Mutation and variability of the human Y chromosome

Thesis submitted for the degree of

Doctor of Philosophy

at the University of Leicester

by

Matthew Hurles BA(Hons) Oxon

Department of Genetics

University of Leicester

October 1999

UMI Number: U123540

All rights reserved

INFORMATION TO ALL USERS
The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript
and there are missing pages, these will be noted. Also, if material had to be removed,
a note will indicate the deletion.

UMI U123540
Published by ProQuest LLC 2013. Copyright in the Dissertation held by the Author.
Microform Edition © ProQuest LLC.
All rights reserved. This work is protected against
unauthorized copying under Title 17, United States Code.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

"I see humanity as a family that has hardly met. I see the meeting of people, bodies, thoughts, emotions or actions as the start of most change. Each link created by a meeting is like a filament which if they were all visible would make the world look as though it is covered by gossamer. Every individual is connected to others, loosely or closely, by a combination of filaments which stretch across the frontiers of space and time."

Theodore Zeldin 'An intimate history of humanity' pp465-466

Mutation and variability of the human Y chromosome

Matthew Hurles

Abstract

The human Y chromosome is inherited from father to son, determining maleness. As a result of its being haploid the vast majority of the Y chromosome does not undergo recombination at meiosis. Consequently polymorphisms on this non-recombining portion represent a simple record of male evolutionary history. It is important to distinguish between polymorphisms that have only arisen once during human evolution and those faster mutating markers, such as microsatellites and minisatellites, that are recurrent. Unique biallelic markers can be used to define monophyletic lineages and recurrent markers used to assay diversity within them. Improving the resolution of Y-chromosomal evolutionary studies is contingent upon discovering more biallelic markers. To this end, a recently-developed high-throughput mutation detection technique is used to screen Y-specific sequences for single nucleotide polymorphisms, resulting in the discovery of a new, African-specific, polmorphism. The utility of adopting a genealogical approach to analysing the different types of Y-chromosomal polymorphic information is investigated. Two population-specific lineages are investigated in depth, requiring the development of new analytical methodologies. Investigating the genetic landscape of a region requires the assaying of multiple lineages in many populations. Local heterogeneities, specifically significant barriers to gene flow, within this genetic landscape can be defined and compared to the geographical and cultural landscapes of the same region. Such an analysis of European Y-chromosomal diversity reveals that genetic barriers to male gene-flow are well correlated with linguistic boundaries. Ever since the discovery that isolated islands in the Pacific Ocean were already densely populated by Polynesians, this region of the world has intrigued anthropologists of all persuasions. An analysis of Y-chromosomal diversity in Southeast Asia and Oceania reveals that the present informative capacity of the human Y chromosome is sufficient to contribute meaningfully to many of the contentious issues debated heatedly amongst Pacific prehistorians.

Publications

Hurles M.E., Irven C., Nicholson J., Taylor P.G., Santos F.R., Loughlin J., Jobling M.A. and Sykes B.C. (1998) European Y-chromosomal lineages in Polynesia: A contrast to the population structure revealed by mtDNA. *American Journal of Human Genetics* **63** 1793-1806

Hurles M.E., Veitia R., Arroyo E., Armenteros M., Bertranpetit J., Pérez-Lezaun A., Bosch E., Shlumukova M., Cambon-Thomsen A., McElreavey K., López de Munain A., Röhl A., Wilson I.J., Singh L., Pandya A., Santos F.R., Tyler-Smith C. and Jobling M.A. (1999) Recent male-mediated gene flow over a linguistic barrier in Iberia suggested by analysis of a Y-chromosomal DNA polymorphism. *American Journal of Human Genetics* **65** 1437-1448

Acknowledgements

Most thanks must always go to a good supervisor but that is not sufficient to describe the all-encompassing support, friendship and guidance offered by Mark. A worthy mentor indeed for life and science, I owe you an unpayable debt of gratitude.

Much thanks also to:

Gabby for co-supervising in an minimalist capacity.

Zoe "Royston" for making the lab a far nicer place to work, lovely dinners and Joining the Losing Crusade against Grammatical Rigour. I'll miss you.

Neale for a precedent, footy and DHPLC solidarity.

Chris Tyler-Smith for getting me started in the first place, much advice and offering the continual pinprick of rigour to the balloons of my half-baked ideas.

Tatiana and Arpita for unpublished information and conference companionship

Penny, Phil and Ali - for bearing with me and making life more bearable

All those at LRC and ULBC who have made rowing so enjoyable; introducing me to multi-dimensional pleasures of weekend regattas, the winning feeling and the Henley Boat Tent, especially Norm, Phil, Tommy, Jaish, Steve, Alan and Rick. I like it!!

John for cleaning up after me, putting up with my blade and many fruitful and fanciful discussions.

My parents for their continuing, unquestioning, support; Mum for the original idea and Dad for getting better.

Jess for the sisterly loan of 'Bob' the laptop.

Collaborative thanks to:

Carole Sargent for primers and sequences

Guido Barbujani, Yuri Dubrova, Chris Tyler-Smith, David Goldstein, Alan Templeton and Fabricio Santos for much needed statistical support.

Guido and Antonio for being accomodating and generous hosts

Bryan, Catherine, John, Jayne and John Clegg for Oceanic discussions, samples and data.

The members of the Iberian and European diversity studies for samples and data.

List of abbreviations

AD - Anno Domini
AFLP - Amplified fragment length polymorphism
AMH - Anatomically modern humans
AMOVA - Analysis of molecular variance
ASD - Average squared distance
ATP - Adenosine triphosphate
AZF - Azoospermia factor
bp - base pairs
BP - Before present
CAR - Central African Republic
CD-CV - Common disease - Common variant
CI - 95% Confidence interval
CIA - Central Intelligence Agency
CIs - Cook Islands
cpDNA - Chloroplast DNA
DBS - DisplayBarrierSig
DHPLC - Denaturing high performance liquid chromatography
DNA - Deoxyribose nucleic acid
dNTPs- Deoxy nucleotide triphosphates
DGGE - Denaturing gradient gel electrophoresis
EBS - EuroBarrierSig
ECT - Evolutionary culture theory
EDTA - Ethylenediamine tetraacetic acid
GCD - Great circle distances
GuSCN - Guanidinium thiocyanate
HCl - Hydrochloric acid
Hg - Haplogroup
HGMD - Human genome mutation database
HLA - Human leukocyte antigen
HMSNL -Hereditary motor-sensory neuropathy type Lom
HPLC - High peformance liquid chromatography
HVS - Hypervariable Segment
IBD - Identity by descent
IBS - Identity by state
IDL - Interactive Data Language

IDS - Inverse distance squared
kb - Kilobase pairs
KK - Kota Kinabalu
Mb - Megabase pairs
MDS - Multi-dimensional scaling
M-J - Median-Joining
MP - Malayo-Polynesian
MRCA - Most recent common ancestor
M-S - Minimum-Spanning
mtDNA - mitochondrial DNA
MVR - Minisatellite variant repeat
NCA - Nested cladistic analysis
OD - Optical density
P-AN - Proto-Austronesian
PAR - Pseudo-autosomal region
PC - Principal component
PCA - Principal components analysis
PCR - Polymerase chain reaction
PNG - Papua New Guinea
P-NP - Proto-Nuclear-Polynesian
P-P - Proto-Polynesian
P-SO - Proto-Samoic-outlier
RAM - Random Access Memory
RAPD - Randomly amplified polymorphic DNA
RFLP - Restriction fragment length polymorphism
RNA - Ribose nucleic acid
SNP - Single nucleotide polymorphism
SRY - Sex-determining region on the Y
SSCP - Single strand conformational polymorphism
TBE - Tris-borate EDTA
TDF - Testis determining factor
TE - Tris-EDTA
TEAA - Triethyl ammonium acetate
UK - United Kingdom
UV - Ultraviolet
VDA - Variant detector arrays
WMP - Western Malayo-Polynesian
3D - three-dimensional

Chapter 1: General introduction to the Human Y chromosome and human molecular evolution

Humans have always been interested in the origins of life. Genesis myths are found in every culture. Unsurprisingly this interest centres primarily on our own origins as a species, as individual races and as geographically structured populations. This hierarchy of interest has traditionally been studied by such disciplines as archaeology, linguistics and palaeontology, utilising records of our past found in our artefacts, our languages and in the remnants of our ancestors respectively (Cavalli-Sforza et al. 1994). Recently a new-found ability to decode a fourth record of our past has resulted in the addition of a new discipline to this trio (Stoneking 1993). This record is that of mutations in our DNA and the discipline is evolutionary genetics. No single discipline is sufficiently informative to deny the utility of the others. But rather used together, in a complementary fashion, we can maximise the resolution of our evolution (Lahr 1994). Colin Renfrew has dubbed archaeology, genetics and linguistics, "the three dimensions of human history".

Each of these independent records of our past has a number of different limitations with respect to the inferences that can be made from them. Whereas archaeology, through radio-isotope dating, can sharply define specific periods, the extant genetic diversity of human populations is a palimpsestual conflation of all previous population histories. In contrast to archaeologists and geneticists who are keen to make inferences over a time depth of hundreds of thousands of years, linguists are unwilling to extrapolate from their data further back in time than approximately 10,000 years (Renfrew 1994; Renfrew and Nettles 1999). Cultures and languages can be learnt and adopted whereas genes can only be inherited and accordingly analytical methods often differ significantly between these fields. Consequently combining these different records into a unified picture of human evolution is not straight forward and has encountered resistance within the different fields.

Before moving on to discuss the historical development of anthropological genetics I shall briefly consider the field of historical linguistics, for comparisons between genetic and linguistic records of our past will form a major part of this thesis.

An introduction to historical linguistics

Historical linguistics is the study of language change and how languages are related. The majority of linguists employ a technique known as the 'comparative method' to investigate language relationships. This involves the comparisons of certain features of language between two languages, and has resulted in a genealogical classification of languages into families and subfamilies (Renfrew 1994). Languages have a number of characteristics than can be used for comparison between them, including vocabulary, phonology, morphology and syntax (Renfrew 1994). For a grouping of languages to be well supported by the data, similarities between languages must occur at a frequency higher than that expected by chance. The comparative method also involves the construction of a 'proto-language' which represents the theoretical ancestral language to a group of extant languages. The subsequent evolution of the extant languages from the proto-language should make sense in the light of the known propensity for certain sound changes to occur (Renfrew 1994). Studying the processes of childhood language acquisition is analogous to investigating the mutational dynamics of a genetic locus and is also important in informing our understanding of how languages change (Don Ringe, personal communication).

Language classification has often taken an explicitly phylogenetic approach. Language families are thus organised into trees rather than simply into groups of related languages. The field of lexicostatistics has provided a quantitative backing for such phylogenetic analyses (Dyen et al. 1992). Lexicostatistics is the quantitative study of cognate words amongst vocabularies of a number of languages. One recent extension of lexicostatistics applied a phylogenetic approach known as network analysis to vocabulary words lists (Forster et al. 1998). It is claimed that networks are better able than trees to cope with the sometimes reticulate nature of language change.

Attempts have been made to produce a global phylogeny of languages comparable to that of populations attainable through genetics. In 1989 an attempt was made to compare qualitatively two such trees to show the correspondence between them (Cavalli-Sforza et al. 1989). Two substantial criticisms were levelled at this ambitious work; firstly linguists disagreed with many aspects of the language phylogeny, and secondly many others noted that there was no statistical support for what could easily have been a fortuitous arrangement of branches. Subsequently tree comparison metrics were successfully used to bolster the original conclusions against the second criticism by showing that the phylogenies were more similar than could be expected by chance (Penny et al. 1993).

A history of human molecular evolutionary studies

The record of the past that we can discern in our genome relies on two basic axioms; that we inherit our DNA solely from our ancestors and that over an evolutionary time scale this DNA accumulates mutations. Given these axioms we can not only assay diversity within and between species but also use the principle of genetic distance to determine the evolutionary relatedness of individuals and populations. This involves comparing homologous sequences in different individuals to assay their relatedness as a function of the genetic distance between them. Closely related individuals have fewer genetic differences within a given sequence than do more distantly related individuals.

The history of molecular studies into human origins can be told from the viewpoint of the molecular data and methodologies used, or from that of the anthropological questions addressed. Here I aim to summarise the former before considering the latter, emphasising the important role that mtDNA has played; subsequently I shall consider the Y chromosome in more detail.

Anthropological genetics falls into a wider field of human molecular evolution. DNA is not the only macromolecule used to analyse human evolution. Initial attempts to assay diversity of humans and other species could only focus on diversity at the protein level (reviewed by (Avise 1994). Immunological methods were first used to detect specific protein alleles (e.g. the ABO blood groupings). Indeed the first geographic study of molecular diversity is purported to be an investigation of blood group frequencies within Europe in 1940 (Haldane 1940). Immunological measures of distance between homologous proteins allowed the first attempts to quantify relationships between species at the molecular level (reviewed by (Avise 1994). The later development of electrophoretic separation of proteins allowed multi-locus investigation of intra-specific diversity and revealed surprisingly high diversity within humans (reviewed by (Avise 1994). Protein data can only assay the coding regions of the genome and only those changes that result in coding differences. The movement towards DNA-based data was initiated by the whole genome comparisons of DNA-DNA hybridisation experiments (Avise 1994). The subsequent discovery of restriction enzymes and their sequence-specific activity allowed sequence variation to be assayed directly at the DNA level for the first time, by Restriction Fragment Length Polymorphism (RFLP). It is only relatively recently that screening for new DNA sequence polymorphism using restriction enzymes has been superseded by direct sequence determination (Cargill et al. 1999). RFLP assays still have a role to play in typing known polymorphisms (Mathias et al. 1994; Torroni et al. 1994).

Allied to these advances in practical methodology was the theoretical work of population genetics, the founding works of which were published well before the data were available to test their hypotheses. Fisher, Wright and Haldane published in the 1930's work which still guides theoretical analyses today (Fisher 1930; Wright 1931; Haldane 1932). Their population models all relate to the distribution of allele frequencies amongst populations. Consequently populations became the *de facto* unit of investigation. Genetic distance measures were devised that combined information from multiple alleles in multiple samples to express the genetic relatedness of pairs of populations: the greater the distance measure, the more distantly related the populations (Nei 1987).

The realisation that not all polymorphisms have a selective effect was critical to the development of the field and underpins the use of molecular diversity to make population historical inferences. The resulting neutral theory of evolution further delineated the factors underpinning temporal changes in neutral allele frequencies (Kimura 1983).

Only once diversity could be assayed at the sequence level was it that the additional information of mutational distance between alleles at the same locus could be analysed. This innovation required that new analytical methods be developed to take account of this additional information. It became no longer sufficient to consider allele frequencies alone (Avise 1989). It was quickly realised that if recombination could be excluded, alleles could be combined into haplotypes and used to construct gene genealogies, thus introducing a new temporal aspect to allelic variation (Avise 1989). In the past few years a whole host of theoretical innovations have sprung up to take advantage of this extra molecular information (Barbujani 1999). I shall argue in this thesis that this approach requires that the lineage, not the population, is adopted as the unit of investigation.

The advantages of non-recombining regions

The majority of the human genome undergoes recombination at meiosis. This shuffling process disrupts the integrity of inherited sequences such that in a given generation of ancestors there are multiple contributors to any sequence of appreciable length in the descendant. The further back in time we look, the shorter the sequence that has maintained its integrity (i.e. can be traced back to a single ancestor). Thus we cease to be able to combine discrete polymorphisms into informative haplotypes and can only relate populations by virtue of variation in allele frequency at individual loci, an intrinsically less powerful analysis (although statistical treatments exist that can

combine allele frequency information from multiple unlinked loci (Womble 1951; Sokal et al. 1992). Consequently regions of the human genome that can exhibit complete linkage can be considered as being simpler, more informative records of the past than regions that undergo extensive recombination (Jobling and Tyler-Smith 1995).

There are two distinct regions of the genome that by their haploid nature escape recombination at meiosis, namely, the sex-determining Y chromosome and mitochondrial DNA (mtDNA). It is also possible to combine polymorphisms on autosomal chromosomes into informative compound haplotypes by isolating a region exhibiting total linkage by selecting a region of low recombination and carefully excluding from the analysis those haplotypes that result from rare recombination events (Tishkoff et al. 1996; Harding et al. 1997).

The principle of coalescence

Coalescence describes the common *a priori* characteristic of all non-recombining regions in that they all have single common ancestors. This is explained below in the case of the Y chromosome but is equally true for any non-recombining locus (Jobling and Tyler-Smith 1995; Templeton 1997).

