet **64**:232-49

MacHugh, D. E., Loftus RT, Bradley DG, Sharp PM and P. Cunningham (1994). "Microsatellite DNA variation within and among European cattle breeds." Proc. R. Soc. Lond. Ser. B **256**: 25-31.

MacHugh, D. E., Shriver MD, Loftus RT, Cunningham P and D.G. Bradley (1997). "Microsatellite DNA variation and the evolution, domestication and phylogeography of taurine and zebu cattle (*Bos taurus* and *Bos indicus*)." Genetics **146**(3):1071-86

MacLysaght, E. (1997) The surnames of Ireland. Irish Academic Press, Dublin.

Maddison W.P. and D.R. Maddison (1992). MacClade version 3: Analysis of phylogeny and character evolution. Sinauer Associates, Sunderland Massachusetts.

Maddison, D. R. (1991). "African origin of human mitochondrial DNA reexamined." Syst. Zool. **40**: 355-363.

Mahtani, M. M and H. F. Willard (1993) "A polymorphic X-linked tetranucleotide repeat locus displaying a high rate of new mutation: implications for mechanisms of mutation at short tandem repeat loci." Hum Mol Genet **2**:431-7

Malaspina, P., Persichetti F, Novelletto A, Iodice C, Terrenato L, Wolfe J, Ferraro M and G. Prantera (1990) "The human Y chromosome shows a low level of DNA polymorphism." Ann Hum Genet **54** (4):297-305

Malaspina, P., Cruciani F, Ciminelli BM, Terrenato L, Santolamazza P, Alonso A, Banyko J, Brdicka R, Garcia O, Gaudiano C, Guanti G, Kidd KK, Lavinha J, Avila M, Mandich P, Moral P, Qamar R, Mehdi SQ, Ragusa A, Stefanescu G, Caraghin M, Tyler-Smith C, Scozzari R and A Novelletto (1998). "Network Analyses of Y-

Chromosomal Types in Europe, Northern Africa, and Western Asia Reveal Specific Patterns of Geographic Distribution." Am J Hum Genet **63**(3): 847-60.

Mallory, J.P. (1991) In search of the Indo-European : Language, Archaeology, and Myths. London, Thames and Hudson

Mallory, J.P. & McNeill, T.E. (1991) The archaeology of Ulster from colonization to plantation. The Institute of Irish Studies, The Queen's University of Belfast, Belfast.

Manly, F. (1986). Multivariate statistical methods: a primer. London, Chapman and Hall.

Mathias, N., M. Bayés, and C. Tyler-Smith (1994) "Highly informative compound haplotypes for the human Y chromosome." Hum. Mol. Genet. **3**, 115-123

McVean, G. T. and L. D. Hurst (1997). "Evidence for a selectively favourable reduction in the mutation rate of the X chromosome." Nature **386**(6623): 388-92.

Mellaart, J. (1967). Çatal Hüyük - a Neolithic Town. New York, McGraw Hill.

Menozzi, P., A. Piazza and L.L. Cavalli-Sforza (1978) "Synthetic maps of human gene frequencies in Europe" Science 201: 786-92

Mitchell, R. J., L. Earl and B. Fricke (1997). "Y-chromosome specific alleles and haplotypes in European and Asian populations: linkage disequilibrium and geographic diversity." Am J Phys Anthropol **104**(2): 167-76.

Morin, P. A., Moore JJ, Chakraborty R, Jin L, Goodall J and D.S. Woodruff (1994). "Kin selection, social structure , gene flow and the evolution of chimpanzees." Science **265**: 1193-1201

Morral, N., Bertranpetit J, Estivill X, Nunes V, Casals T, Gimenez J, Reis A, Varon-Mateeva R, Macek M Jr, Kalaydjieva L, *et al.* (1994). "The origin of the major cystic firbrosis mutation (Δ-F508) in European populations." Nature Genetics **7**: 169-175.

Morral, N., Nunes V, Casals T, Chillon M, Gimenez J, Bertranpetit J and X. Estivill (1993). "Microsatellite haplotypes for cystic fibrosis: mutation frameworks and evolutionary tracers." Hum Mol Genet **2**(7): 1015-22.

Mountain, J. L. and L. L. Cavalli-Sforza (1994). "Inference of human evolution through cladistic analysis of nuclear DNA restriction polymorphisms." Proceedings of the National Academy of Sciences, USA **91**: 6515-6519.

Naumova, A. K., Olien L, Bird LM, Slamka C, Fonseca M, Verner AE, Wang M, Leppert M, Morgan K and C. Sapienza (1995). "Transmission-ratio distortion of X chromosomes among male offspring of females with skewed X-inactivation." Dev Genet **17**(3): 198-205.

Naumova, A. K., Olien L, Bird LM, Smith M, Verner AE, Leppert M, Morgan K and C. Sapienza (1998). "Genetic mapping of X-linked loci involved in skewing of X chromosome inactivation in the human." Eur J Hum Genet **6**(6): 552-62.

Neel, J. V. (1973). ""Private" genetic variants and the frequency of mutation among South American Indians." Proc Natl Acad Sci U S A **70**(12): 3311-5.

Nei, M. (1972). "Genetic distance between populations." Am. Nat. **106**: 283-292.

Nei, M. (1987). "Molecular evolutionary genetics." Columbia University Press, New York.

Nei, M. (1995). "Genetic support for the out-of-Africa theory of human evolution." Proc. Natl. Acad. Sci. USA **92**: 6720-6722.

Nei, M. (1995). "Genetic support for the out-of-Africa theory of human evolution." Proc. Natl. Acad. Sci., USA **92**: 6720-6722.

Nei, M. and D. Graur (1984). "Extent of protein polymorphism and the neutral mutation theory." Evol. Biol. **17**: 73-118.

Nei, M. and A. Roychoudhury (1993). "Evolutionary relationships of human populations on a global scale." Molecular Biology and Evolution **10**: 927-943.

Nei, M. and A. K. Roychoudhury (1972). "Gene differences between Caucasian, Negro and Japanese populations." Science **177**: 434-435.

Ngo, K. Y., Vergnaud G, Johnsson C, Lucotte G and J. Weissenbach (1986). "A DNA probe detecting multiple haplotypes of the human Y chromosome." Am J Hum Genet **38**(4): 407-18.

Oakey, R. and C. Tyler-Smith (1990). "Y chromosome DNA haplotyping suggests that most European and Asian men are descended from one of two males." Genomics **7**(3): 325-30.

Ohta, T. and K. Kimura (1973). "A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a genetic population." Genet. Res. **22**: 201-204.

Orstavik, K. H., Orstavik RE, Naumova AK, D'Adamo P, Gedeon A, Bolhuis PA, Barth PG and D. Toniolo (1998). "X chromosome inactivation in carriers of Barth syndrome." Am J Hum Genet **63**(5): 1457-63.

Oudet, C., von Koskull H, Nordstrom AM, Peippo M and J.L. Mandel (1993). "Striking founder effect for the fragile X syndrome in Finland." Eur J Hum Genet **1**(3): 181-9.

Paabo, S. (1995) "The Y chromosome and the origin of all of us." Science **26**;268(5214):1141-2

Palsboll, P.J., M. Berube and H. Jorgensen (1999) "Multiple levels of single-strand slippage at cetacean tri- and tetranucleotide repeat microsatellite loci" Genetics **151**: 285-96

Papiha, S. S., S. S. Bhattacharya and D.F. Roberts (1988). "DNA polymorphisms, identified by an X-chromosome short-arm probe L 1.28 (DXS7), in different racial groups." Hum Hered **38**(2): 72-5.

Papiha, S. S., Mastana SS, Roberts DF, Onyemelukwe GC and S.S. Bhattacharya (1991). "Population variation in molecular polymorphisms of the short arm of the human X chromosome." Am J Phys Anthropol **85**(3): 329-34.

Pardue, M.L., Lowenhaupt K, Rich A and A. Nordheim (1987) "(dC-dA)n•(dG-dT)n sequences have evolutionary conserved chromosomal locations in Drosophila with implications for roles in chromosome structure and function. EMBO J. **6**: 1781-1789

Passarino, G., Semino O, Quintana-Murci L, Excoffier L, Hammer M and A.S. Santachiara-Benerecetti (1998). "Different genetic components in the Ethiopian population, identified by mtDNA and Y-chromosome polymorphisms." Am J Hum Genet **62**(2): 420-34.

Pauling, L., H. A. Itano *et al.* (1949) "Sickle-cell anemia, a molecular disease." Science **110**: 543-48

Pena, S. D., de Souza KT, de Andrade M and R. Chakraborty (1994). "Allelic associations of two polymorphic microsatellites in intron 40 of the human von Willebrand factor gene." Proc Natl Acad Sci U S A **91**(2): 723-7.

Peretz, H., Mulai A, Usher S, Zivelin A, Segal A, Weisman Z, Mittelman M, Lupo H, Lanir N, Brenner B, Shpilberg O and U. Seligsohn (1997). "The two common mutations causing factor XI deficiency in Jews stem from distinct founders: one of ancient Middle Eastern origin and another of more recent European origin." Blood **90**(7): 2654-9.

Pérez-Lezaun, A., , Calafell F, Mateu E, Comas D, Bosch E and J. Bertranpetit (1997). "Allele frequencies for 20 microsatellites in a worldwide population survey." Hum Hered **47**(4): 189-96.

Pérez-Lezaun, A., Calafell F, Mateu E, Comas D, Ruiz-Pacheco R and J. Bertranpetit (1997). "Microsatellite variation and the differentiation of modern humans." Hum Genet **99**(1): 1-7.

Pérez-Lezaun, A., Calafell F, Seielstad M, Mateu E, Comas D, Bosch E and J. Bertranpetit (1997). "Population genetics of Y-chromosome short tandem repeats in humans." J Mol Evol **45**(3): 265-70.

Persichetti F, Blasi P, Hammer M, Malaspina P, Jodice C, Terrenato L and A. Novelletto (1992) "Disequilibrium of multiple DNA markers on the human Y chromosome." Ann Hum Genet **56** (4):303-10

Pesso, R., Barkai G, Ravia Y, Gak E, Frydman M, Goldman B and E. Friedman (1997). "No founder effect detected in Jewish Ashkenazi patients with fragile-X syndrome." Hum Genet **101**(2): 186-9.

Piazza, A., Rendine S, Minch E, Menozzi P, Mountain J and L.L. Cavalli-Sforza (1995) "Genetics and the origin of European languages" Proc Natl Acad Sci U S A **13**:5836-40

Pollock DD, and D.B. Goldstein *(1995)* A comparison of two methods for constructing evolutionary distances from a weighted contribution of transition and transversion differences. Mol Biol Evol **12**(4):713-7

Poloni, E. S., Semino O, Passarino G, Santachiara-Benerecetti AS, Dupanloup I, Langaney A and L. Excoffier (1997). "Human genetic affinities for Y-chromosome P49a,f/TaqI haplotypes show strong correspondence with linguistics." Am J Hum Genet **61**(5): 1015-35.

Primmer, C.R., Saino N, Moller AP and H. Ellegren (1996) " Directional evolution in germline microsatellite mutations." Nat Genet **13**:391-3

Rahuel, C, Ngo KY, Cartron JP and G. Lucotte (1984) "DNA polymorphis of the RC8 sequence on the short arm of the X chromosome in a French population". Hum. Hered. **34**: 59-61

Raymond, M. and F. Rousset (1995). "GENEPOP (Version 1.2): Population genetics software for exact tests and ecumenicism." J. Hered. **86**(3): 248-249.

Reich, D. E. and D. B. Goldstein (1998). "Genetic evidence for a Paleolithic human population expansion in Africa." Proc Natl Acad Sci U S A **95**(14): 8119-23.

Relethford, J. H., F.C. Lees and M.H. Crawford (1980) " Population structure and anthropometric variation in rural western Ireland: migration and biological differentiation" Ann Hum Biol **7**:411-28

Relethford, J. H. (1988). "Effects of English admixture and geographic distance on anthropometric variation and genetic structure in 19th-century Ireland." Am J Phys Anthropol **76**(1): 111-24.

Relethford, J. H. (1997). "Mutation rate and excess African heterozygosity." Hum Biol **69**(6): 785-92.

Relethford, J. H. and M. H. Crawford (1995). "Anthropometric variation and the population history of Ireland." Am J Phys Anthropol **96**(1): 25-38.

Relethford, J. H. and H. C. Harpending (1994). "Craniometric variation, genetic theory, and modern human origins." Am J Phys Anthropol **95**(3): 249-70.

Relethford, J. H. and L. B. Jorde (1999) "Genetic evidence for a larger African population size during recent human evolution." Am J Phys Anthropol **108:** 251-60.

Renfrew, C. (1987) Archaeology and language: The puzzle of Indo-European origins. London, Jonathan Cape

Renfrew, C. (1989). "The Origins of Indo-European languages." Scientific American **261**(4): 106-114.

Richards, M., Corte-Real H, Forster P, Macaulay V, Wilkinson-Herbots H, Demaine A, Papiha S, Hedges R, Bandelt HJ and B. Sykes (1996). "Paleolithic and neolithic lineages in the European mitochondrial gene pool." Am J Hum Genet **59**(1): 185-203.

Richards, R. I., Holman K, Friend K, Kremer E, Hillen D, Staples A, Brown WT, Goonewardena P, Tarleton J, Schwartz C, *et al.* (1992). "Evidence of founder chromosomes in fragile X syndrome." Nat Genet **1**(4): 257-60.