The Y chromosome functions to determine maleness and as such is inherited uniparentally from father to son. All sons have fathers but do not necessarily go on to have sons themselves. This simple fact results in Y-chromosomal lineages dying out over time. This is illustrated in figure 1.1, for a population of constant size. As not all lineages in a given generation are represented in the next we can see that in preceding generations fewer and fewer Y chromosomes are ancestral to the extant Y chromosomes. Many are ancestral to lineages that died out before the present generation. Thus if we trace all extant Y-chromosomal lineages back through time we find them coalescing into fewer and fewer lineages until at some point in the past we can trace all extant human Y chromosomes back to a single common ancestor.

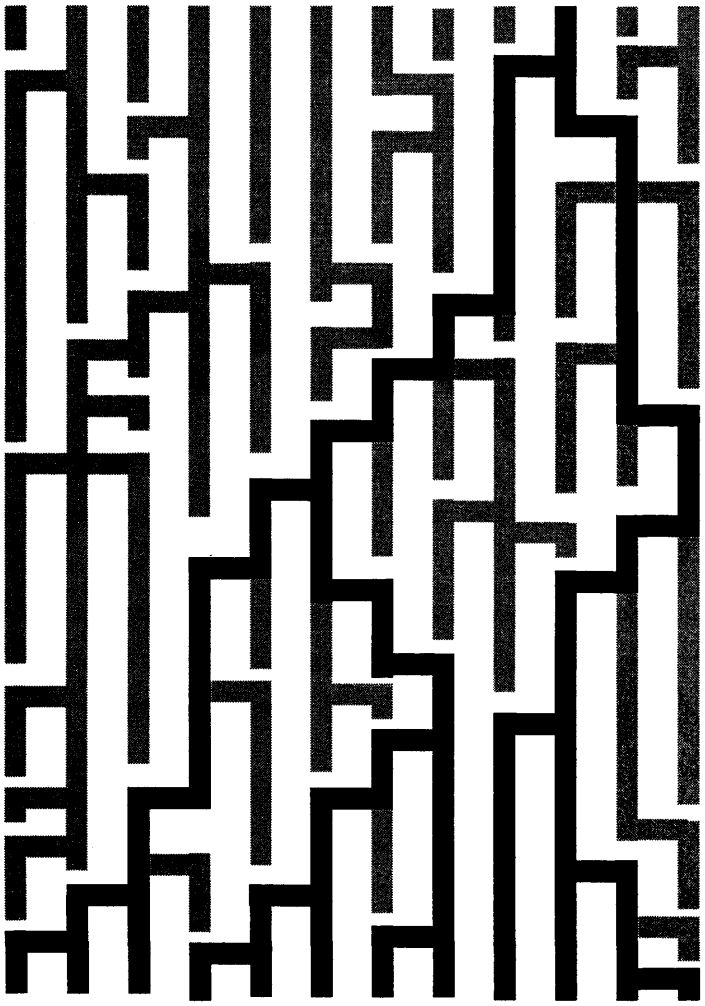


Figure 1.1: Coalescence of non-recombining lineages. Lineage extinction in a constant population of 12 chromosomes results in a recent common ancestor.

We can say nothing *a priori* about the date, place or indeed species in which this common ancestor existed. However under neutral expectations we expect that the time taken to reach a common ancestor is dependent on the effective population size. The larger the population size the longer it takes for lineages to coalesce back to a single common ancestor. This common ancestors of extant Y chromosomes and mtDNA have been dubbed 'Adam' and 'Eve' respectively, in reference to the biblical characters.

Mitochondrial DNA

The mitochondrion is an organelle involved in cellular energy production and is found in multiple copies within the cell. It contains within it several copies of a gene-rich circular DNA molecule, 16.5kb in size. Like the Y chromosome, mtDNA escapes recombination at meiosis as a result of its being haploid and it also has a uniparental pattern of inheritance (reviewed in (Stoneking and Soodyall 1996)). Many people think that the portion of a sperm that enters the ovum on fertilisation does not contribute any mitochondria to the resulting zygote. Consequently mtDNA is passed down through maternal lineages alone. The principle of coalescence also extends to mtDNA, resulting in a common maternal ancestor of all modern mtDNA, misleadingly titled 'Eve'. This has led to confusion in reconciling the fact that mitochondrial 'Eve' (and Y-chromosomal Adam) is only the common ancestor of our mtDNA (Y chromosome) and had to coexist within a much larger population containing ancestors contributing to all other human nuclear loci, with the Biblical interpretation of Adam and Eve as the sole male and female ancestors of us all. Indeed it is highly unlikely that our genetic 'Adam' and 'Eve' ever met.

MtDNA has a mutation rate generally considered to be ten times that found at nuclear loci, resulting in a wealth of easily-assayed polymorphisms. The non-coding D-loop mutates at a rate even higher than the rest of this genome (Parsons et al. 1997). The D-loop is divided into two hypervariable sequences (HVS), named HVS-I and HVS-II. The cellular abundance of mtDNA, in conjunction with this high mutability, has meant that of the two haploid non-recombining loci in the human genome, mtDNA analysis has advanced ahead of Y-chromosomal analysis.

Certain problems exist with mtDNA as an evolutionary genetic tool (Stoneking and Soodyall 1996). Firstly, there is a limited range of polymorphism types available for analysis in mtDNA: base substitutions are commonly typed along with a few insertion/deletion events. An ideal locus for human evolutionary work would contain a range of different polymorphism types that have differing mutation rates, so that discrimination can be achieved at the level of both populations (where slowly mutating markers are more informative) and individuals (where faster mutating markers are more informative). Secondly, the high base substitution mutation rate means that certain mutations have occurred independently multiple times which hampers attempts to draw a single most parsimonious tree (Bandelt et al. 1995). Thirdly, due to the high mutation rate, chimpanzee mtDNA has diverged from human mtDNA to such an extent that determination of ancestral state is problematic. Finally, heteroplasmy of mtDNA may cause problems (see Ivanov et al. 1996). More recently it has been suggested that recombination between maternal and paternal

mtDNA may also account for the high rate of homoplasy within mtDNA phylogenies (Hagelberg et al. 1999; EyreWalker et al. 1999).

MtDNA has been used extensively in ecological genetics (Avise 1994). Extensive conservation of sequences across many species allows PCR to give immediate access to molecular diversity in species that have been poorly characterised genetically. In addition mtDNA diversity, in combination with nuclear data, allows investigations into the genetic impact of the different mating structures and novel behaviours that litter the field of molecular ecology (Avise 1994).

Man's place amongst the primates

Traditional morphological studies prior to Darwin placed man within the primates. It was thought that the Old World Apes were our closest cousins, but the common consensus was that the split between us and them occurred tens of millions of years ago. In 1967 Sarich and Wilson used distances from immunological comparisons of homologous proteins from different primates to show that chimpanzees, gorillas and man are related roughly equally to one another (Sarich and Wilson 1967). Under the assumption that the mutation rate of a protein does not vary between different primate species and using dated splits in primate phylogeny from the fossil record they calibrated the age of this split to only 5-7 million years. Other subsequent molecular comparisons reinforced this recent common ancestry. On the weight of current evidence chimpanzees rather than gorillas would seem to be our closest primate relative.

Recent work has challenged this age for the common ancestor of humans and chimps. More accurately determined splits within mammalian phylogeny have been used to recalculate the age of the human/chimp split to between 10 and 13 million years (Arnason et al. 1998). Much calibration of intra-specific ages is done relative to this human/chimp split therefore it seems surprising that people have been unwilling to adopt this new, theoretically more accurate, calibration.

The origins of modern humans

One of the most obvious questions to address with molecular data is "where do humans come from?". This layman's question can quickly be refined to the more testable "How did anatomically modern humans (AMH) evolve?". Answers to this question have polarised into two

major hypotheses (Lahr 1994). Firstly, that within the past 200,000 years AMH arose in Africa and subsequently migrated out of Africa replacing the pre-existing *Homo erectus* populations, the so-called 'Out of Africa' hypothesis. It is thought from palaeontological and archaeological data that *Homo erectus* came out of Africa around a million years ago and spread across the Old World, as far as some of the Indonesian islands. Alternatively it is thought that indigenous transformation of archaic hominid groups to AMH occurred with sufficient gene flow between them to maintain species integrity, the so-called 'multi-regional hypothesis' (reviewed in (Lahr 1994). These two explanations are summarised in figure 1.2. This debate began amonst palaeontologists over the question of whether regional characteristics of *Homo erectus* cranial morphology could be seen to be perpetuated within AMH from the same regions.

Many global studies of different genetic loci have addressed this question. Most, but not all (see (Harding et al. 1997)), of these studies have claimed to support the 'Out of Africa' hypothesis, with evidence coming from mitochondrial DNA (mtDNA) being widely cited (Cann et al. 1987). Trees of mtDNA diversity have been used to conclude that an ancestor of all extant mtDNA existed in Africa within the past 200,000 years (Stoneking and Soodyall 1996). However it has been argued that despite their compatibility with a recent African origin for AMH, these conclusions are not capable of distinguishing between the two hypotheses (see below; Templeton 1993; Templeton 1997).

It is a common finding at many different loci that diversity within African populations is greater than outside of Africa, and although this result is compatible with the 'Out of Africa' hypothesis, it does not exclude the 'multi-regional' hypothesis, as the effects of differential demographic histories on extant diversity are well known (Templeton 1997). Greater diversity may be accounted for by a larger long term effective population size. The determination of an ancient Neanderthal mtDNA sequence markedly different from any known human sequence has provided the best support for their being no genetic input from archaic human populations to extant human diversity (Krings et al. 1997). However arguments can be marshalled against even this evidence (Templeton 1997). In conclusion the genetic support for the 'Out of Africa' hypothesis comes more from the weight of compatible evidence than the absolute exclusion of the alternative hypothesis.

Whilst this is possibly the single most important question in the field of human evolution as a whole, two factors suggest that studies of extant Y-chromosomal diversity, like analogous mtDNA studies, will resolve the debate little, if at all.

(1) Neutral theory predicts that from what we know about our demographic history, in terms of the likely constant effective population size, genetic drift alone would result in both the Y chromosome and mtDNA coalescing to a common ancestor about 200,000 years ago, independent of the true picture of human origins (Templeton 1997). Thus both the Y chromosome and mtDNA can only be used to address questions of human evolution within the last 200,000 years.

(2) The Y chromosome and mtDNA only represent single loci. The discrepancy that can exist between gene trees and species trees is a well documented one (Avise 1994). To get an accurate picture of our true species tree we need to sum the evidence from as many different loci as possible.

The best way to resolve this debate from the point of view of genetics is to construct gene trees based on compound haplotypes for multiple autosomal loci that comprise a small number of polymorphisms within a few kilobases that exhibit complete linkage and analysing them to determine the date and place of the common ancestor of each locus (Tishkoff et al. 1996; Harding et al. 1997; Jin et al. 1999; Kaessmann et al. 1999).

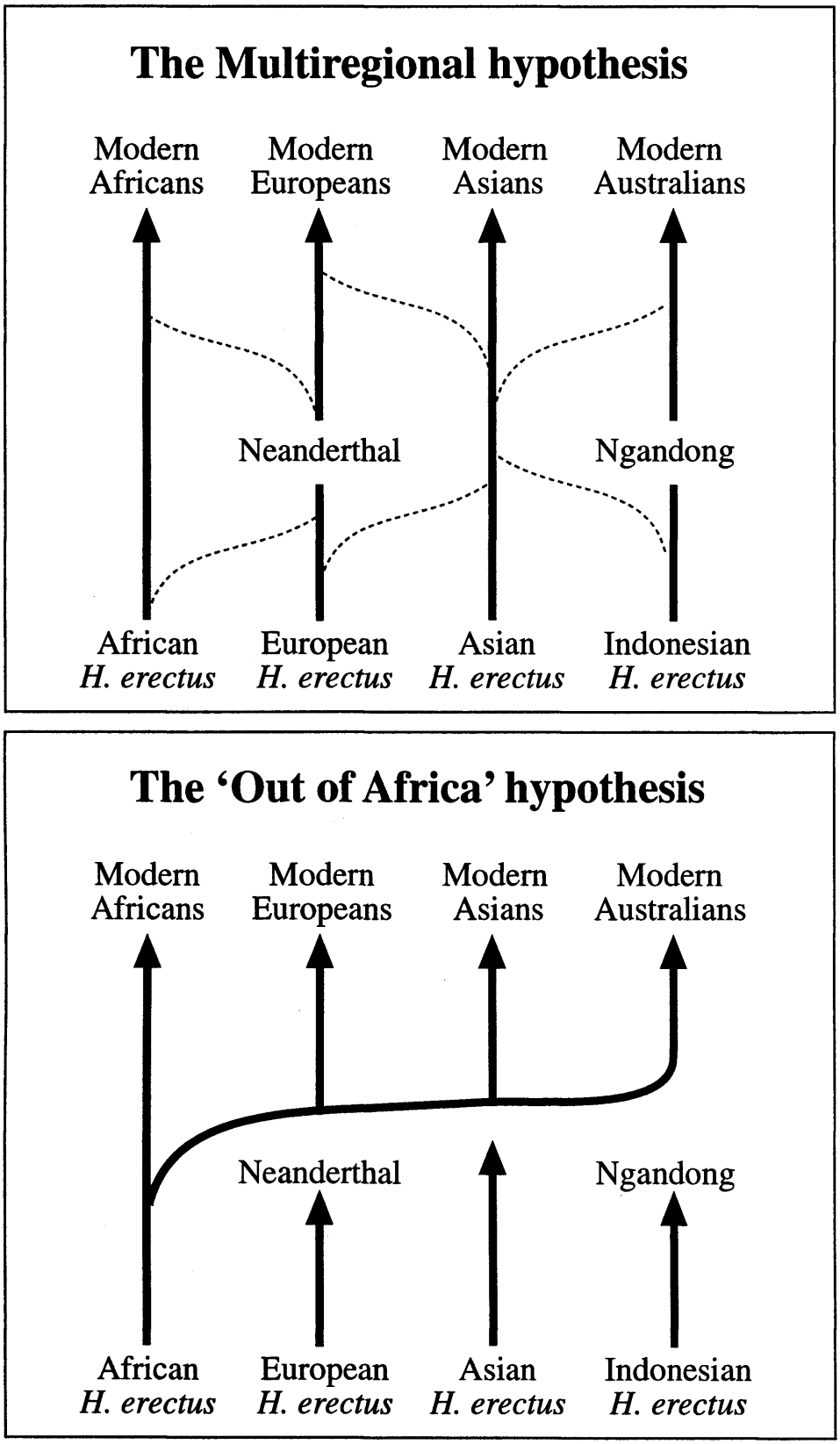


Figure 1.2: Two competing theories for the origins of anatomically modern humans, dotted lines indicate gene flow between different regional populations.

Regional studies of diversity

As well as studies of global diversity, molecular information has also been used to address questions of regional prehistory. Human prehistory has been investigated at many different scales from the continent-wide (Richards et al. 1996) to the country-specific (Barbujani and Sokal 1991). Much was accomplished with data from allele frequencies, and large databases covering all regions of the world were accumulated (Sokal et al. 1991; Cavalli-Sforza et al. 1994). New analyses were developed to investigate the spatial distribution of this diversity and its relationship to the other records of the human past (reviewed by (Barbujani 1999)). The culmination of much of this work was the extensive work by Cavalli-Sforza et al. published in 1994 (Cavalli-Sforza et al. 1994). More recently however attention has focused on those non-recombining regions of the genome that can supply extra inferential information in the form of phylogenies (Avise 1989). Attention has focused on issues of prior anthropological interest. Major areas of focus have included the peopling of the Americas (Forster et al. 1996; Bianchi et al. 1997), the colonisation of the Pacific (Spurdle et al. 1994; Sykes et al. 1995) and the impact of agriculture in Europe (Ammerman and Cavalli-Sforza 1984; Richards et al. 1996).

Before considering the evolutionary applications of the human Y chromosome some background to this unique locus must be introduced.

Structure of the Y chromosome

The human Y chromosome is about 60Mb in length and, as with other mammalian Y chromosomes, functions to determine maleness (reviewed by (Goodfellow and Lovell-Badge 1993). The viability of 45,XO individuals indicates the likely lack of other major non-male-specific functional roles for the Y chromosome. The distal part of the long arm of the chromosome is heterochromatic, consisting of highly repeated sequences, and is polymorphic in length with an average size of 30Mb (Verma et al. 1978). The rest of the long arm and the whole of the short arm is euchromatic and contains all the known functional genes on the Y chromosome. A substantial portion of the euchromatic portion of the Y chromosome also consists of repeated sequences; both multiple copies of functional genes and non-coding repeat sequences. It also includes two large

regions of homology with the X chromosome (reviewed by (Affara and Lau 1994) and (Vogt et al. 1997).

The Y chromosome divides genetically into three distinct regions, namely the two short pseudoautosomal regions (PARI and PARII) at either end of the chromosome and the intervening Y-specific region. PARI is 2.6Mb in length (Brown 1988), at the distal end of Yp, and is the site at which pairing between the X and Y chromosomes is initiated during meiosis (Chandley et al. 1984). Recombination between PARI and its corresponding region on the X chromosome (distal Xp) is obligatory during the successful segregation that is required during meiosis to ensure there is only one sex chromosome per gamete. PARII is smaller, only 320kb in length and situated at the distal end of Yq (Kvaløy et al. 1994). Recombination at PARII is not obligatory for proper meiotic segregation and chiasmata are rarely formed here during male meioses (Li and Hamer 1995). An ideogram of the Y chromosome is shown in figure 1.3.

It is the functional role of the Y chromosome in sex determination which results not only in its being haploid and thereby escaping recombination but also its unique pattern of inheritance. The Y chromosome is passed down through the generations solely from father to son.

Evolution of the Y chromosome

Chromosomal systems of sex-determination have evolved independently multiple times in evolution, though many operate by a dosage-related system and not via a dominant haploid chromosome. A common evolutionary path has been proposed for such heterogametic systems whereby the two sex chromosomes diverge from an ancestral, homologous, pair of autosomes (Ohno 1967). The origins of the human Y chromosome can be traced back to the ancestor of all mammals, as mammals generally share the same X-Y heterogametic sex-determination system (Marshall Graves 1995).

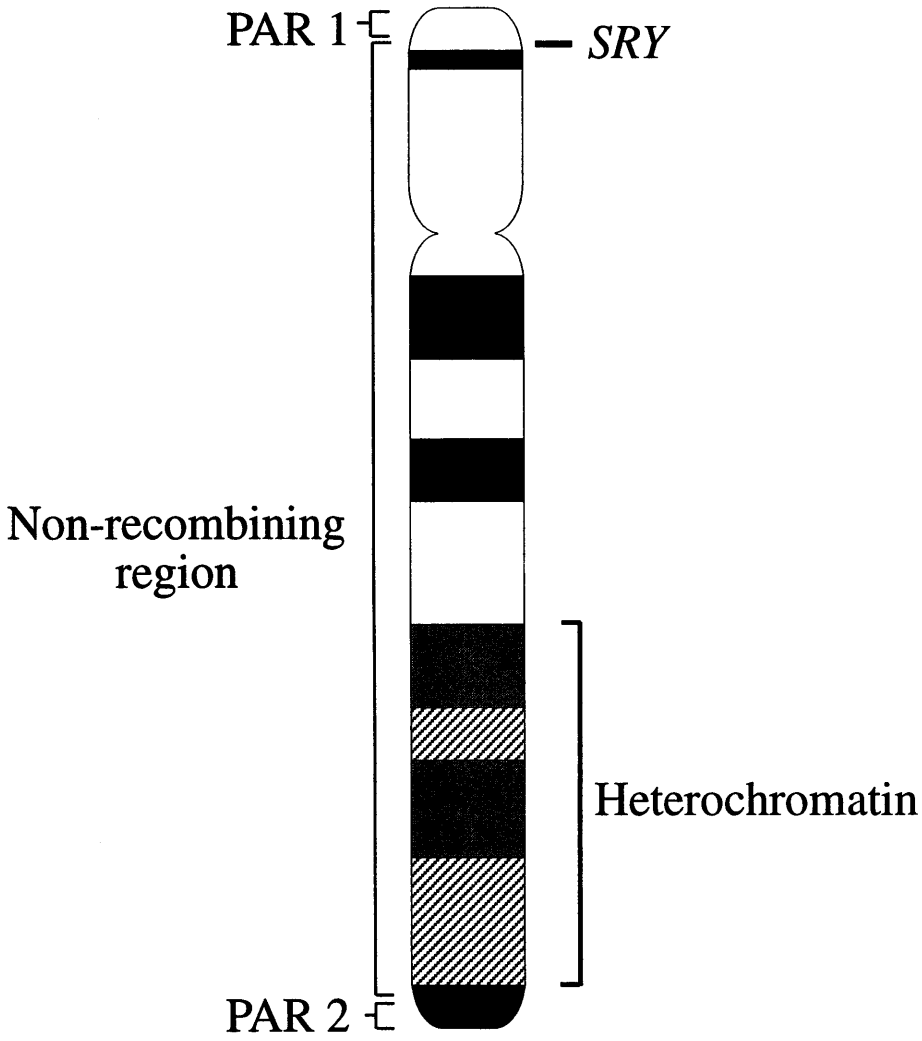


Figure 1.3: Ideogram of the Y chromosome. Showing the locations of the pseudoautosomal regions, the testis determining gene, SRY, and the long arm heterochromatin.

The haploid chromosome of a chromosomal sex-determining system is exposed to the effects of Muller's Ratchet, the stochastic elimination of Y chromosomes containing the fewest deleterious mutations, resulting in a net accumulation of these mutations. A prognosis of perpetual attrition is predicted by this theory (Jegalian and Page 1998). Accordingly the Y chromosome contains many degraded copies of X-linked genes. To maintain similar expression of these genes between the sexes, one of the X-linked copies of such genes undergoes inactivation and is thus transcriptionally silent. However not all Y copies of X-Y homologous genes do undergo degradation. Obviously for such genes two copies are a necessity. Correspondingly the X-chromosome homologues of these genes are not inactivated (Jegalian and Page 1998).

A second process has dominated mammalian Y chromosome evolution, namely the acquisition of spermatogenic genes. It is thought that this results from the fact that such genes can enhance male reproductive fitness whilst being neutral or even detrimental to females (Lahn and Page 1997). In this situation selective pressures would favour the accumulation of such genes in male-specific regions of the genome. Alternatively it has been suggested that the presence of a gene on the Y chromosome predisposes it to acquiring a male-specific function and that this alone can account for the presence on the Y chromosome of multiple spermatogenic genes (Marshall Graves 1995). Under the first hypothesis no X-linked homologues of Y-linked spermatogenic genes are expected, whereas under the latter hypothesis we should expect all Y-linked genes to have X homologues. In actuality some, but not all, genes performing male-specific functions have X-linked homologues (Lahn and Page 1997). This might indicate that both mechanisms have operated during the evolution of mammalian Y chromosomes.

Genes on the Y chromosome

In the last ten years the Testis Determining Factor (*TDF*), the gene product of which directs the indifferent gonads to form testes during male development, has been localised through mapping experiments in XX males to a specific gene on the Y specific portion of the Y chromosome, *SRY* (Sinclair et al. 1990). The *SRY* gene was confirmed as the *TDF* by the finding of *de novo* mutations in XY females (Berta et al. 1990; Jager et al. 1992) and by sex-reversal of a female mouse transgenic for a 14kb segment of DNA containing the murine *SRY* homologue. The *SRY* gene has been placed 5kb from the pseudoautosomal boundary between PARI and the Y-specific portion of the Y chromosome (Sinclair et al. 1990).

PARI is relatively gene rich containing at least nine genes including genes coding for metabolic functions: *ASMT* (Yi et al. 1993), *ASMTL* (Ried et al. 1998) & *ANT3* (Schiebel et al. 1993); genes coding for cytokine receptors: *IL3RA* (Milatovich et al. 1993) & *CSF2RA* (Gough et al. 1990); genes coding for cell surface antigens: *PBDGX* (Ellis et al. 1994) & *MIC2* (Smith and Goodfellow 1994); a homeobox gene involved in stature: *SHOX* (Rao et al. 1997); and a gene of unknown function: *XE7* (Ellison et al. 1993). The X homologues of these genes, as with all X homologues of Y-linked genes, have been shown to escape X inactivation. PARII is known to contain a least two genes, a cytokine receptor gene: *IL9RA* (Kermouni et al. 1995) and a synaptobrevin-like gene: *SYBL1* (D'Esposito et al. 1996).

Until recently there were 10 known genes on the Y-specific portion of the Y chromosome. However a recent publication has expanded this number to 20 genes and gene families, many of which are thought to have a role in spermatogenesis (Lahn and Page 1997). Some of these spermatogenic genes, such as *RBM* (Ma et al. 1993) *DAZ* (Saxena et al. 1996) and *TSPY* (Manz et al. 1993), are present in multiple copies dispersed along the chromosome. It is not known how many copies are functional, or even whether the presence of multiple copies has a functional role. Other spermatogenic genes have only been found in single copies. Single copy spermatogenic genes tend to have X-linked homologues whereas the repeated genes do not. This might be explained by the former class of gene acquiring a male-specific function *in situ* whereas the latter class of gene was acquired by the Y chromosome by virtue of its pre-existing male reproductive function and then subsequently underwent multiple rounds of duplication and divergence. Non-spermatogenic genes include the ribonuclear protein gene *RPS4Y* (Fisher et al. 1990) and the initiation factor *EIF1AY* (Lahn and Page 1997), which are expressed in multiple tissues and are thought to have vital house-keeping functions, explaining the maintenance of the Y-linked copy.

In the rest of this thesis the phrase 'Y chromosome' is used as short hand for what is more properly known as the Y-specific region of the Y chromosome (i.e. excluding PARI and II).

Genetic variation on the Y chromosome

In contrast to mtDNA, the Y chromosome should have a mutation rate similar to that of other nuclear loci as it is subject to the same repair processes, although it never passes through female meioses. In addition it should contain the same types of polymorphic loci found on the autosomes.

Unique, biallelic polymorphisms

Some biallelic polymorphisms on the Y chromosome, such as base substitutions and certain insertion/deletion events, are commonly thought to represent unique events in human evolution. Although base substitutions occur at low rates in the nuclear genome, if we consider the number of extant human genomes in relation to the known mutation rate of these polymorphisms we see that recurrent mutation is bound to have occurred in some individuals. However any sampling of the

world's populations has such a low density of coverage that these recurrent mutations are very unlikely to be picked up, and thus will have a negligible effect on the analysis. In effect, they can indeed be considered to represent unique events.

Early surveys of diversity on the Y chromosome found little polymorphism. These surveys searched for Restriction Fragment Length Polymorphisms (RFLPs) by digesting human DNA with a panel of restriction enzymes and probing with single and low-copy Y-specific probes (Jakubiczka et al. 1990; Malaspina et al. 1990; Spurdle and Jenkins 1992; Jobling 1994). More recently studies have focused on sequencing the same region in a number of individuals from diverse populations. One study found no variation in 729bp sequenced in 38 individuals (Dorit et al. 1995), another found three base substitutions when sequencing 18.3kb in five individuals (Whitfield et al. 1995) whilst a third found three base substitutions when sequencing 2.6kb in 16 individuals (Hammer 1995). The most recent screen for Y-chromosome sequence polymorphism identified 19 new polymorphisms in a screen of roughly 19kb in 21-53 individuals (Underhill et al. 1997). This level of diversity is low compared to that found on the autosomes of one base substitution every 235bp for 12-20 chromosomes (Nickerson et al. 1992) and can be explained in a number of ways. Firstly it raised the possibility of a recent selective sweep on the Y chromosome. This entails a recent fixation of an advantageous allele of a Y-chromosomal gene resulting in the fixation of a specific allele for every polymorphic locus in complete linkage with it. An alternative explanation for this apparent lack of diversity is that the population size of Y chromosomes is a quarter that of any autosome and one third that of the X chromosome. Neutral theory predicts the diversity apparent in the Y chromosome to be similarly reduced, purely as a function of this reduced population size. A further explanation for reduced Y chromosome diversity is that cultural mating practices have in the past resulted in certain males being over-represented in the next generation. Other explanations rely on a lack of recombination causing a reduction in mutation rate on the Y chromosome although this is not borne out by the limited available data on mutation rates of Y-linked polymorphisms (Underhill et al. 1997).

Biallelic polymorphisms of unique origin on the Y chromosome can be combined into monophyletic compound haplotypes, haplogroups, which are related to one another phylogenetically by a single most parsimonious tree (Mathias et al. 1994). It is simple to construct this tree by hand from a character state table. Ancestral state can be determined by looking at the homologous region in an appropriate outgroup, the chimpanzee. If the ancestral state of all polymorphisms used to construct the tree is known, the tree can be rooted. Dating the branchpoints

of the tree is of prime importance in relating tree topology to population movements in prehistory. It can also be inferred indirectly by comparing human sequence diversity to that seen between humans and apes and using estimates of the date of the human-ape divergence to calibrate the measurement. This second approach has been used in five recent papers which give non-overlapping 95% confidence intervals for the age of the Y chromosome common ancestor of 0-800,000 years (Dorit et al., 1995), 51,000 to 411,000 years (Hammer, 1995), 39,000 to 47,000 years (Whitfield et al., 1995) and 69,000-372,000 years (Underhill et al., 1997). Four trees based on these biallelic polymorphisms, as they were presented in their different publications, are shown in figure 1.4. Note that each study used a different set of markers and that there are a number of different nomenclatures. In addition the trees look very different, due to the manner in which they were originally drawn and whether they are rooted or not. Figure 1.5 shows the unrooted tree of haplogroups on which much of this thesis is based. This tree comprises some markers that can be typed by PCR and others that can not, and was constructed mainly by Arpita Pandya and Chris Tyler-Smith with help from Fabricio Santos, Mark Jobling and myself.

Microsatellites

Useful measures of diversity require markers that are polymorphic in all populations. Unique, biallelic polymorphisms on the Y chromosome are relatively population-specific (e.g.Seielstad et al. 1994; Hammer et al. 1997). In addition, with the presently available panel of such markers, it is often difficult to discriminate between two closely related populations. Microsatellites are tandem repeats of 2-6 bp that are widespread in the human nuclear genome. They are highly variable in allele length, are polymorphic in all populations and are notably used in linkage analysis (Hearne et al. 1992) and identification of individuals (Urquhart et al. 1995). Recently their utility in evolutionary studies has been investigated, initially on autosomes (Bowcock et al. 1994).

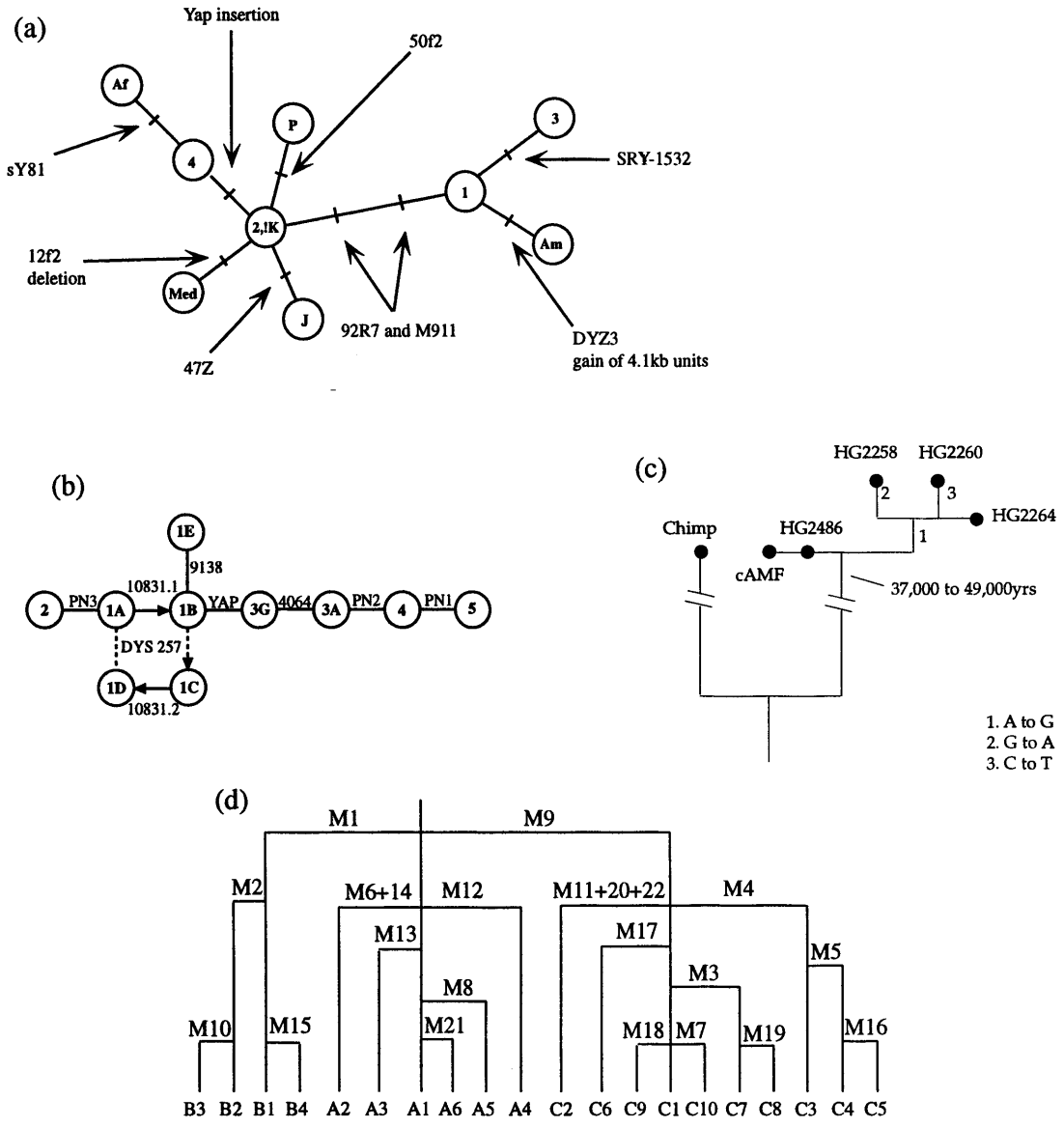


Figure 1.4: Four trees constructed from unique, biallelic polymorphisms. Tree **a** is from (Jobling and Tyler-Smith, 1995), is unrooted due to uncertainty about ancestral states and uses nine unique polymorphisms to define nine compound haplotypes (haplogroups) represented by circles. Tree **b** is from (Hammer et al., 1998) and uses seven base substitutions and an *Alu* insertion to define ten compound haplotypes in an unrooted tree. Tree **c** is from (Whitfield et al., 1995) and uses three base substitutions to define 4 compound haplotypes in a rooted tree with chimps shown as an outgroup. Tree **d** is from (Underhill et al., 1997) and uses twenty-two markers to define 19 lineages in a rooted tree.