Richards, R.I. and G.R. Sutherland (1994) "Simple repeat DNA is not replicated simply." Nat Genet **6**:114-6

Richards M.B,. Macaulay VA, Bandelt HJ and B.C. Sykes (1998) "Phylogeography of mitochondrial DNA in western Europe." Ann Hum Genet **62**: 241-60

Roewer, L., Kayser M, Dieltjes P, Nagy M, Bakker E, Krawczak M and P. de Knijff (1996). "Analysis of molecular variance (AMOVA) of Y-chromosome-specific microsatellites in two closely related human populations." Hum Mol Genet **5**(7): 1029-33.

Rogers, A. and H. Harpending (1992). "Populations growth makes waves in the distribution of pairwise differences." Molecular Biology and Evolution **9**: 552-569.

Rogers, A.R, Fraley AE, Bamshad MJ, Watkins WS and L.B. Jorde (1996) "Mitochondrial mismatch analysis is insensitive to the mutational process." Mol Biol Evol **13**(7):895-902

Rogers, A. and L. Jorde (1995). "Genetic evidence on modern human origins." Human Biology **67**: 1-36.

Ruhlen, M (1987) A guide to the world's languages. Stanford, Stanford University Press

Ruiz Linares, A., Nayar K, Goldstein DB, Hebert JM, Seielstad MT, Underhill PA, Lin AA, Feldman MW and L.L. Cavalli Sforza (1996). "Geographic clustering of human Y-chromosome haplotypes." Ann Hum Genet **60**(Pt 5): 401-8.

Ruiz-Linares, A., Ortiz-Barrientos D, Figueroa M, Mesa N, Munera JG, Bedoya G, Velez ID, Garcia LF, Perez-Lezaun A, Bertranpetit J, Feldman MW and D.B. Goldstein DB (1999) Microsatellites provide evidence for Y chromosome diversity among the founders of the New World. Proc Natl Acad Sci U S A **96**(11):6312-7

Ryan, W. and W. Pitman (1998). Noah's Flood. New York, Simon and Schuster.

Saitou, N. and M. Nei (1987). "The neighbour-joining method: a new method for reconstructing phylogenetic trees." Molecular Biology and Evolution **4**: 406-425.

Sajantila, A., Salem AH, Savolainen P, Bauer K, Gierig C and S. Paabo (1996). "Paternal and maternal DNA lineages reveal a bottleneck in the founding of the Finnish population." Proc Natl Acad Sci U S A **93**(21): 12035-9.

Sajantila, A., M. Lukka and A.C. Syvanen (1999) "Experimentally observed germline mutations at human micro- and minisatellite loci." Eur J Hum Genet **7**:263-6

Salem, A. H., Badr FM, Gaballah MF and S. Paabo (1996) "The genetics of traditional living: Y-chromosomal and mitochondrial lineages in the Sinai Peninsula." Am J Hum Genet **59**(3):741-3

Santos, F. R., N. O. Bianchi and S.D. Pena (1996). "Worldwide distribution of human Y-chromosome haplotypes." Genome Res **6**(7): 601-11.

Santos, F. R. & Tyler-Smith, C. (1996) "Reading the human Y chromosome: the emerging DNA markers and human genetic history". Braz. J. Genet. **19**, 665-670.

Santos, F. R., Pandya A, Tyler-Smith C, Pena SD, Schanfield M, Leonard WR, Osipova L, Crawford MH and R.J. Mitchell (1999) "The central Siberian origin for native American Y chromosomes." Am J Hum Genet **64**(2):619-28

Savatier, P., Trabuchet G, Faure C, Cheboune Y, Gouy M, Verdier G and V.M. Nigon (1985) "Evolution of the primate beta-globin gene region: High rate of variation in CpG dinucleotides and in short repeated sequences between man and chimpanzee" J. Mol. Biol. **182**: 21-29

Sawada, I., Willard C, Shen CK, Chapman B, Wilson AC and C.W. Schmid (1985) "Evolution of Alu family repeats since the divergence of human and chimpanzee." J Mol Evol **22**(4):316-22

Saxena, R., Brown LG, Hawkins T, Alagappan RK, Skaletsky H, Reeve MP, Reijo R, Rozen S, Dinulos MB, Disteche CM and D.C. Page (1996) "The DAZ gene cluster on the human Y chromosome arose from an autosomal gene that was transposed, repeatedly amplified and pruned." Nat Genet **14**(3):292-9

Sayos, J., Wu C, Morra M, Wang N, Zhang X, Allen D, van Schaik S, Notarangelo L, Geha R, Roncarolo MG, Oettgen H, De Vries JE, Aversa G and C. Terhorst (1998). "The X-linked lymphoproliferative-disease gene product SAP regulates signals induced through the co-receptor SLAM." Nature **395**(6701): 462-9.

Schlessinger D, Little RD, Freije D, Abidi F, Zucchi I, Porta G, Pilia G, Nagaraja R, Johnson SK, Yoon JY, *et al.* (1991) "Yeast artificial chromosome-based genome mapping: some lessons from Xq24-q28." Genomics **11**(4):783-93

Schlotterer, C. and D. Tautz (1992) "Slippage synthesis of simple sequence DNA." Nucl. Acids. Res. **20**: 211-215

Schneider, S., J. Kueffer, *et al.* (1997). Arlequin ver. 1.1: A software for population genetic analysis. Genetics and Biometry Laboratory,University of Geneva,Switzerland.

Scozzari, R., Cruciani F, Malaspina P, Santolamazza P, Ciminelli BM, Torroni A, Modiano D, Wallace DC, Kidd KK, Olckers A, Moral P, Terrenato L, Akar N, Qamar R, Mansoor A, Mehdi SQ, Meloni G, Vona G, Cole DE, Cai W and A. Novelletto (1997). "Differential structuring of human populations for homologous X and Y microsatellite loci." Am J Hum Genet **61**(3): 719-33.

Seielstad M. T. (1994). "Construction of human Y-chromosomal haplotypes using a new polymorphic A to G transition." Human Molecular Genetics **3**(12): 2159-2161.

Seielstad, M.T., E. Minch, and L. L. Cavalli-Sforza (1998) "Genetic evidence for a higher female migration rate in humans." Nature Genet. **20**, 278-280..

Semino, O., Passarino G, Brega A, Fellous M and A.S. Santachiara-Benerecetti (1996). "A view of the neolithic demic diffusion in Europe through two Y chromosome-specific markers." Am. J. Hum. Genet. **59**, 964-968

Sherman, S. L., Jacobs PA, Morton NE, Froster-Iskenius U, Howard-Peebles PN, Nielsen KB, Partington MW, Sutherland GR, Turner G and M. Watson (1985). "Further segregation analysis of the fragile X syndrome with special reference to transmitting males." Hum Genet **69**(4): 289-99.

Sherry, S. T., Rogers AR, Harpending H, Soodyall H, Jenkins T and M. Stoneking (1994). "Mismatch distributions of mtDNA reveal recent human population expansions." Human Biology **66**: 761-775.

Shriver, M. D., L. Jin and E. Boerwinkle (1993) "VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach" Genetics **134**: 983-993

Shriver, M. D., Jin L, Boerwinkle E, Deka R, Ferrell RE and R. Chakraborty (1995). "A novel measure of genetic distance for highly polymorphic tandem repeat loci." Molecular Biology and Evolution **12**(5): 914-920.

Shriver, M. D., Jin L, Ferrell RE and R. Deka (1997). "Microsatellite data support an early population expansion in Africa." Genome Res **7**(6): 586-91.

Shurmann, M., R. Warneke and E. Schwinger (1987) "Four DNA polymorphisms on the short arm of the X chromosome: allele frequencies in a German and in a Turkish population." Hum. Hered. **37**: 329-33

Sibley, C. G. and J. E. Ahlquist (1984). "The phylogeny of the hominid primates, as indicated by DNA-DNA hybridization." J. Mol. Evol. **20**: 2-15.

Slatkin, M. (1994) "Linkage disequilibrium in growing and stable populations" Genetics **137**: 331-36

Slatkin, M. (1995). "A measure of population subdivision based on microsatellite allele frequencies." Genetics **139**: 457-462.

Slatkin, M. and L. Excoffier (1996). "Testing for linkage disequilibrium in genotypic data using the Expectation-Maximization algorithm." Heredity **76**(Pt 4): 377-83.

Slatkin, M. and R. R. Hudson (1991) "Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations." Genetics **129**: 555-62

Slatkin, M. and B. Rannala (1997). "Estimating the age of alleles by use of intraallelic variability." American Journal of Human Genetics **60**: 447-458.

Smyth, W. J. (1997). In Search of Ireland, a cultural geography. London, Routledge.

Sokal, R. R., N. L. Oden, and C. Wilson (1991) "Genetic evidence for the spread of agriculture in Europe by demic diffusion." Nature **351**(6322):143-5

Sokal, R.R., N. L. Oden and B.A. Thompson (1999) "The problem with synthetic maps" Hum Biol. **71**(3):447-53.

Spiegel R, La Spada AR, Kress W, Fischbeck KH and W. Schmid (1996) "Somatic stability of the expanded CAG trinucleotide repeat in X-linked spinal and bulbar muscular atrophy." Hum Mutat **8**:32-7

Spurdle, A., and T. Jenkins (1992) "Y chromosome probe p49a detects complex PvuII haplotypes and many new TaqI haplotypes in southern African populations." Am J Hum Genet **50**(1):107-25

Spurdle, A. B. and T. Jenkins (1992). "The Y chromosome as a tool for studying human evolution." Curr Opin Genet Dev **2**(3): 487-91.

Spurdle, A. B. and T. Jenkins (1993) "Complex polymorphisms are revealed by Y chromosome probe 49a with BglII, HindIII, PstI and SstI." Ann Hum Genet **57** (1):41-53

Spurdle, A. B., Woodfield DG, Hammer MF and T. Jenkins T (1994) "The genetic affinity of Polynesians: evidence from Y chromosome polymorphisms." Ann Hum Genet **58** (Pt 3):251-63

Starr, T. and S. Wood (1987). "Isolation and characterization of DNA probes from the short arm of the human X chromosome that detect restriction fragment length polymorphisms." Genome **29**(1): 201-5.

Stephan, W. and Y. Kim (1998). "Persistence of microsatellite arrays in finite populations." Mol Biol Evol **15**(10): 1332-6.

Stephens, J. C., Reich DE, Goldstein DB, Shin HD, Smith MW, Carrington M, Winkler C, Huttley GA, Allikmets R, Schriml L, Gerrard B, Malasky M, Ramos MD, Morlot S, Tzetis M, Oddoux C, di Giovine FS, Nasioulas G, Chandler D, Aseev M, Hanson M, Kalaydjieva L, Glavac D, Gasparini P, Dean M, *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.

Stoneking, M. and R. L. Cann (1989). African origin of human mitochondrial DNA. Princeton, N.J., Princeton University Press.

Stringer, C. B. (1985). Middle Pleistocene variability and the origin of late Pleistocene ancestors. New York, A.R. Liss.

Stringer, C. B. (1986). Documenting the origin of modern humans. Cambridge, Cambridge University Press.

Stringer, C. B. (1995) "The evolution and distribution of later Pleistocene human populations". In Paleoclimate and evolution with emphasis on human origins. E.S. Verba, G.H. Denton, T.C. Partridge and L.H. Burckle, eds. New Haven, Yale University Press.

Stringer, C. B. and P. Andrews (1998) "Genetic and fossil evidence for the Origin of Modern Humans" Science **239**: 1263-1268

Summers, K. M. (1987). "Restriction fragment length polymorphisms on the X chromosome in populations from Papua New Guinea, Thailand, northwestern Australia and China." Gene Geogr **1**(3): 207-17.

Sutherland G. R., and R.I. Richards (1995) "Simple tandem DNA repeats and human genetic disease." Proc Natl Acad Sci U S A **92**:3636-41

Swisher, C. C., Curtis GH, Jacob T, Getty AG, Suprijo A and Widiasmoro (1994). "Age of the earliest known hominids in Java, Indonesia." Science **263**(5150): 1118-21.

Sykes, B. (1999) "The molecular genetics of European ancestry" Phil Trans R Soc Lond B **354**: 131-139

Tai, D. Y., J. P. Pan and G.J. Lee-Chen (1998). "Identification and haplotype analysis of apolipoprotein B-100 Arg3500-- >Trp mutation in hyperlipidemic Chinese." Clin Chem **44**(8 Pt 1): 1659-65.

Takezaki, N. and M. Nei (1996). "Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA." Genetics **144**: 389-399.

Tautz, D. (1989) "Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucl. Acids. Res. **17**: 6463-6471

Tautz, D. and M. Renz (1984) "Simple sequences are ubiquitous repetitive components of eukaryotic genomes" Nucl. Acids. Res. **12**: 4127-4138

Thomas, M. G., Skorecki K, Ben-Ami H, Parfitt T, Bradman N and D.B. Goldstein (1998) "Origins of Old Testament priests." Nature **394**(6689):138-40

Tills D., Teesdale P. and A. E. Mourant (1977) "Blood groups of the Irish." Ann Hum Biol 4:23-4

Tishkoff, S. A., Dietzsch E, Speed W, Pakstis AJ, Kidd JR, Cheung K, Bonne-Tamir B, Santachiara-Benerecetti AS, Moral P and M. Krings (1996). "Global patterns of linkage disequilibrium at the CD4 locus and modern human origins." Science **271**: 1380-1387.

Tishkoff, S. A., Goldman A, Calafell F, Speed WC, Deinard AS, Bonne-Tamir B, Kidd JR, Pakstis AJ, Jenkins T and K.K. Kidd (1998). "A global haplotype analysis of the myotonic dystrophy locus: implications for the evolution of modern humans and for the origin of myotonic dystrophy mutations." Am J Hum Genet **62**(6): 1389-402.