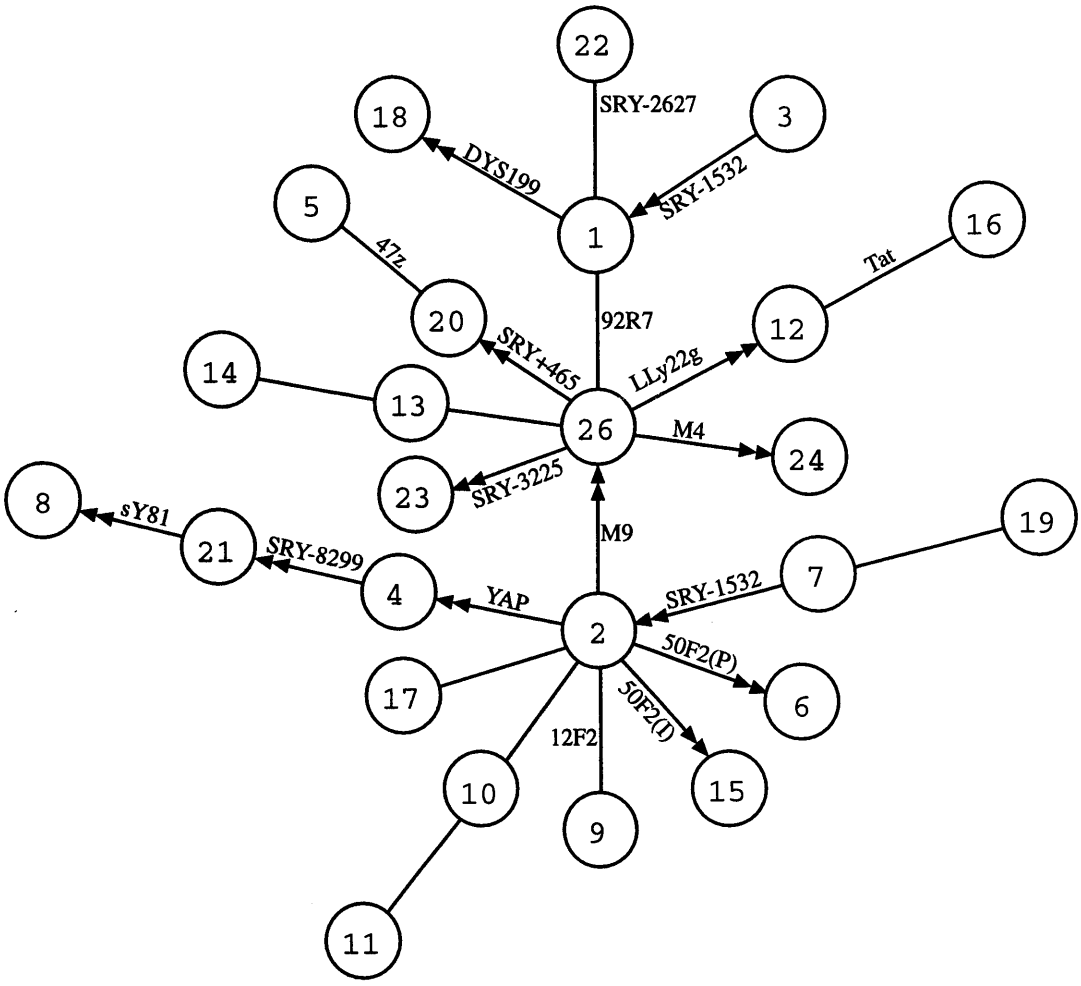


Figure 1.5: An unrooted tree of haplogroups developed from that in figure 1.4a. Circles represent haplogroups and lines individual mutational steps between them. Arrows indicate ancestral state, where known. Mutational steps are named where they can be typed by PCR

Microsatellites on the Y chromosome show roughly equivalent diversity to autosomal loci (Roewer et al. 1992; Goldstein et al. 1996). Because they are not independently assorted by recombination, Y-linked microsatellite allele lengths show greater population structure than do their autosomal counterparts (Jobling et al. 1997). As with unique, biallelic polymorphisms, alleles at multiple microsatellite loci on the Y chromosome can be combined into compound haplotypes. However, because of the likely nature of their mutational mechanism (slippage), recurrent mutation is common (Ciminelli et al. 1995; Cooper et al. 1996). This leads to the construction of very many equally parsimonious trees (Roewer et al. 1996), that are of limited phylogenetic use due to the

inability to discriminate between those linkages due to recurrent mutation (Identity By State - IBS) and those resulting from single mutational events from ancestral haplotypes (Identity By Descent - IBD).

There are at least 34 known microsatellite loci on the Y chromosome (Jobling et al., 1997; White et al. 1999), C. Tyler-Smith personal communication). Of these, seven loci have been widely used for both evolutionary and forensic purposes by virtue of their high levels of diversity and unambiguous allele designation (Roewer et al. 1996; Kayser et al. 1997).

Mutation rates for autosomal microsatellites have been estimated from pedigree analysis. One study gives a mutation rate of $2x10^{-3}$ for tetranucleotide repeats (Weber and Wong 1993). Preliminary data on mutation rates of Y-chromosomal microsatellites come from deep-rooting pedigrees and indicate rates of a similar magnitude to those found at autosomal microsatellites (Heyer et al. 1997). Knowledge of microsatellite mutation rates is of potential utility in dating the branchpoints found in trees defined by unique, biallelic polymorphisms (Zerjal et al. 1997). A haplogroup is founded by a single male with zero diversity. We therefore expect that recurrent polymorphisms will retain some phylogenetic information and can anticipate that haplotypes of recurrent polymorphisms will show some statistical correspondance with the tree of haplogroups.

Different measures of genetic distance have been developed to cope with the use of microsatellites as tools in human evolutionary genetics (Goldstein et al. 1995; Zhivotovsky and Feldman 1995) and it is often found that the picture obtained differs significantly depending on the measure of diversity used (Perez-Lezaun et al. 1997).

Minisatellites

Minisatellites are tandem repeats of especially GC-rich 10-60bp units that are preferentially located at the distal ends of chromosomes (Armour and Jeffreys 1992). Their utility for evolutionary work is limited if allele length is the only polymorphic quality assayed because of high length homoplasy. An additional source of polymorphic information content is the sequence variation of repeat units within the array. The positions of these variant repeats can be mapped using the technique of Minisatellite Variant Repeat (MVR) PCR (Jeffreys et al. 1991). This technique has been used to good effect for the evolutionary study of an autosomal minisatellite

(Armour et al. 1996), as well as for the study of mutational processes operating at these loci (Jeffreys et al. 1995).

The only known minisatellite on the Y chromosome is MSY1 (*DYF155S1*). Unusually it is AT-rich, and comprises 48-114 repeats of a 25bp repeat (Jobling et al. 1994; Jobling et al. 1998). The vast majority of repeats are of one of three types, which tend to occur in blocks within the array. The order of the blocks along the array constitutes its modular structure, for example the MSY1 allele shown in figure 1.5 has the modular structure (3,1,3,4). MSY1 exhibits extreme diversity, with a mutation rate calculated from diversity measures, of 2-11% and directly estimated from pedigree studies of ~6% per generation (Paul Taylor, personal communication). MSY1 codes of repeat types fall into distinct subtypes defined by different modular structures which, as expected, correlate reasonably well with haplogroups defined by biallelic polymorphisms (Bouzekri et al. 1998; Jobling et al. 1998) . Whilst individual block sizes fluctuate rapidly in size, similar to microsatellites, modular structures are much more stable. Once more is known about the mutation mechanisms operating at this locus it has potential utility as a measure of diversity within haplogroups defined by unique, biallelic polymorphisms and may allow better dating of tree branchpoints than that achievable using microsatellites. Interestingly the MSY1 repeat sequence is predicted to form an almost perfect hairpin structures (Jobling et al. 1998).

MVR-PCR typing of MSY1

The development of an MVR-PCR system for typing MSY1 was essential to unlocking the polymorphic information contained within it. Figure 1.6 shows schematically the principle of MVR-PCR: single radioactively-labelled discriminator primers are used in each reaction. Codes from MVR-PCR reveal MSY1 to be the most diverse polymorphic locus on the Y chromosome with a vanishingly small chance of picking two identical codes at random (Jobling et al. 1998). Given the complexity of the mutational processes operating on this locus it is not easy to envisage how to construct a tree from MSY1 data alone. It is likely that such a tree would involve multiple reticulations given the microsatellite-like evolution of block sizes. In addition it is known that many of the modular structures are recurrent, occurring in multiple lineages defined by unique markers. Consequently trees constructed on the basis of modular structures alone would involve considerable confusion of IBD and IBS.

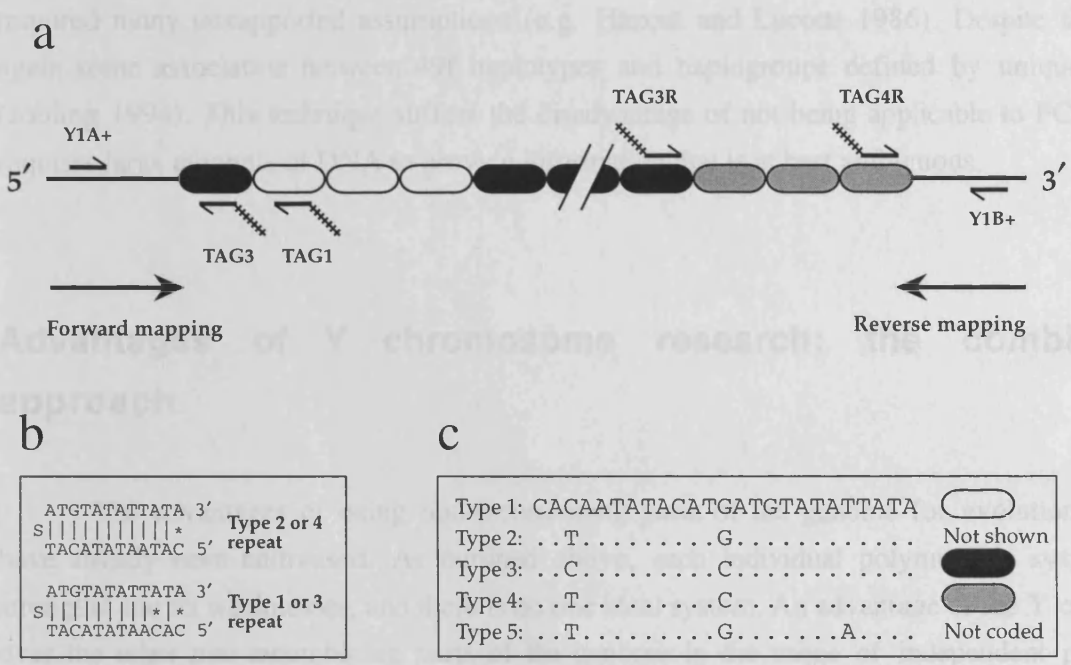


Figure 1.6: Schematic diagram of the MVR-PCR system on MSY1, sequence and structure of variant repeats. (a) A schematic allele of MSY1 is shown which contains type 1,3 and 4 repeats which are coded in forward (repeat types 1 and 3) and reverse (repeat types 3 and 4). (b) Predicted hairpin structure of MSY1 repeats. (c) Five variant MSY1 repeats identified by sequencing.

Additional polymorphic systems on the Y chromosome

The alphoid heteroduplex polymorphic system has some features of both recurrent and unique mutations (Santos et al. 1995). Co-amplification of variant alphoid satellite units of 281-285bp gives a pattern of multiple heteroduplexes that define at least 32 different haplotypes. Although simple patterns are likely to have undergone recurrent mutation, alphoid heteroduplex haplotypes associate well with haplogroups defined by unique mutations (Santos et al. 1995). This polymorphic system suffers from ambiguity in assigning heteroduplex patterns to haplotypes and also in not knowing how different haplotypes are related to one another.

Multiple complex haplotypes can be defined by probing genomic *Taq*I digests with the Y-specific probe 49f (Ngo et al. 1986; Lucotte et al. 1990) which detects a polymorphic repeated sequence on Yq. Over a hundred different haplotypes can be distinguished by this probe (Poloni et al. 1997). Relationships between the different haplotypes are unknown and analysis of these has

required many unsupported assumptions (e.g. Hazout and Lucotte 1986). Despite this there is again some association between 49f haplotypes and haplogroups defined by unique mutations (Jobling 1994). This technique suffers the disadvantage of not being applicable to PCR and thus requires large amounts of DNA to provide information that is at best ambiguous.

Advantages of Y chromosome research: the combinatorial approach

The advantages of using non-recombining parts of the genome for evolutionary studies have already been addressed. As outlined above, each individual polymorphic system has its strengths and its weaknesses, and there is no one ideal system. An advantage of the Y chromosome over the other non-recombining parts of the genome is the range of independent polymorphic systems available for informative typing. A combinatorial approach combining both unique and recurrent mutational systems allows optimum discrimination of both individuals and populations (Jobling and Tyler-Smith, 1995; (Santos and Tyler-Smith 1996); (Mitchell and Hammer 1996). These markers are best combined in a hierarchical fashion, defining lineages with unique markers prior to investigating the intra-lineage diversity with faster mutating recurrent markers (de Knijff et al. 1997). This kind of approach has been called 'genealogical' (Richards et al. 1997). Many recurrent mutations observed within a microsatellite network of a population can be eliminated if networks are instead constructed for individual haplogroups defined by unique mutations, as shown in figure 1.7. This is due to the fact that all chromosomes within a haplogroup are derived from a common ancestral chromosome that is likely to be far younger than the ancestral chromosome of a geographically defined population (de Knijff et al. 1997). The amount of diversity observed within a haplogroup defined by unique mutations, assayed using recurrent mutations with much higher mutation rates, such as microsatellites or minisatellites, gives an idea of the demographic history of that haplogroup (de Knijff et al. 1997).