Torroni, A., Semino O, Scozzari R, Sirugo G, Spedini G, Abbas N, Fellous M and A.S. Santachiara Benerecetti (1990) "Y chromosome DNA polymorphisms in human populations: differences between Caucasoids and Africans detected by 49a and 49f probes." Ann Hum Genet 54 (4):287-96

Torroni, A., Chen YS, Semino O, Santachiara-Beneceretti AS, Scott CR, Lott MT, Winter M and D.C. Wallace (1994). "mtDNA and Y-chromosome polymorphisms in four Native american populations from southern Mexico." Am J Hum Genet **54**(2): 303-18.

Torroni, A. and D. C. Wallace (1994). "Mitochondrial DNA variation in human populations and implications for detection of mitochondrial DNA mutations of pathological significance." J Bioenerg Biomembr **26**(3): 261-71.

Torroni, A., Bandelt HJ, D'Urbano L, Lahermo P, Moral P, Sellitto D, Rengo C, Forster P, Savontaus ML, Bonne-Tamir B and R. Scozzari (1998). "Mitochondrial

DNA analysis reveal a major late Paleolithic population expansion from southwestern to northeastern Europe. Am. J. Hum Genet. 62: 1137-52

Underhill, P. A., Jin L, Zemans R, Oefner PJ and L.L. Cavalli-Sforza (1996). "A pre-Columbian Y chromosome-specific transition and its implications for human evolutionary history." Proc Natl Acad Sci U S A **93**(1): 196-200.

Underhill, P.A., Jin L, Lin AA, Mehdi SQ, Jenkins T, Vollrath D, Davis RW, Cavalli-Sforza LL and P.J. Oefner (1997) "Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography." Genome Res. **7:** 996-1005.

Valdes, A. M., M. Slatkin and N.B. Freimer (1993). "Allele frequencies at microsatellite loci: the stepwise mutation model revisited." Genetics **133**: 737-749.

Vallades, H., J. L. Reyss, *et al.* (1988). "Thermoluminescence dating of Mousterian 'Proto-Cro-Magnon' remains from Israel and the origin of modern man." Nature **331**: 614-616.

Vandermeersch, B. (1985). The origin of the Neanderthals. New York, A.R. Liss.

Veitia, R., Ion A, Barbaux S, Jobling MA, Souleyreau N, Ennis K, Ostrer H, Tosi M, Meo T, Chibani J, Fellous M and K. McElreavey (1997). "Mutations and sequence variants in the testis-determining region of the Y chromosome in individuals with a 46, XY female phenotype." Human Genetics **99**(5): 648-652.

Vetrie, D. (1993) "Isolation of the defective gene in X linked agammaglobulinaemia." J Med Genet **30**(6):452-3

Vetrie, D., Kendall E, Coffey A, Hassock S, Collins J, Todd C, Lehrach H, Bobrow M, Bentley DR and A. Harris (1994). "A 6.5-Mb yeast artificial chromosome contig incorporating 33 DNA markers on the human X chromosome at Xq22." Genomics **19**: 42-47.

Vigilant, L., , Stoneking M, Harpending H, Hawkes K and A.C. Wilson (1991). "African populations and the evolution of human mitochondrial DNA." Science **253**: 1503-1508.

Von Haeseler, A., A. Sajantila and S. Paabo (1996) "The genetic archaeology of the human genome. Nat. Genet. **14**: 135-40

Waddel, J. (1998). The prehistoric archaeology of Ireland. Galway, Galway University Press.

Wadhwa, R., Papiha S, Lester D, Ray V, Saha N and S. Bhattacharya (1989). "X chromosome restriction fragment length polymorphisms in five racial groups: rare variant

detected with the RC8 (DXS9) probe in the Marathi population, India." Hum Hered **39**(6): 309-12.

Wang, B., Fu S, Sun Y, Huang C, Yang H, Xue Y, Wang Q, Zhang G and P. Li (1999) " Absence of the derived allele at the DYS199 locus in 3 NE Chinese populations. " Hum Hered **49**(1):63-4

Watterson, G.A., and H.A. Guess (1977) "Is the most frequent allele the oldest?" Theor Popul Biol **11**(2):141-60

Weber, J.L. (1990) "Informativeness of human (dC-dA)n•(dG-dT) polymorphisms." Genomics **7**: 524-530

Weber, J.L. and P.E. May (1989) "Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction." Am. J. Hum. Genet. **2**: 1123-1128

Weber, J. L. and C. Wong (1993). "Mutation of human short tandem repeats." Human Molecular Genetics **2**: 1123-1128.

Weidenreich, F. (1947). "The trend of human evolution." Evolution **1**: 436-442.

Weir, B. S. and C. C. Cockerham (1984). "Estimating F-statistics for the analysis of population structure." Evolution **38**: 1358-1370.

Weiss, K.M. (1973) "Demographic models for anthropology." Am. Antiquity **38**, 1-186

Weissenbach J, Gyapay G, Dib C, Vignal A, Morissette J, Millasseau P, Vaysseix G and M. Lathrop (1992) "A second-generation linkage map of the human genome." Nature 359:794-801

Wenke, R.J. (1990) Patterns in prehistory: Humankind's first three million years. New York and Oxford, Oxford University Press.

Whitfield, L. S. (1995). "41kb of analysed sequence from the pseudoautosomal and sex determining regions of the short arm of the human Y-chromosome." Genomics **27**(2): 306-311.

Willard HF, and J.R. Riordan (1985) "Assignment of the gene for myelin proteolipid protein to the X chromosome: implications for X-linked myelin disorders." Science **230**(4728):940-2

Willems, P.J. (1994) "Dynamic mutations hit double figures." Nat Genet **8**:213-5.

Wilson, A. C. and V. M. Sarich (1969). "A molecular time scale for human evolution." Proc. Natl. Acad. Sci., USA **63**: 1088-1093.

Wolpoff, M. H. (1989). Multiregional evolution: The fossil alternative to Eden. Princeton, Princeton University Press.

Wolpoff, M. H., X. Wu and A.G. Thorne (1984) "Modern Homo sapiens origins: A general theory of hominid evolution involving fossil evidence from East Asia. In The origins of Modern Humans, A world survey of the fossil evidence. F.H. Smith and F. Spencer, eds. New York: Alan R. Liss

Wood, B. (1996) "Human Evolution" Bioessays **18**: 945-54

Woodman, P. C. (1985). Excavations at Mount Sandel, 1973-77. Belfast.

Worton, R.G., P.N. Ray et al (1988) "The problem of Duchenne muscular dystrophy." Philos Trans R Soc Lond B Biol Sci **319**(1194):275-84

Wright, S. (1951). "The genetical structure of populations." Ann. Eugenics **15**(323-354).

Zei, G., Barbujani G, Lisa A, Fiorani O, Menozzi P, Siri E and L.L. Cavalli-Sforza (1993) "Barriers to gene flow estimated by surname distribution in Italy." Ann Hum Genet **57** (Pt 2):123-40

Zerjal, T. (1997). "Genetic relationships of Asians and Northern Europeans, revealed by Y-chromosomal DNA analysis." American Journal of Human Genetics **60**(5): 1174-1183.