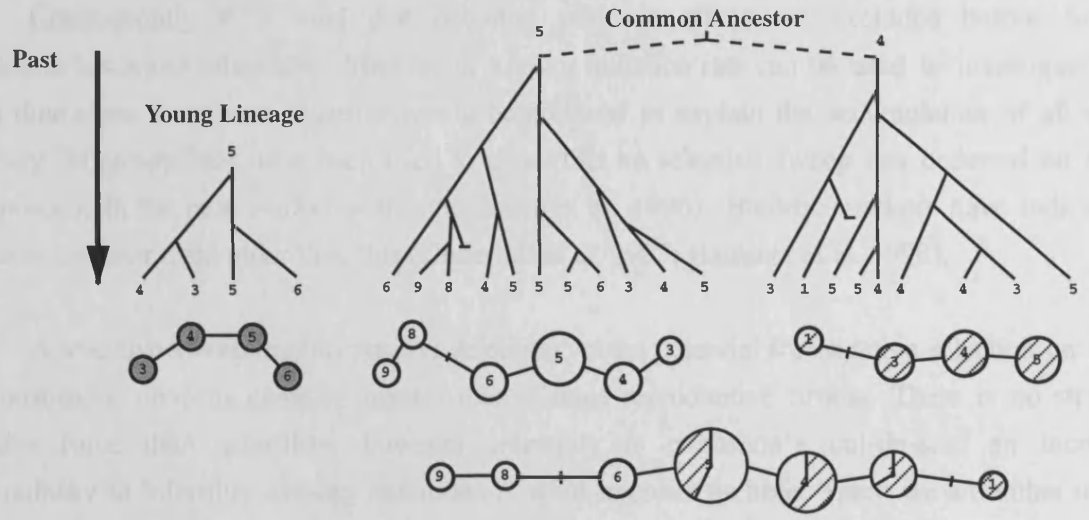


Figure 1.7: Schematic diagram illustrating the minimising of confusion between IBS and IBD by adopting a genealogical approach. The evolution of a single microsatellite locus is followed through three different lineages. Numbers indicate the length of the microsatellite allele in repeat units. Networks under each lineage are constructed from the extant diversity at the single microsatellite locus. Circles represent individual microsatellite alleles the length of which is written inside the circle. Circle area is proportional to the frequency at which each allele is observed. Note that the network of the young lineage faithfully recapitulates the true evolutionary connections, where as the networks of the older lineages confuse IBD and IBS to a limited degree, but not as much as the lowest network which represents the sum of the two older networks, effectively ignoring the existence of the marker which distinguishes between the two lineages.

Selection on the Y chromosome

Selective effects at any locus obscure the relationship between extant diversity and population history. This effect is magnified greatly at non-recombining loci for the following reason. Should selection result in the fixation of a certain Y chromosome it would wipe the slate clean in terms of the diversity apparent within extant Y chromosomes, for a single allele at every polymorphic locus on the Y chromosome would become fixed through hitch-hiking. This would result in an unexpectedly recent common ancestor for all Y chromosomes (Whitfield et al. 1995). Under neutral expectation and given the likely long term effective population size of the human Y chromosome of 5000 (Hammer 1995) we would expect to find a common ancestor for all extant Y chromosomes, in the absence of selection, around 200,000 years ago.

Consequently it is vital that potential selective effects be excluded before making population-historical inferences. Markers of known mutation rate can be used to investigate how much time since a common ancestor would be required to explain the accumulation of all extant diversity. Microsatellites have been used to show that no selective sweep has occurred on the Y chromosome in the past 74,000 years (Goldstein et al. 1996). Biallelic markers have indicated a common ancestor even older than this (Underhill et al. 1997; Hammer et al. 1998).

A selective sweep implies positive selection yet the potential for negative selection on the Y chromosome is obvious given its pivotal role in male reproductive fitness. There is no stronger selective force than infertility, however infertility is evolution's cul-de-sac, an increased susceptibility to infertility-causing mutations is what interests us here. There are a number of XY translocations that cause sex-reversal which results in infertile XX males. A recent study revealed that one specific subclass of these XY translocations was associated with a chromosomal inversion found only within a subset of Y-chromosomal lineages (Jobling et al. 1998). However comparing the low prevalence of this translocation to the relatively small effective population size of the Y chromosome indicated that it was very unlikely to have any impact on the distribution of the lineage, as drift would have a significantly greater impact than any selective forces.

There are other, less rare, forms of infertility-causing mutations. Multiple classes of Y-chromosomal microdeletion have been identified that result in reduced fertility or even infertility (Vogt et al. 1996). A number of studies have sought to investigate the distribution of these classes of microdeletions amongst different Y-chromosomal lineages (Previdere et al. 1999; McElreavey and Krausz 1999) with little success in identifying predisposing lineages. One recent study has claimed that mean sperm concentration varies significantly between different lineages (Kuroki et al. 1999).

Apart from infertility a number of other phenotypic effects have been ascribed to the Y chromosome (Kittles et al. 1999; Lau 1999). Recently one family of repeated genes on the Y chromosome have become implicated in gonadoblastoma, testicular and prostate cancers (Lau 1999). In addition certain behavioural traits have been associated with the Y chromosome (Kittles et al. 1999). The selective implications of either these cancers or behaviour remains unquantified.

Applications of the Y chromosome

Studies of Y-chromosomal diversity can be used for a variety of purposes other than to address evolutionary issues.

Infertility studies of the Y chromosomes have localised three regions (AZFa, AZFb and AZFc) in Yq11 that are often deleted *de novo* in azoospermic, oligozoospermic and infertile normospermic males (Vogt et al. 1996). All three regions contain multiple known genes, though there is no guarantee that other, presently unknown, genes do not also lie in these regions (McElreavey and Krausz 1999). Often the known genes represent subsets of the families of repeated spermatogenic genes. It has proven difficult to localise the phenotypic effects of these microdeletions to individual genes. The identification of lineages predisposed to undergoing certain microdeletions will aid the discovery of the physical structures that underpin these infertility-causing mutations and provide insight into the processes of spermatogenesis itself (Jobling et al. 1998).

The male-specificity of the Y chromosome makes it potentially useful for certain forensic applications (reviewed in Jobling et al. 1997; Kayser et al. 1997). Mixed male assailant and female victim DNA samples are a special case in which the Y chromosome provides advantages over autosomal markers. The initial lack of known polymorphisms of the Y chromosome has recently been rectified by the discovery of many more markers in the past few years which have markedly improved the discriminating power of this locus (Kayser et al. 1997). Multiallelic, highly mutable markers are especially applicable for discriminating between individuals. However the present discriminating capacity of the Y chromosome remains much less than that of the multiplexed autosomal microsatellite markers that are the mainstay of forensic DNA typing. Consequently whilst the Y chromosome is powerful for exclusions, the lower discriminating capacity and likely effects of the known substantial population substructuring make inclusions more problematic (Jobling et al. 1997). One further disadvantage of the Y chromosome is the fact that identical haplotypes are carried by many of the suspect's male relatives, such as brothers, sons, father and paternal uncles (Jobling et al. 1997; Foster et al. 1999).

However, this sharing of identical Y-chromosomal haplotypes amongst male relatives makes the Y chromosome particularly useful for paternity testing where the alleged father is not available (Jobling et al. 1997; Foster et al. 1998), although this does render the analysis sensitive to non-paternities in other father-son pairs and many of the problems with inclusion remain. A recent

publication used Y-chromosomal markers to resolve a 200 year old paternity case involving the US president and 'secular saint', Thomas Jefferson (Foster et al. 1998).

The co-inheritance of Y chromosomes with surnames in many cultures around the world makes it of potential interest to those interested in genealogical studies (Jobling and Tyler-Smith 1995). Surname distributions themselves have been used to investigate spatial patterns of genetic diversity (Lasker 1985; Sokal et al. 1992), though European surnames are generally thought to have their origins in the 13th century (Lasker 1985). Whilst many surnames have multiple origins, others are likely to be monophyletic. The Y chromosome provides a method of discriminating between these two hypotheses. The non-recombining patrilineal inheritance of the Y chromosome means that it can resolve deep-rooting pedigrees in a manner not possible for autosomal markers (Jobling et al. 1999). This is useful both to verify non-paternities within deep-rooting pedigrees used for other purposes (Jobling et al. 1999) and to ascertain whether two males with the same surname come from the same monophyletic lineage of Y chromosomes and thus share a recent common ancestor (Mark Jobling, personal communication).

The differentiation of human Y chromosomes

A locus can be said to exhibit geographical differentiation if it varies between geographically separated populations more than could be expected if the populations were panmixic and any differences result solely from sampling (Nei 1987). Genetic distances between populations that are larger than could be expected under a null hypothesis of random mating, are taken as indication that a locus exhibits significant differentiation (Roewer et al. 1996). The amount of geographical differentiation of a locus can be compared to other loci by investigating how genetic distances vary within geographical distance when the loci are typed in the same populations (Seielstad et al. 1998). These can be compared by means of a graph as shown in figure 1.8.

The extent of differentiation is conditional on a number of factors other than population history, for example the mutation rate of the loci and its effective population size (Nei 1987; Stoneking 1998). The lower the effective population size, the greater the effect drift has on allele frequencies and the more rapidly populations diverge, in both time and space. Consequently under neutral expectations mtDNA and the Y chromosome in having the same effective population sizes should exhibit similar, high, levels of geographic differentiation - significantly greater than autosomal loci. A recent paper has indicated that the Y chromosome shows greater differentiation

both globally and within Africa than either mtDNA or autosomes (Seielstad et al. 1998). It has been suggested that this is evidence for a higher prehistoric rate of female migration. At first sight this might seem counter-intuitive given that males often seem to be more mobile. However the crucial parameter is that of inter-generational movement. Patrilocality is the practice by which children resulting from matings between two people from different areas are brought up in the paternal home. It is estimated that the majority of human populations practise patrilocality (Seielstad et al. 1998). Under this scenario the Y chromosome has remained stationary whilst the mtDNA has moved. Over many generations it can be envisaged that this will have a homogenising influence on the mtDNA genetic landscape. By contrast the geographical differentiation of the Y chromosome is such that many lineages defined by unique markers are found to have population-specific distributions (Jobling et al. 1997).

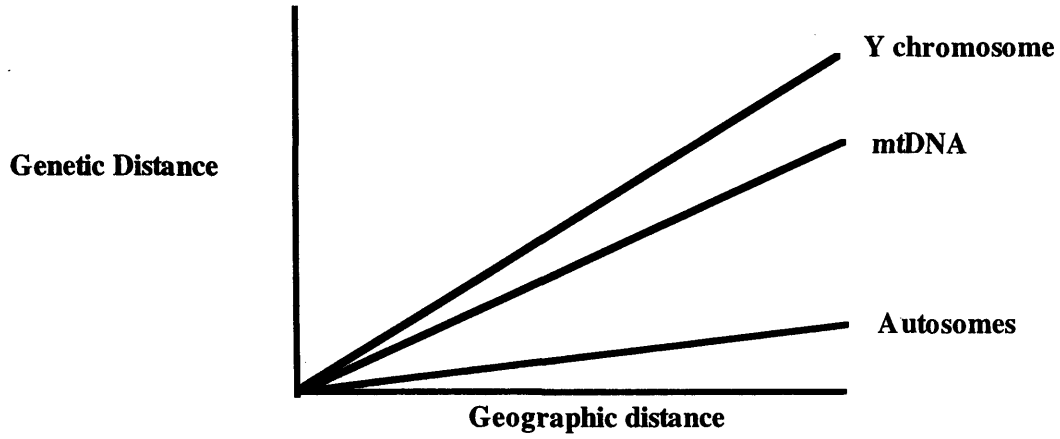


Figure 1.8: Schematic graph indicating the greater geographic differentiation of the Y chromosome compared to both mtDNA and autosomal loci. (Taken from Seielstad et al., 1998, see text)

The above conclusion of higher female migration rates reminds us that the Y chromosome only provides information about a single, albeit large, locus and only about one of the two sexes. Therefore it should not be viewed as being in competition with the other regions of the human genome for evolutionary study but as one part of a collaborative picture containing complementary parts. Conflict between mitochondrial and Y-chromosomal pictures of evolution does not necessarily imply that one is wrong, but that there are potentially interesting differences between male and female migration patterns or intriguing mating structures which have occurred in prehistory.

Another major study that sought to investigate global differences in mtDNA and Y-chromosomal diversity, addressed the issue of language correlations (Poloni et al. 1997). It found a better correlation of language affiliation with paternal lineages than with maternal lineages, and suggested that this might be due to children adopting their father's tongue rather than their mother's, in contrast to the phrase 'mother tongue' (Poloni et al. 1997).

Aims of Y chromosome human evolutionary studies

Two main approaches have been adopted when using the Y chromosome for evolutionary studies. Though both provide inferences on population history, they differ in focus.

The first approach focuses on an individual lineage: its aims are to characterise its world-wide distribution and assay the diversity within it. The analytical tools used have been developed relatively recently. During the period when few Y-chromosomal lineages had been defined this was the only approach available. Although lineages to be investigated can be chosen on the basis of a distribution within populations of interest, often they were chosen by virtue of being the only new lineage not yet characterised. As such it is difficult to see how this approach can be purely question-driven. A number of papers have consequently become involved in certain anthropological issues almost by accident (Underhill et al. 1996; Zerjal et al. 1997; Bergen et al. 1999).

The second approach to Y-chromosomal evolutionary studies is to attempt to address a specific anthropological issue by assaying all known lineages within the relevant populations. This question-based, population-focused approach has only become feasible in Y chromosome research through the recent delineation of markers polymorphic in all regions of the world. The aims of this research are to determine the haplotypic makeup of individual populations at the finest possible resolution. The analytical tools used are little different from those applied to allele frequencies. Some studies have attempted to investigate migrational movements at deep time depths by assaying populations from around the world (Poloni et al. 1997; Hammer et al. 1998). Popular regional research topics have proven to be those that have previously been addressed using other loci and include investigating population isolates such as the Finns (Sajantila et al. 1996; Lahermo et al. 1999) and Basques (Lucotte and Hazout 1996; Perez-Lezaun et al. 1997), the colonisation of the Americas (Santos et al. 1999; Karafet et al. 1999) and the spread of agriculture and possibly of the farmers themselves into Europe (Semino et al. 1996).

Outline of this study

This thesis aims to apply and improve both of the above approaches, and in part to provide a unification of them in order to bring the advantages of lineage analysis to bear on the resolution of question-driven, population-focused studies. At the time of the initiation of this work in late 1996, Y-chromosomal lineage analysis was in its infancy, few lineages were known and sampling was sparse. Y-chromosomal research was still very much the younger, puerile brother to the older, wiser mtDNA sister.

Chapter 3 details the development of the novel mutation detection technique, Denaturing High Performance Liquid Chromatography (DHPLC), to aid identification of more unique markers capable of defining monophyletic lineages. Whilst the technique itself proved successful the return of a single new polymorphism from a screen of Y-specific sequences was disappointing. In chapter 4 the tools for analysing an individual lineage are tested against one another and adapted for application to intra-lineage diversity of the minisatellite MSY1. These tools are then applied to two lineages with restricted population distributions, to infer about the population histories of two population isolates within Europe; the Basques and the Gypsies. Chapter 5 discusses the incorporation of geographical information into regional studies of multiple populations and develops a new method for ascertaining significant barriers to gene flow within a genetic landscape. This method is applied to a high-resolution dataset of European Y-chromosomal lineages to indicate that language boundaries are significant barriers to male gene-flow within Europe. Chapter 6 describes the collection of a large database of Y-chromosomal diversity within Oceania and Southeast Asia. Polymorphic information from biallelic markers, microsatellites and MSY1 is fed into the analyses developed previously to investigate the population history of this region, specifically the colonisation of Polynesia.

Chapter 2: General materials and methods

Materials

Buffers

Loading buffer (5x) - 5x TBE, 15% (w/v) Ficoll-400 (Pharmacia), 0.05% (w/v) bromophenol blue or xylene cyanol (Sigma)

MVR-PCR stop solution - 98% (v/v) formamide, 20mM Na_2EDTA, 0.05% (w/v) bromophenol blue, 0.05% (w/v) xylene cyanol (Sigma)

PCR buffer (11.1x) - 40mM Tris, 10mM $(NH_4)_2SO_4$, 4mM $MgCl_2$, 6mM β–Mercaptoethanol, 1mM each dNTP, 4nM Na_2EDTA (Jeffreys *et al.*, 1990).

PCR buffer (AB 10x) - 750mM Tris-HCl (pH 8.8), 200mM $(NH_4)_2SO_4$, 0.1% (v/v) Tween-20

PCR buffer (Tbr 1x) - 10mM Tris-HCl pH8.8, 50mM KCl, 1.5mM $MgCl_2$, 0.1% (v/v) Triton X-100 (Sigma)

10x TBE - 0.89M Tris-borate, 2mM Na_2EDTA (pH 8.3)

10x TBE (sequencing) - 1M Tris-borate, 2mM Na_2EDTA (pH 8.3)

Genomic DNA samples

Genomic DNA was used as a template for PCR at a concentration of 5-50ng/μl. Genomic DNA samples were kindly supplied by Mark Jobling and Chris Tyler-Smith. Other suppliers of DNA samples are acknowledged in the relevant chapters.

Methods

PCR

Primer precipitation

Primers were made by the Leicester University Protein and Nucleic Acid Chemistry Laboratory (PNACL), and supplied as a crude stock in ammonium hydroxide solution.

1) Add 30µl of 3M sodium acetate (pH 5.2) to 300µl of crude primer stock and mix.

2) Add 990µl of absolute ethanol and mix

3) Spin at 13,000rpm for 1 minute

4) Wash pellet with 1ml of 80% (v/v) ethanol for 5 minutes at room temperature

5) Spin at 13,000 rpm for 1 minute

6) Pipette off supernatant and let pellet air dry

7) Resuspend pellet in 100µl water

8) Make up 1ml solutions of 1000x and 500x dilutions of this primer stock in water

9) Measure the optical density of these dilutions at a wavelength of 260nm, against a water blank, using quartz cuvettes.