Zhong, N., Kajanoja E, Smits B, Pietrofesa J, Curley D, Wang D, Ju W, Nolin S, Dobkin C, Ryynanen M and W.T. Brown (1996). "Fragile X founder effects and new mutations in Finland." Am J Med Genet **64**(1): 226-33.

Zietkiewicz, E., Yotova V, Jarnik M, Korab-Laskowska M, Kidd KK, Modiano D, Scozzari R, Stoneking M, Tishkoff S, Batzer M and D. Labuda (1998). "Genetic structure of the ancestral population of modern humans." J Mol Evol **47**(2): 146-55.

Zori, R. T., Gardner JL, Zhang J, Mullan MJ, Shah R, Osborn AR, Houlden H, Wallace MR, Roberts S and T.P. Yang (1998). "Newly described form of X-linked arthrogryposis maps to the long arm of the human X chromosome." Am J Med Genet **78**(5): 450-4.

Zschocke, J., Mallory JP, Eiken HG and N.C. Nevin (1997). "Phenylketonuria and the peoples of Northern Ireland." Human Genetics **100**: 189-194

Zvelebil, M. (1986) Hunters in Transition: Mesolithic societies of Temperate Eurasia and their transition to farming. Cambridge, Cambridge University Press

Zvelebil, M. and P.R. Rowley-Conwy (1986) "Foragers and farmers in Atlantic Europe" In Hunters in transition: Mesolithic societies of Temperate Eurasia and their transition to farming. M. Zvelebil ed., Cambridge, Cambridge University Press

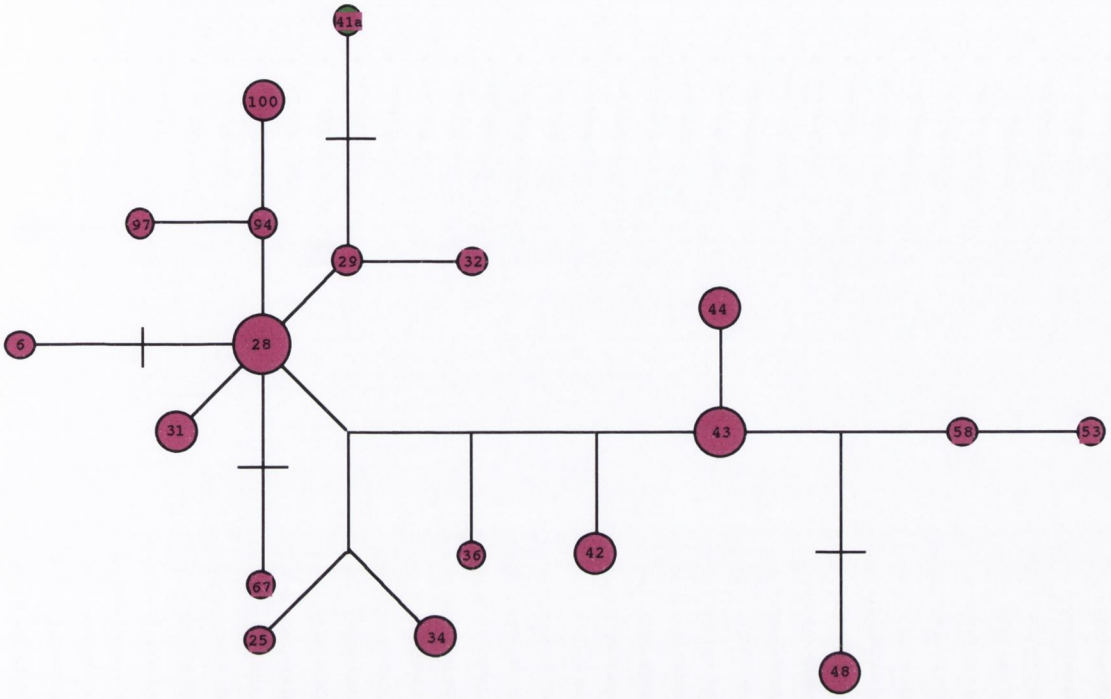
Appendices

Appendix A

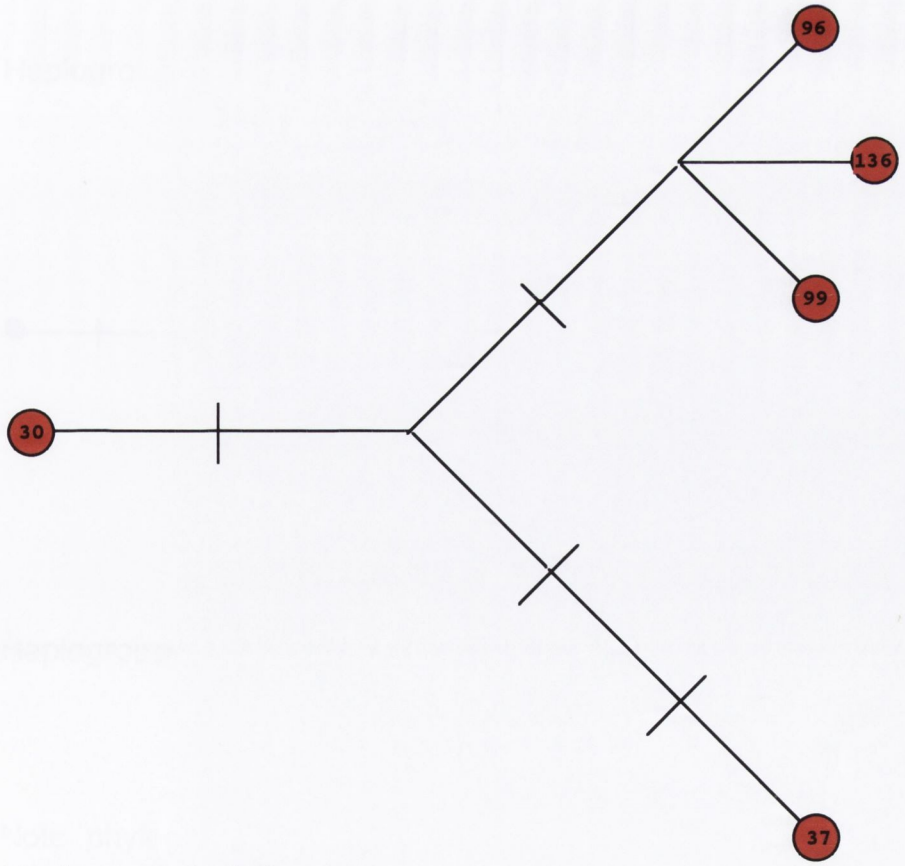
Real Ht	Sampling Ht	177e8-F3	177e8-F7	177e8-F13	177e8-F14	177e8-F29
1	11	160	166	113	131	154
2	13	160	168	111	131	162
3	26	160	172	111	131	154
4	52	160	174	111	131	160
5	53	160	174	111	131	162
7	3	162	156	115	131	150
8	4	162	162	111	131	154
9	5	162	164	111	131	154
10	6	162	164	111	131	156
11	7	162	164	111	134	154
12	8	162	166	105	131	156
16	9	162	168	111	131	154
17	10	162	170	115	134	148
18	11	162	170	115	134	154
20	35	162	172	115	131	154
21	41	162	172	115	131	156
22	36	162	172	115	134	154
23	42	162	172	115	134	156
26	1	164	152	115	134	152
27	2	164	152	115	134	154
30	6	164	162	115	131	154
32	7	164	168	115	131	154
35	8	164	168	115	134	154
40	9	164	170	115	131	154
41	10	164	170	115	131	156
42	11	164	170	115	131	158
46	12	164	170	119	131	156
48	13	164	172	111	131	166
50	14	164	172	113	131	152
55	15	164	172	115	131	150
56	37	164	172	115	131	154
57	43	164	172	115	131	156
58	45	164	172	115	131	158
59	47	164	172	115	131	160
62	48	164	172	115	134	152
63	38	164	172	115	134	154
65	49	164	172	117	131	156
66	48	164	172	117	134	154
68	50	164	172	119	131	156
69	51	164	172	119	134	156
70	54	164	174	111	131	164
72	55	164	174	115	131	150
74	56	164	174	115	131	154
75	57	164	174	115	131	156
77	58	164	174	115	131	162
81	59	164	174	117	131	154
82	60	164	174	117	134	154
85	63	164	176	115	131	162
87	64	164	178	115	131	154
89	65	164	178	115	134	154
90	5	166	162	113	134	156
93	6	166	168	119	131	154
94	7	166	170	115	131	156
95	8	166	172	113	131	154
96	39	166	172	115	131	154
98	34	166	172	115	134	152
99	40	166	172	115	134	154
100	46	166	172	115	134	158
104	61	166	176	115	134	154
106	62	166	176	115	137	154
109	63	168	168	119	131	156
110	64	168	170	111	134	154
unknown	65	164	172	111	131	164
unknown	66	164	172	115	131	152
unknown	44	164	172	115	134	156

Appendix A: Haplotype identities from entire sample set and the corresponding haplotypes sampled in the subset of samples. Three of the haplotypes that are present in the smaller sample, and not in the entire sample, may have been detected after the larger analysis had been performed.

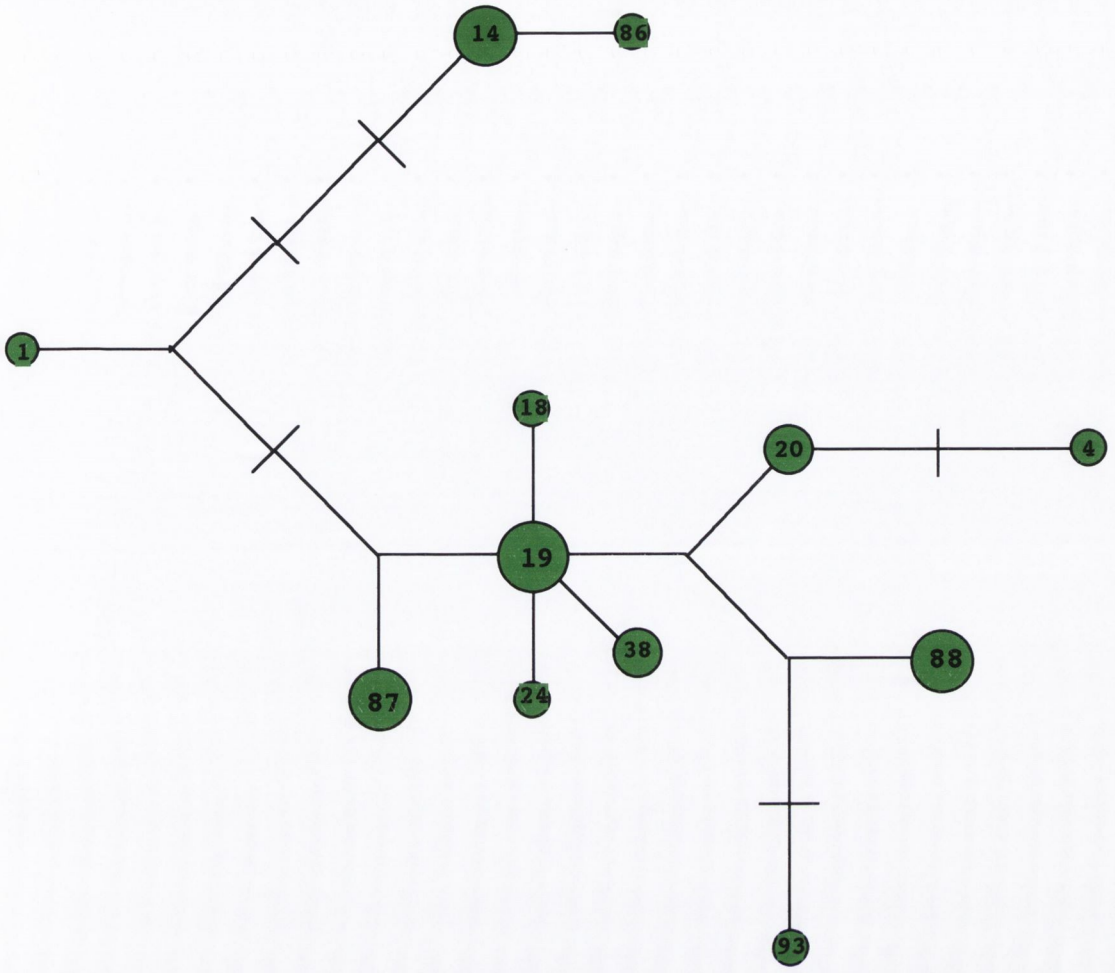
Appendix B



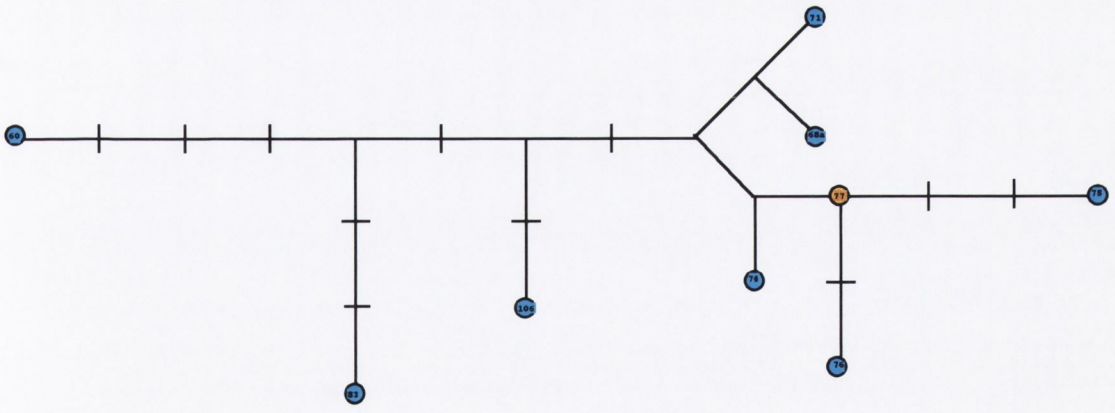
Haplogroup 8 phylogeny



Haplogroup 3 phylogeny