10) Calculate the concentration of the primer stock in µM using the formula below:

$$\frac{OD^{260} \text{ (of 500x dil}^{n}\text{)} \times 30770}{\text{length of primer in bases}}$$

Primer labelling (for MVR-PCR)

The four repeat variant specific primers for MVR-PCR were end-labelled with γ-[^{33}P]-ATP (NEN Dupont) in the following reaction mix to provide sufficient primer for 24 MVR-PCR amplifications in a 96 well plate (24 reactions x4 primers=96 wells).

5x Forward buffer (Gibco)	5.6µl
primer (10µM)	2.8µl
T4 Kinase (Gibco)	1.4µl
^{33}P γ-ATP	2.8µl
ddH_2O	15.4µl

This mix was then incubated at 37°C for at least 30 mins. Unincorporated ATP was not removed prior to PCR.

Primer annealing temperature

Approximate primer annealing temperatures were calculated using the rule A/T = 2°C and C/G = 4°C and summing the total for each primer sequence. If no specific product was obtained then the annealing temperature was titrated to give the best amplification.

High throughput PCR-RFLP

All 16 of the polymorphisms described in table 2.1 can be typed using similar PCR protocols. Although the cycling programs differ, the reaction volume (10µl) and composition is the same. 10-20ng of genomic DNA are added to a PCR buffer solution containing, 1µM each primer, 0.5U *Taq* DNA polymerase (Advanced Biotechnologies) and 111ng/µl of BSA,

All PCR reactions were done in 96-well Thermowell M microtitre plates in an MJR PTC-200 machine. For RFLP analysis one unit of the relevant restriction enzyme in 10µl of 2 times digestion buffer was added directly to the PCR product and incubated at the relevant temperature for two hours. All 96 digest/PCR products were run out on a single agarose gel (1-3%), using a gel running system that was made by the Leicester University Biological Sciences Workshop. All pipetting operations were performed using a 12-channel pipette. An example of such a gel is shown in figure 2.1.

Name	Reference	Primers (5´- 3´)	PCR Conditions / (temp) time	Assay	Ancestral
YAP	Hammer and Horai, 1995	cag ggg aag ata aag aaa ta act gct aaa agg gga tgg at	(94)30sec, (53)30sec, (72)30sec 34 cycles	Size difference: YAP+ 455; YAP- 150bp	YAP -
sY81	Seielstad et al., 1994	agg cac tgg tca gaa tga ag aat gga aaa tac agc tcc cc	(94)60sec, (60)60sec, (72)60sec 32 cycles	209bp: NlaIII gives 144bp G and 102bp A allele (internal site)	A to G
DYS199	Underhill et al., 1995	taa tca gtc tcc tcc cag ca agg tac cag ctc ttc cca att	(94)20sec, (59)20sec, (72)30sec 36 cycles	202bp: MfeI gives 182 and 20bp Forced in site	C to T site present
Tat	Zerjal et al., 1997	gac tct gag tgt aga ctt gtg a gaa ggt gcc gta aaa gtg tga a	(94)30sec, (60)30sec, (72)30sec 33 cycles	112bp: NlaIII/Hsp92II (T allele) or MaeII (C) gives 85 and 27bp	T to C
92R7	Hurles et al., 1999	gac ccg ctg tag acc tga ct gcc tat cta ctt cag tga ttt ct	(94)30sec, (62)30sec, (72)60sec 33 cycles	709bp: HindIII partial cleavage gives 197 and 512bp	C/T unclear
47z	Shinka et al., 1999	ttt gct tct cat ttc atc tg tta gat gaa ttg ttg tgc tg	(94)60sec, (54)60sec, (72)120sec 36 cycles	1.6kb: StuI gives 850 and 750bp, and 1.6kb from X locus	StuI absent
SRY -1532	Whitfield et al., 1995;	tcc tta gca acc att aat ctg g aaa tag caa aaa atg aca caa ggc	(94)30sec, (59)30sec, (72)30sec 34 cycles	167bp: DraIII gives 112 and 55bp	A to G site absent
SRY -3225	Whitfield et al., 1995	caa ctg ttg aga aat agt cat c ccc aga tgc ata tat tac agg	(94)30sec, (58)30sec, (72)60sec 34 cycles	580bp: HaeIII gives 414+166bp or 294+120+166bp (internal site)	C to T site present
SRY -8299	Whitfield et al., 1995	aca gca cat tag ctg gta tga c tct ctt tat ggc aag act tac g	(94)30sec, (62)30sec, (72)60sec 33 cycles	509bp: BsrBI gives 362 and 147bp	G to A site present
SRY +465	Shinka et al., 1999	gcc gaa gaa ttg cag ttt gc gtt gat ggg cgg taa gtg gc	(94)30sec, (58)30sec, (72)30sec 33 cycles	148bp: BsoFI gives 99 and 49bp	C to T site present
SRY -2627	Veitia et al., 1997 Hurles et al., in press	agg tct ttt ttg cct tct ta atg cac ggt ttc ttt tga	(94)30sec, (52)30sec, (72)120sec 33 cycles	1242bp: HgiAI (Bsi HKA 1) gives 298 & 944bp	can't tell from chimp
DYS234	Underhill et al 1997 Hurles et al., 1998	tcc tag gtt atg att aca gag cg tgc aga aca ttt gta ctg ttc c	(94)30sec, (60)30sec, (72)45sec 34 cycles	273bp: NdeI gives 200 and 70bp	G to C site absent
Lly22g	Righetti and Tyler-Smith, unpublished	cca ccc agt ttt atg cat ttg ata gat ggc gtc ttc atg agt	(94)30sec, (55)60sec, (72)60sec 33 cycles	850bp: HindIII gives 500, 230 and 120bp (C) or 650, 500, 230 and 120bp (A)	C to A site present
M9	Underhill et al 1997	gca gca tat aaa cat ttc agg aaa acc taa ctt tgc tca agc	(94)30sec, (58)30sec, (72)30sec 33 cycles	3411bp: HinfI gives 182, 93 and 66bp (C) or 248 and 93bp (G)	C to G site present
DYS257	Hammer et al 1998	gaa ctt gtc ggg agg caa t tga tac act tcc tcc ttt agt gg	(94)30sec, (60)30sec, (72)30sec 33 cycles	BanI gives 182, 106, 63 and 43 bp (G) or 288, 182, 106, 63 and 43bp (A)	G to A site present
12f2	Jobling, Shlumukova and Hurles, unpublished	ctg act gat caa aat gct tac aga t tct tct aga att tct tca cag aat t	(94)30sec, (59)30sec, (72)45sec 34 cycles	Presence/absence of 500bp product; coamplify with a larger product	undeleted

Table 2.1: Details of biallelic polymorphisms used in this thesis

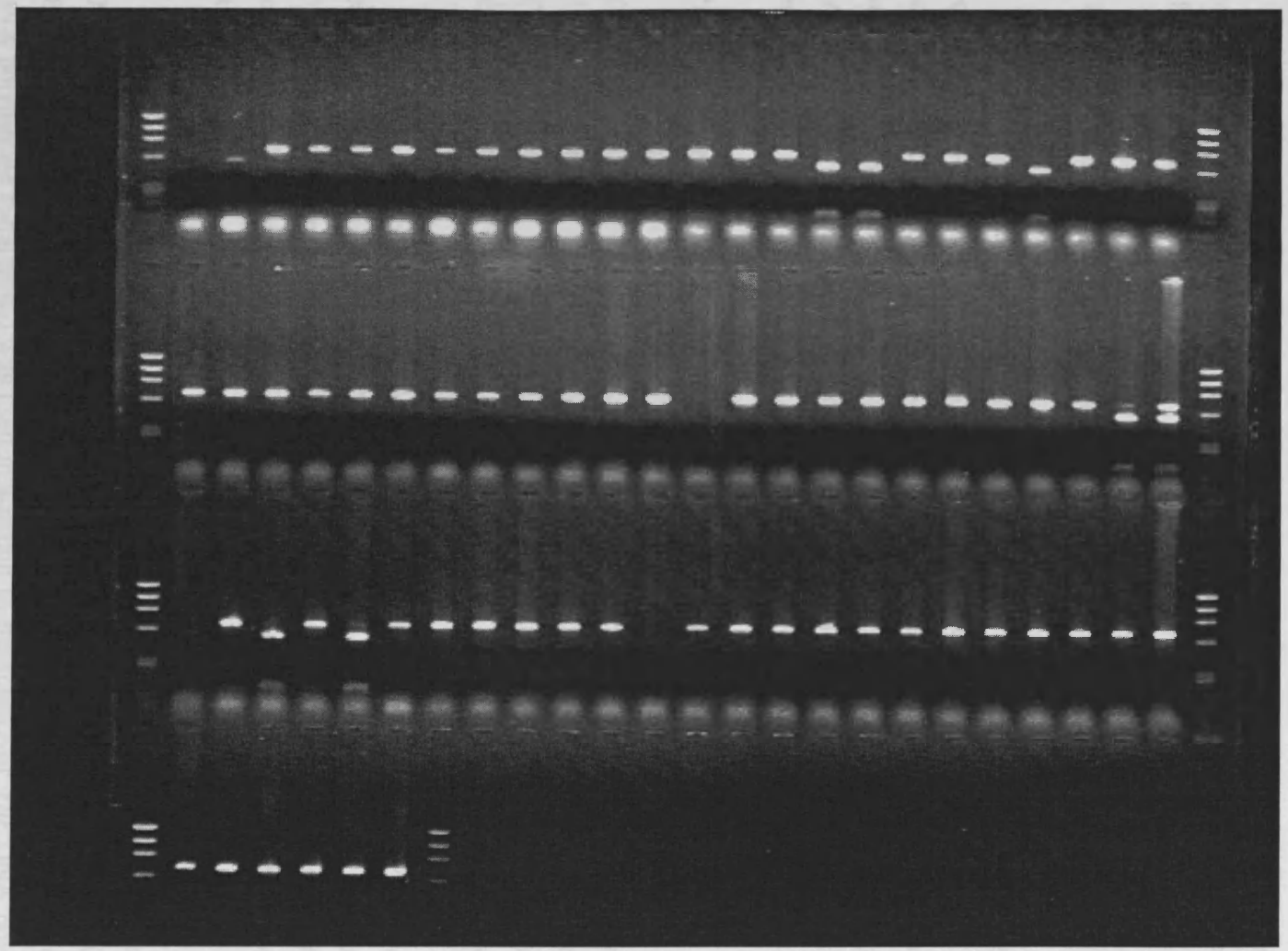


Figure 2.1: Example of PCR-RFLP typing using agarose gel electrophoresis, 78 samples are loaded on this gel, flanked by size ladders. A polymorphic base substitution generates a restriction enzyme site which is only cleaved to give two products in a subset of samples.

For allele specific amplifications the products of each primer pair were loaded in the same well of the gel, staggered at an interval of 10-15 minutes. Thus only a single band should be seen in each track (an example is shown in figure 3.6).

DNA masterplates were set up in Thermowell M 96-well plates with dilution factors of genomic DNA stocks titrated to give solutions containing 5-10ng/µl of DNA so as to minimise the effects of sample variation.

Three-state Minisatellite **Variant Repeat PCR (MVR-PCR)**

A flanking PCR was performed using the primers Y1A+ and Y1B+ (Jobling *et al.*, 1998) on 50ng of genomic DNA and run out on a 1% agarose gel. The variant MSY1 band (~2-3kb) was cut out and incubated for at least 16 hours at 4°C in 400µl of water. 2µl of this eluate were then used as template in four 10µl MVR-PCR reactions which contain one of the flanking primers and one ^{33}P (NEN Dupont) end-labelled discriminator primer that recognises only one of the three repeat types.

Primers used	**Repeat type detected**	**Direction**
Y1A+ and TAG1D	1	Forward
Y1A+ and TAG3C	3	Forward
Y1B+ and TAG3R3	3	Reverse
Y1B+ and TAG4R3	4	Reverse

After amplification 12µl of stop solution were added to each MVR-PCR product. 6µl of each reaction were run out on a 50cm 2.5% denaturing polyacrylamide gel (Sequagel, National Diagnostics) after denaturation at 80°C for 2 minutes. The gels were fixed with a solution of 10% acetic acid and 12% methanol (Sigma), then dried and exposed to X-ray film (Fuji) for 16-72 hours at room temperature.

The buffer for the flanking PCR reaction was that supplied with the *Thermus brockianus* (*Tbr*) polymerase (NBL), and was supplemented with 200µM of each dNTP and 200µg/ml BSA (Boehringer). The Y1A+ and Y1B+ primers were used at 100nM concentration and 0.02U of *Pfu* DNA polymerase (Stratagene) was used in addition to the 0.66U of *Tbr* DNA polymerase. The reaction volume was made up to 10µl with water.

The buffer for the MVR-PCR was also that supplied with the *Tbr* polymerase and was again supplemented with 200μM of each dNTP and 200μg/ml BSA (Boehringer). 0.5U of *Tbr* DNA polymerase alone was used per reaction. The flanking primer was used at a concentration of 100nM and the labelled discriminator primers at concentrations of 100nM (TAG1D and TAG3C), 50nM (TAG3R3) and 200nM (TAG4R3).

Both reactions were carried out using Thermowell M plates (MJR) in an MJR PTC-200 thermal cycler using the programs detailed below, under calculated mode. When DNA quantities were limiting or quality poor, only 5-10ng of template was added to the flanking reaction and step 9 was extended by two cycles.

Flanking PCR:

1 - 96°C for 40 secs
2 - 94°C for 8 secs
3 - 68°C for 1 min -0.5°C each cycle
4 - 68°C for 3 mins
5 - Goto 2 10 times
6 - 94°C for 8 secs
7 - 63°C for 1 min
8 - 68°C for 3 mins +4 secs each cycle
9 - Goto 6 25 times
10 - 5°C for ever

MVR-PCR:

1 - 96°C for 40 secs
2 - 94°C for 8 secs
3 - 62°C for 1 min
4 - 68°C for 3 mins
5 - Goto 2 2 times
6 - 94°C for 8 secs
7 - 68°C for 30 secs +4 secs each cycle
8 - 68°C for 3 mins 30 secs
9 - Goto 6 20 times
10 - 5°C for ever

Primer sequences (5' to 3'):

Y1A+: ACA GAG GTA GAT GCT GAA GCG GTA TAG C
Y1B+: GCA ACT CAA GCT AGG ACA AAG GGA AAG G

TAG1D: tca tgc gtc cat ggt ccg gaT GTG TAT AAT ATA CAT CAT GTA TAT TG
TAG3C: tca tgc gtc cat ggt ccg gaT GTG TAT AAT ATA CAT GAT GTA TAT TG
TAG3R3: tca tgc gtc cat ggt ccg gaC ATC ATG TAT ATT ATA CAC AAT ATA CAT C
TAG4R3: tca tgc gtc cat ggt ccg gaC ATC ATG TAT ATT ATA CAT AAT ATA CAT C

Lower case bases are the 'tag' sequence (Jeffreys *et al.*, 1991) used to decouple discrimination from amplification (see Jobling *et al.*, 1998).

An example of an autoradiograph showing MSY1 codes generated by MVR-PCR is shown in figure 2.2

Sequencing of PCR products

Preparation of template DNA

Initially the PCR product to be sequenced was amplified and run out on an agarose gel; the correct band was cut out and incubated in 400µl of water for 16 hours at room temperature. The eluate was then used as template DNA in a subsequent reamplification that was cycle-titrated to maximise the yield of the specific product. A number of 10µl reamplification reactions were pooled and the DNA purified using a Qiagen PCR purification kit as per the manufacturers instructions. An aliquot of the final DNA containing eluate was run out on a gel to determine the concentration.

Automated sequencing

Automated fluorescent sequencing was performed by the Leicester University Protein and Nucleic Acid Chemistry Laboratory using Big Dye Terminators and the products were run out on a ABI 377 machine.