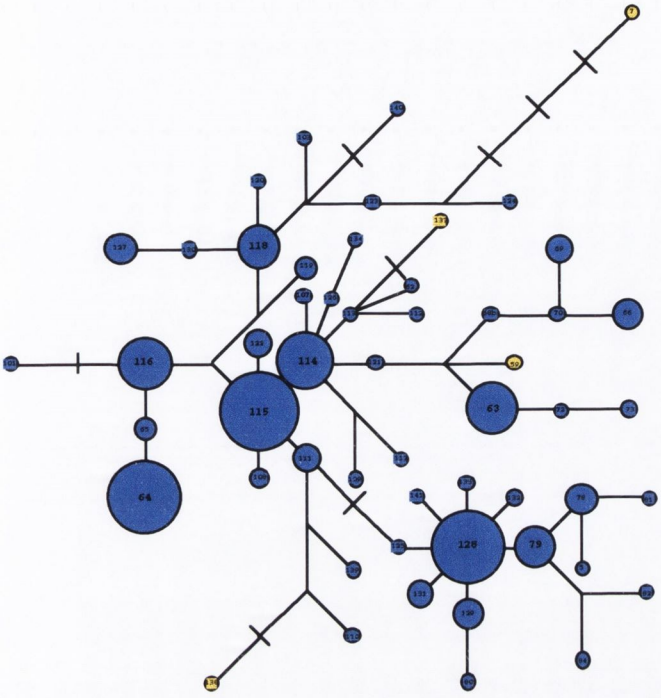
Haplogroup 21 phylogeny



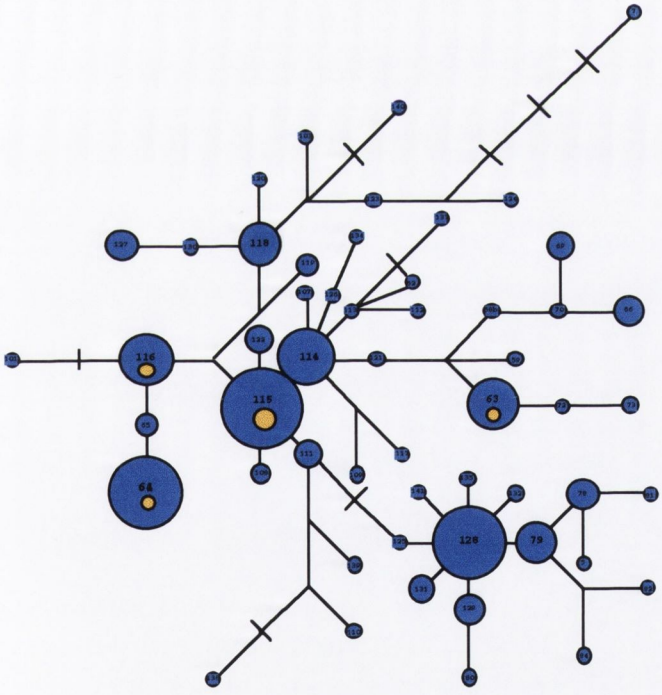
Haplogroup 26 phylogeny

Note: phylogenies in appendix B are not to scale

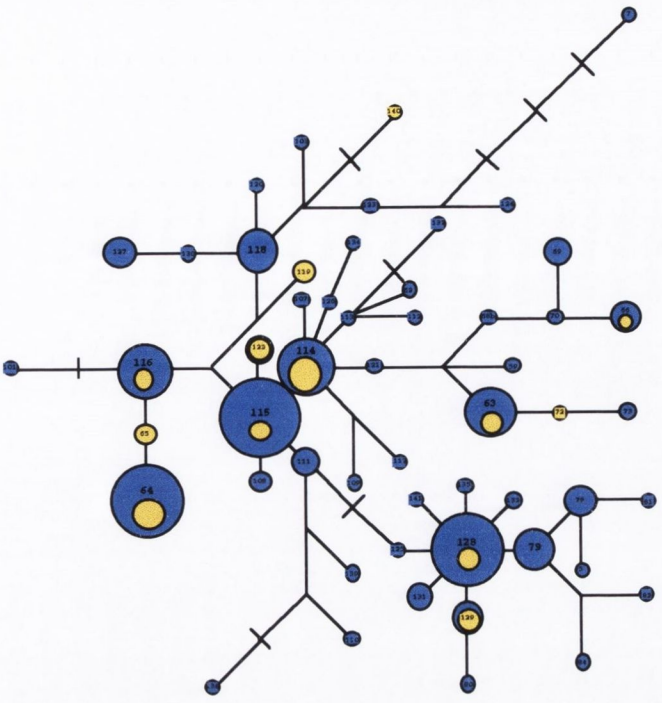
Appendix C



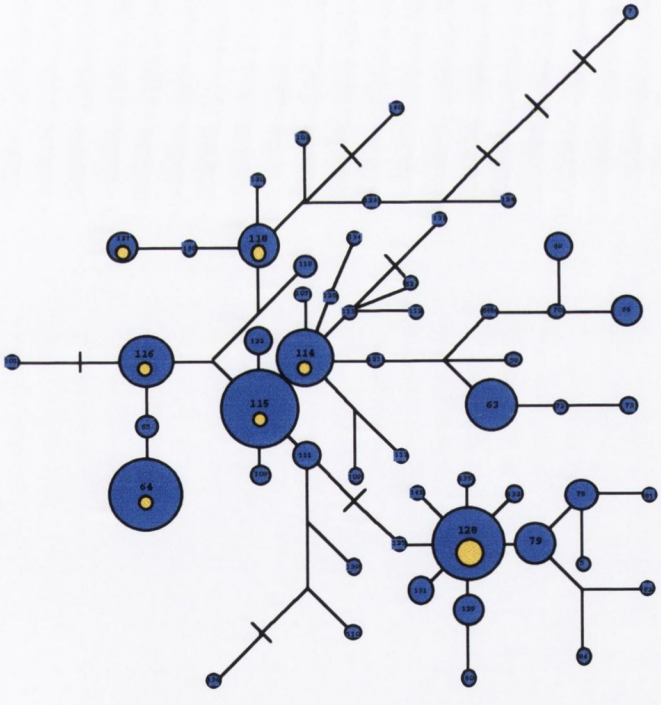
Hg 1 chromosomes in the Turkish sample



Hg 1 chromosomes in the Danish sample



Hg 1 chromosomes in the English sample



Hg 1 chromosomes in the Scottish sample