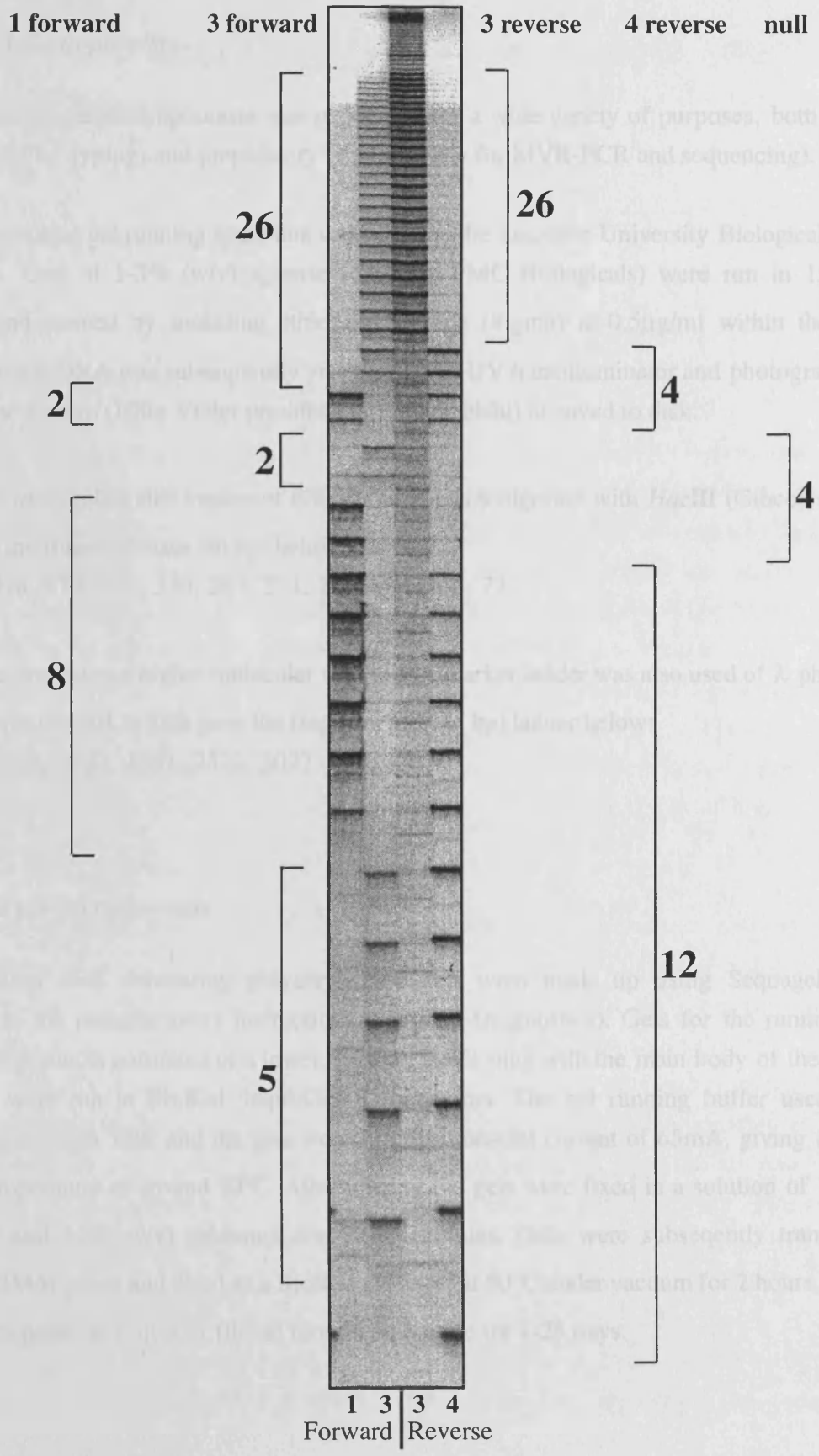


Figure 2.2: Example of a 2.5% polyacrylamide gel of MVR-PCR products displaying an entire MSY1 code: (3)5(1)8(3)2(1)2(3)26(4)4(0)4(4)12

Agarose gel electrophoresis

Agarose gel electrophoresis was performed for a wide variety of purposes, both analytical (e.g. PCR-RFLP typing) and preparatory (e.g. template for MVR-PCR and sequencing).

Horizontal gel running apparatus was made by the Leicester University Biological Sciences Workshop. Gels of 1-3% (w/v) agarose (Seakem, FMC Biologicals) were run in 1x TBE at 7.5V/cm and stained by including ethidium bromide (Sigma) at 0.5μg/ml within the gel and running buffer. DNA was subsequently visualised on a UV transilluminator and photographs taken with a video system (Ultra Violet products, Inc./ Mitsubishi) or saved to disk.

For most gels a size marker of ϕX174 phage DNA digested with *Hae*III (Gibco) was used. This gives the fragment sizes (in bp) below:
- 1353, 1078, 872, 603, 310, 281, 271, 234, 194, 118, 72

In some cases a higher molecular weight size marker ladder was also used of λ phage DNA digested with *Hin*dIII, which gave the fragment size (in bp) ladder below:
- 23130, 9416, 6557, 4361, 2322, 2027, 564, 125

Acrylamide gel electrophoresis

0.4mm thick denaturing polyacrylamide gels were made up using Sequagel reagents according to the manufacturers instructions (National Diagnostics). Gels for the running out of MVR-PCR products consisted of a lower, inch deep, 6% plug with the main body of the gel being 2.5% and were run in BioRad Sequi-Gen II apparatus. The gel running buffer used was 1x sequencing strength TBE and the gels were run at a constant current of 65mA, giving a constant running temperature of around 50°C. After running the gels were fixed in a solution of 10% (v/v) acetic acid and 12% (v/v) methanol for 10-15 minutes. Gels were subseqently transferred to Whatman 3MM paper and dried in a BioRad gel dryer at 80°C under vacuum for 2 hours. Once dry gels were exposed to Fuji R-X film at room temperature for 1-25 days.

Chapter 3: Searching for new single nucleotide polymorphisms

Introduction

Single nucleotide polymorphisms in the human genome

Prior to having any knowledge about intraspecific genetic diversity the expectation was that all genomes would have been shaped by selection. Consequently any polymorphism that might be found should be due either to balancing selection or incomplete fixation of a soon to be fixed allele (Avise 1994). The discovery of far more diversity than could readily be explained from a selectionist stance caused wide consternation beacause the genetic load would be far too great for all this diversity to be maintained by selection. The publication of the 'neutral theory' by Motoo Kimura in 1969 provided an alternative hypothesis; that most polymorphism was selectively neutral and that allele frequencies fluctuated through random genetic drift (Kimura 1983). Subsequently arguments raged between the two camps of selectionist and neutralists. It is now apparent to most people that polymorphism within the human genome results from the forces of both selection and neutral processes.

It is estimated that only 20% of polymorphisms in the human genome are repeat polymorphisms whereas 80% of polymorphism involves sequence changes at a single nucleotide (Haff and Smirnov 1997). Single nucleotide polymorphisms (SNPs) are distributed all over the genome, although it is known that they are non-uniformly distributed and that there is a higher nucleotide diversity within sequences that can form replication origins and within subtelomeric regions (M. Fullerton, personal communication; Baird et al. 1995). Within coding regions base changes that result in a change in amino-acid sequence are called 'non-synonymous', whereas those that do not are 'synonymous'. Two recent large scale screens for SNPs have estimated a frequency of one every 350bp in the autosomes, with no significant difference in frequency between coding and non-coding regions, consistent with a population evolving according to neutral expectation (Cargill et al. 1999). However the non-coding regions focused on in one of the studies were flanking genes and therefore might still be under direct or indirect selective constraints, thus restricting polymorphism (Cargill et al. 1999). Indeed the diversity apparent in non-coding regions

was only half that at four-fold degenerate sites, where any nucleotide change is synonymous, which might reasonably be expected to approximate to the rate of neutral mutation. The role of selection within coding SNPs is apparent in the type of SNP present. Nonsynonymous SNPs are present as a lower fraction of coding SNPs than would be expected by chance, the ratio of nonconservative to conservative nonsynonamous SNPs is also lower than expected. In addition there is a decreasing gradient of nucleotide diversity going from four-fold degenerate to two-fold degenerate to non-degenerate sites within coding regions.

Germ-line mutations causing SNPs are thought to result from error-prone endogenous processes relating to either chemical or enzymatic mechanisms (Krawczak et al. 1998). The methylation-mediated deamination of 5-methylcytosine in CpG dinucleotides is a prime example of a chemical mechanism. It results in the replacement of cytosine by thymine and is responsible for the higher mutability of CpG dinucleotides over all other dinucleotides (Krawczak et al. 1998). Mismatch repair and exonucleolytic proof-reading are examples of enzymatic processes thought to be responsible for germline SNP mutations. Another mutagenic model for SNPs relies on transient template-primer misalignment resulting in the incorporation of a noncomplementary nucleotide (Kunkel 1990).

Single nucleotide polymorphism and disease

The vast majority of all known lesions causing human genetic disease are single base pair substitutions (Krawczak et al. 1998). Repeat sequence mutations play a relatively rare role in disease and tend to be limited to certain genes and hence certain diseases, for example the triplet repeat diseases. The human gene mutation database (HGMD) contains roughly 8000 different lesions in the coding regions of over 500 genes, only 20 of which involve repeat mutation (Krawczak et al. 1998).

Whilst the majority of SNPs that cause disease are nonsynonymous base substitutions within the coding regions of a gene, SNPs in noncoding regions of genes can also be pathogenic whether it be at intronic splice sites or in flanking regulatory regions (Krawczak et al. 1998). All nonsynonamous base substitutions are not equal in their pathogenicity. Those substitutions that result in the conservative replacement of an amino acid residue of similar chemistry to the lost residue are less likely to have a pathogenic effect than those that replace amino acids of differing chemistries. The genetic code has evolved to minimise the deleterious effect of base substitutions

through both degeneracy that results in many base substitutions being synonymous and closely related codons having similar chemistry that minimises the chance of nonconservative amino acid replacement (Hurst 1999).

Given that the efficiencies of some of the mechanisms that generate germline single base mutations are sequence dependant (Krawczak et al. 1998) it has been suggested that the mutability of a given nucleotide may be dependant on the sequence context in which it is found. A recent survey of all mutations in the HGMD found that apart from the well known five-fold higher mutability of CpG dinucleotides there are only subtle and locally confined effects of surrounding sequence on single nucleotide mutability (Krawczak et al. 1998). This study also identified a slightly higher likelihood of a nucleotide mutating to be identical to one of its flanking bases. This provides limited support for the mutagenic model whereby nucleotide substitutions are caused by transient template-primer misalignment resulting in the incorporation of a noncomplementary nucleotide (Kunkel 1990).

Applications of single nucleotide polymorphisms

Despite their wide and frequent distribution within the genome SNPs have not been studied as intensively as other polymorphisms. It is only recently that large scale initiatives have been set up to screen the genome for SNPs. This is partly due to the technical difficulties involved but also to their perceived relative lack of utility when compared to other polymorphisms. Multiallelic markers have long been the polymorphisms of choice for forensics and linkage analysis. Although screens for SNPs are involved in many studies, they are not being sought for future use as a tool, but to provide an answer, for example a cis-acting regulatory mutation (Monckton et al. 1994) or pathogenic variation within a gene.

SNPs as tools have been applied in evolutionary studies. Their low mutation rate within the human genome means that they can often be considered unique events which as we shall see in later chapters makes for more powerful analyses. As long as recombination events can be excluded, individual SNPs can be combined into haplotypes which provide greater information than allele frequencies alone (Jobling and Tyler-Smith 1995). This has been exploited readily in evolutionary studies as evidenced by many mtDNA studies (e.g. (Torroni et al. 1994) and the rapidly growing number of papers investigating prehistoric migrations using autosomal haplotypes (Harding et al. 1997; Tishkoff et al. 1996; Tishkoff et al. 1998; Kaessmann et al. 1999; Jin et al. 1999). SNP

haplotypes are broken down by recombination and recent studies investigating recombination at the molecular rather than genetic level have utilised this fact (Jeffreys et al. 1998). SNP haplotype analysis is also useful to medical geneticists when fine mapping a region thought to be involved in genetic disease.

Single nucleotide polymorphisms for whole genome linkage disequilibrium mapping

The past few years have shown a dramatic increase in the identification of genes responsible for single gene disorders (Krawczak et al. 1998). Many different approaches have been taken towards identifying these genes. These can be usefully classified into those that concentrate on specific genes and those that search the whole genome. The former method, the candidate gene approach, relies on identifying from outside evidence genes relevant to the disease in question. Often this involves knowing the underlying pathology of the disease and relating it to genes involved in those physiological processes. The second approach has been revolutionised by the application of whole genome linkage maps of microsatellite markers. Maps are available that have an average spacing of less than 1cM (Schmitt and Goodfellow 1994).

Most of these whole genome studies have used a linkage analysis approach; they seek to identify chromosomal regions that have been coinherited with the disease in family pedigrees. In contrast, candidate gene studies often take an association approach, that is they seek statistically significant differences in the frequencies of specific alleles between unrelated affected individuals and controls. Association studies can also be used to refine the location of a disease gene within an interval previously defined by linkage studies. They rely on linkage between the marker locus and the disease-causing mutation resulting in a association between a marker allele and the disease phenotype. The phenomenom whereby two markers are found together at a frequency significantly higher than expected if they were unlinked is known as linkage disequilibrium. Linkage disequilibrium is broken down by recombination over time (Kruglyak 1999). Consequently linkage disequilibrium can only be detected over much shorter genetic distances than can linkage, which is only broken down by recombination events within the pedigree itself. Linkage disequilibrium can also inform about the demographic history of the population being studied (Laan and Pääbo 1997).

Common diseases are more often than not multifactorial, that is their aetiology involves the input of many genetic and environmental factors. This genetic heterogeneity reduces the power of linkage analysis to define the chromosomal regions involved. Whole genome association studies

have potentially much greater statistical power than linkage analysis to detect regions involved in common disease (Kruglyak 1999). However it has been estimated from coalescent modelling that in the general population linkage disequilibrium, on average, is unlikely to extend more than 3kb away from the disease causing allele (Kruglyak 1999). Microsatellites are thought to be present in the genome at an average of one every 10kb. Thus the only type of marker with sufficiently dense coverage of the genome to allow this kind of whole genome association study is the SNP. In addition, the low mutation rate of SNPs reduces the likelihood of recurrent mutation reducing the power to detect disease causing variants.

It should be noted that the coalescent modelling of the extent of linkage disequlibrium depends on a number of poorly quantified population genetic and demographic parameters and several unrealistic assumptions. Consequently there is some disagreement within the field concerning both the level of linkage disequilibrium that can be expected within the general population and the type of populations to use to maximise linkage disequilibrium. Isolated populations have often been touted as useful for studying common diseases due to the likelihood of reduced genetic heterogeneity (Sheffield et al. 1998). Whether they can be expected to exhibit extended linkage disequilibrium as well is unclear and is likely to be dependant on their demographic history.

Whereas rare, mendelian diseases are caused by infrequent variants, one hypothesis of common disease-causing mutations has proposed that variants that contribute significantly to the genetic risk of these diseases may in fact be common within the population (Cargill et al. 1999). This common disease-common variant (CD-CV) hypothesis is supported by the finding of alleles that can account for a significant proportion of the genetic risk for a common disease whilst having an appreciable allele frequency within the population. For example the APO*4E allele accounts for half of the population-attributable risk of Alzheimer's disease whilst existing within the population at a frequency of 10-20% (Roses et al. 1993).

It has been proposed, on the back of the human genome project, that a sufficiently dense map of SNPs be constructed to allow whole genome association studies. Indeed a number of initiatives towards this aim have recently been launched by both large pharmaceutical companies and publically funded institutions. Detection of association between a variant and a disease may be either indirect or direct. The former approach assumes that the SNPs within the dense map are selectively neutral and will merely identify small regions within which lie the disease-predisposing variants. The latter approach suggests that many of the SNPs used in the map may themselves be responsible for significant risk of disease (Cargill et al. 1999; Kruglyak 1999). The CD-CV

hypothesis suggests that if the SNPs that make up the dense map for whole genome association studies are drawn from sequences within and surrounding genes then there is a good chance of picking up the disease-related variants directly.

What is certain is that the construction and use of a dense genome-wide SNP map will require methods for both high throughput screening for and typing of SNPs. I shall consider the mutation detection techniques that may allow such a map to be constructed later on in this introduction.

SNP typing methodologies have traditionally been used on a small scale, often typing a single SNP at a time. Initially methods to detect sequence variation were performed immunologically at the protein level. Obviously such methods underestimated the underlying DNA variation by not detecting synonymous base substitutions. Methods to detect sequence variation at the DNA level initially focused on SNPs that cause changes in restriction enzyme recognition sites and therefore could be detected by Restriction Fragment Length Polymorphism (RFLP). The advent of PCR allowed allele specific amplification to type, in principle, any base substitution. Since then the goal has been to develop techniques to allow the typing of multiple SNPs at the same time. Multiplex PCR has been used in conjunction with a number of detection methods, with a move away from the traditional gel-based methods, to try to increase throughput. One common method is the use of multiplex PCR followed by the immobilisation of the PCR product to a filter which can then be sequentially probed by sequence specific oligonucleotides in a method known as sequence specific oligo-hybridisation (SSO) (Arguello et al. 1996; Comas et al. 1999). A number of more sophisticated sequence variation detection methods have only recently emerged and promise far greater throughput.

A combination of photo-lithographic techniques and solid state chemistry has allowed sequence specific oligonucleotides to be positioned in high density arrays on a 'DNA chip' (Chee et al. 1996). This technology, developed principally by the Californian company Affymetrix, allows the products of multiplex PCR to be hybridised to these arrays which contain sequence-specific oligos for both alleles of multiple SNPs. A specialised confocal microscope 'chip reader' can then be used to score all polymorphisms for which there are oligos on the chip. Typing SNPs is only one of a number of applications of this technology (Chee et al. 1996; Wang et al. 1998).

A number of recent methods for the high throughput typing of SNPs use the principle of primer extension whereby after the initial PCR amplification of a polymorphism-containing amplicon, a primer is annealed directly upstream of the polymorphic site such that the first base to

be added to the primer uses the polymorphic base as template (Hoogendoorn et al. 1999). The extension of the primer can be detected in a number of ways, for example the change in mass of the primer can be detected by HPLC (Hoogendoorn et al. 1999) or Mass Spectrometry (Haff and Smirnov 1997). The advantage of this approach is that the extension reaction is highly robust and a single set of conditions can be used for all SNPs thus allowing easy mulitplexing. In addition many of the detection methodologies can be easily automated.

Mutation detection techniques

Mutation detection of sufficient SNPs on a scale large enough to construct a map dense enough to allow whole genome association studies has never been attempted before, largely due to the lack of sufficiently high throughput mutation detection methodologies. There are a number of different methodologies for mutation detection that can be broadly categorised into two groups; those that rely on the physical changes generated by SNPs and those that recognise and cleave mismatches in duplex DNA or RNA (Cotton 1997). The information output of these two groups of methods is also different. Cleavage methods include positional information as to the site of the mutation whereas most physical methods merely identify whether there is a mutation within a sequence or not. Additional methods that do not fall into either of these two groups are Sanger sequencing and 'DNA chip' technology.

The diversity of mutation detection methodologies is in part due to the different requirements (and budgets) of their users (Cotton 1997). There are a number of considerations when choosing the right method for a given study, some are more important to different groups than others. Cost and throughput are often major factors which have to be considered. The extent to which detection efficiency can be compromised is also often important. For example medical genetics often requires that detection efficiency be as high as possible and high throughput is less important, whereas in evolutionary genetics, as long as a sufficient number of markers can be found throughput may be more important than detection efficiency. An additional factor to consider is whether the sequence of the amplicon is known or not. 'DNA chip' technology is dependant on knowing the sequence of the DNA fragment to be screened. However the eagerly awaited publication of a draft copy of the human genome sequence will soon negate this consideration.

Denaturing Gradient Gel Electrophoresis (DGGE) and Single Strand Conformational Polymorphism Analysis (SSCP) have traditionally been the major mutation detection methods that

rely on physical changes caused by SNPs. DGGE exploits the fact that DNA duplexes differing by a base pair will denature under slightly different conditions. Upon partial denaturation migration of the duplex through a gel is halted. Consequently DGGE relies upon detecting mobility differences between duplex DNA when run on a gel down which there is a denaturing gradient, either chemical or thermal. SSCP utilises the fact that a single base change will cause single stranded DNA to adopt a new folding conformation. Consequently mobility differences should be noted when single stranded products are run on a non-denaturing gel. Whilst the efficiency of SSCP mutation detection can vary widely from study to study, the use of multiple conditions increases the detection efficiency markedly, although it is obviously time consuming (Vidalpuig and Moller 1994). One major factor limiting the throughput of these physical methods is that in order to get reasonable detection efficiencies the sizes of DNA fragments must be limited to about 300bp.

Recently a new physical method has been developed that promises higher throughput and high efficiency of mutation detection. It uses High Performance Liquid Chromatography (HPLC) to resolve differences between heteroduplex molecules under partially denaturing conditions (Underhill et al. 1997). This method is considered in detail further on in this chapter.

Cleavage methods of mutation detection can be subdivided into those that use enzymes such as cruciform resolving enzymes for DNA mismatch recognition and those that use chemical methods to detect mismatches, with different chemistries applied to mismatches involving C and T mispairing (Veitia et al. 1997). Cleavage methods can generally screen larger regions of DNA than physical methods and new enzymes capable of mismatch detection are constantly coming onto the market.

Sequencing is often considered to be the gold standard of mutation detection probably because it is used to confirm putative mutations found by other methods (Cotton 1997). However sequencing does not guarantee 100% detection efficiency. The deficiencies of sequencing with respect to heterozygous positions and compressions is well known. In addition sequencing is often costly and high throughput requires considerable capital investment.

'DNA chip' technology can be used to screen for mutations within a DNA fragment of known sequence (Wang et al. 1998). Overlapping sequence specific oligos for every possible SNP variant within the sequence can be arrayed in defined positions on high density chips. This can be considered as a window of interrogation, shifting a single base at a time, moving along the entire sequence. Differences in the hybridisation patterns of wild type and mutant DNA can be detected with the 'chip reader' and the position and nature of an SNP determined. At present the

commercially available technology is prohibitively expensive, consequently similar systems are being developed in house by various laboratories (R. Villems, personal communication), and so is likely to remain the preserve of well funded large institutions for some time yet.

It is worth pointing out that as well as detecting base substitutions, some but not all of the mutation detection methods described here are capable of detecting small insertions and deletions of a few base pairs and that these also seem to be a relatively abundant source of sequence diversity. Of the 19 novel Y chromosome mutations discovered by Underhill *et al.* using DHPLC, three were mutations of this type (Underhill et al. 1997). Obviously such mutations are less likely to be present in coding regions of genes than base substitutions due to their disruption of the reading frame.

DHPLC

The mutation detection method of Denaturing High Performance Liquid Chromatography (DHPLC) developed from the work of Peter Oefner towards improving pre-existing matrices for the size separation of DNA restriction fragments by ion-pair reversed-phase HPLC. Other forms of liquid chromatographic separation of DNA restriction fragments had previously had difficulty separating fragments of greater than 200bp. Oefner introduced a matrix of nonporous alkylated poly(styrene-divinylbenzene) particles and showed that a resolution comparable to agarose gel electrophoresis of DNA restriction fragments of up to 2kb in size could be accomplished in a few minutes (Huber et al. 1995). Other advantages over competing column matrices include the short equilibration and regeneration times which allow high throughput and the physical and chemical robustness of the matrix which allows many samples to be run on the same column thus giving low cost analysis (Huber et al. 1995).

The basis for size separation of DNA fragments by DHPLC is the use of an ion-pairing reagent to bind the DNA to the column matrix followed by the elution of DNA by an increasing gradient of organic solvent. An ion pairing reagent comprises both a positively charged modality to bind the negatively charged DNA and a non-polar modality to bind the alkyl chains of the matrix. Triethylammonium acetate (TEAA) at a concentration of 0.1M was found to be the ideal ion-pairing reagent (Society 1997). The optimum temperature for size-dependant resolution was found to be 50°C (Huber et al. 1995). Acetonitrile was used as the organic solvent for elution DNA fragments from the matrix.

DHPLC became applicable to mutation detection when it was shown by Oefner and Peter Underhill that at temperatures approaching the melting temperature of a 100-1500bp piece of duplex DNA it was possible to resolve between a homoduplex molecule and a heteroduplex molecule resulting from a single base pair mismatch (Society 1997; Underhill et al. 1997). Single-stranded nucleic acids elute earlier from the matrix than do double stranded nucleic acids. Thus under partially denaturing conditions heteroduplex molecules will elute earlier than homoduplex molecules. Consequently multiple elution peaks (2-4 depending on the amplicon) are observed where previously, at lower (non-denaturing) temperatures, there had only been a single peak. The number of additional peaks depends on whether the two different heteroduplex and homoduplex species can be separated from one another, in which case four peaks can be seen. Resolving between heteroduplex and homoduplex molecules requires that two parameters be optimised; firstly the column temperature must be partially denaturing and secondly the organic elution gradient must be suitably shallow and have the correct start and end points such that the products are eluted in the most sensitive part of the elution profile (Underhill et al. 1997). It has been shown that the window of temperature within which mutations could be detected is about 4°C wide (Society 1997).

DHPLC as a mutation detection technique was pioneered on the Y chromosome as a result of Peter Underhill's interest in Y chromosomal evolution. Prior to the advent of DHPLC, in the previous ten years since the first biallelic polymorphism was found on the Y chromosome only some 15 polymorphisms had been discovered (Jobling and Tyler-Smith 1995) despite some reasonably large sequencing surveys (Dorit et al. 1995; Whitfield et al. 1995). In 1997 a paper was published by Oefner and Underhill, amongst others, detailing the discovery of some 19 new SNPs (Underhill et al. 1997). Since then they have accelerated their rate of discovery and their eagerly anticipated publication of some 200 Y-chromosomal SNPs will revolutionise the field (Peter Underhill, personal communication). Meanwhile attempts to infer population history from Y-chromosomal diversity are limited by the lack of sufficient SNPs to define new lineages. Consequently other laboratories are pursuing their own, smaller scale, SNP screens on the Y chromosome.

It is worth noting that the haploid nature of the Y chromosome means that to detect heteroduplexes, PCR amplicons of the same sequence from two potentially variant individuals must first be denatured and reannealed to generate potential heteroduplexes, whereas at a diploid autosomal locus heteroduplexes result from heterozygosity and so samples from different individuals need not be mixed.

DHPLC has also been applied to medical genetics, high throughput is attractive to those groups studying large linkage intervals containing many candidate disease genes, and the list of publications in which it has been used is growing rapidly (Liu et al. 1998; Ophoff et al. 1996). DHPLC was recently used in a SNP screen of 106 genes related to cardiovascular disease, endocrinology and neuropsychiatry and compared to the competing 'DNA chip' mutation detection technology of variant detector arrays (VDA) (Cargill et al. 1999). This study showed the two technologies having roughly equal efficiency at detecting SNPs, 87% and 85% for DHPLC and VDA respectively.

The efficiency of DHPLC at detecting all SNPs has been estimated in different studies at between 87% and 97% (Society 1997; Cargill et al. 1999). Such a high detection efficiency compares favourably with nearly all other mutation detection methodologies. In addition DHPLC can be used to type known mutations in a relatively high throughput manner and so is unusual amongst mutation detection techniques (excepting VDA). The potential for this typing technology to be multiplexed has not yet been properly exploited.

Although it is possible to adapt existing biocompatible HPLC instrumentation to perform DHPLC mutation detection simply by the addition of a column oven, this is made difficult by the practice of the company holding the sole licence to supply the columns to refuse to support such apparatus. Consequently most people buying into this technology find themselves paying for an entire package of HPLC hardware and software. Whilst initially this seemed unnecessary it appears that to get the most out of the technology, in terms of automation, sensitivity and adaptability, integrated function is preferable.

This chapter describes the use of a 'home-made' DHPLC apparatus to screen a sample of diverse Y chromosomes for SNP variation.

Strategy for the detection of Y-chromosomal single nucleotide polymorphisms by DHPLC

A peltier-cooled column oven was linked to a biocompatible HPLC apparatus to evaluate the ability to use such apparatus for DHPLC mutation detection. Subsequently a panel of diverse DNA samples representing all known Y chromosome lineages would be assembled and a number of Y-specific single copy sequences PCR-amplified from these samples. After annealing to a reference

amplicon from a common DNA sample these samples would be run on the DHPLC apparatus at temperatures determined either experimentally or *in silico*. The resulting chromatograms would be compared to chromatograms from known homoduplexes. Amplicons which gave chromatograms displaying multiple peaks in the putative heteroduplex chromatogram but not the homoduplex chromatogram would be sequenced to confirm the presence of a polymorphism. Once defined such polymorphisms would be further investigated to determine their value in evolutionary studies.

Materials

Buffers

Eluent A - 0.1M TEAA HPLC-grade (Fluka), 5% Acetonitrile HPLC-grade (Sigma), 0.1mM Na_4EDTA (Sigma) in MilliQ water

Eluent B - 0.1M TEAA HPLC-grade (Fluka), 25% Acetonitrile HPLC-grade (Sigma), 0.1mM Na_4EDTA (Sigma) in MilliQ water

Samples

The panel of DNA samples assembled for mutation screening is shown in table 3.1. All major regions of the world are represented, with an intentional bias towards samples from East Asia to detect mutations informative for inferences about the colonisation of the Pacific (see chapter 6). All known Y chromosomal lineages defined by biallelic markers, with the exception of two rare haplogroups (14 and 25) for which samples were unavailable, are represented. In cases where more than one samples is taken from a given lineage, sample with different MSY1 modular structures were chosen. Multiply represented lineages were biased towards those ancestral haplogroups which can be expected to be older and therefore more diverse than their derived haplogroups. All samples were kindly supplied by Mark Jobling and Chris Tyler-Smith.

Primers

PCR primers for the amplicons GMGY6, GMGY26 and GMGY34 were kindly provided by Carole Sargent.

Haplogroup	Sample	MSY1 structure	Population
1	m20	1,3,4	UK
1	m204	1,3,4	Mongolian
1	m456	3,1,3,4	Indian
1	m464	3,1,3,4,3,4	Indian
2	CI198	...4,0,4	Polynesian
2	m119	1,3,1,3,4	Australian Aborigine
2	m98	3,1,3,4	Norwegian
2	m346	3,1,3,4,0,4	Basque
3	m29	1,3+,4	English
4	m211	0,1,0,3,0,4	Bulgarian
4	m219	0,3,1,3,4	Khalkh
5	m27	3,1,3,4	Japanese
6	m47	0,2,0,4,0,4	CAR Pygmy
6	nYCC28	0,4,0,2,0,4,0,2	Tsumkwe
6	nYCC40	0,2,0,2,0,4	Zulu
7	m82	2,0,2,0,4	San
8	m714	0,3a,1a,3a,4a,4	Bantu
8	m73	3a,1a,3a,1a,3a,4a,4	Kenyan
9	m40	3,1,3,4,3,4	Iraqi Jewish
9	m94	3,1,3,1,3,1,3,1,3,4	Portuguese
10	m224	0,3,1,3,4	Mongolian
10	m252	0,1,3,4	Mongolian
11	m250	0,3,1,3,4	Mongolian
12	m300	3,1,3,4	Chinese
12	LGL5176	3,1,3,1,3,4	Finnish
13	m39	3,1,3,4	Chinese
14	n.a.		
15	m622	3,1,3,4	Indian
16	m724	3,1,3,4	Yakut
16	m227	1,3,4	Mongolian
17	m249	n.k.	Mongolian
18	m707	3,1,3,4	Maian
19	m720	0,4,0,4	San
20	m65	3,1,3,4	Japanese
21	m125	1,3,1,3,1,3,4	Berber
21	m6	0,1,0,3,4	Scottish
22	Sp77	1,3,4	Spanish
23	m87	1,3,0,1,3,4	Melanesian
24	m88	3,1,3,4	Melanesian
25	n.a.		
26	m89	3,1,3,4	Cambodian
n.k.	I915	3,1,3,1,3,1,3,4	Indonesian
n.k.	I903	3,1,3+,4-	Indonesian
n.k.	I906	1,3,4-	Indonesian
n.k.	nYCC8	?,4	Zaire
n.k.	m38	3,1,3,1,3,4	Chinese
n.k.	T12	0,1,3,4	Chinese
n.k.	T7	0,3,1,3,4	Chinese
n.k.	m117	3,1,3,4	Nigerian

Table 3.1 - A list of the 47 diverse samples used in the screen for SNPs. n.a. = none available, n.k. = not known.

Methods

Preparation of Eluents

Eluents A and B were made up with MilliQ water, which has been specially purified to be suitable for HPLC and other applications requiring greater than usual water purity, with much reduced levels of both inorganic and organic contamination. Metal ion contamination severely reduces the life of the column matrix. All liquid handling operations were performed using dedicated or sterile disposable equipment so as to minimise contamination from other sources.

Acetonitrile was included in both eluents so as to eliminate the possibility that an error could result in the organic solvent percentage dropping beneath the critical level of 5%, at which damage to the column matrix becomes a risk.

0.1M Tetrasodium EDTA (Na_4EDTA) was made up from powder and filtered through a 0.4μm filter before addition to the eluents. The purpose of Na_4EDTA was to chelate any metal ions present in the HPLC system before contact with the column.

Before being run through the HPLC system both eluents were degassed by being placed in a sonicating waterbath for 10-15 minutes.

Preparation of heteroduplex DNA

47 diverse genomic DNA samples representing all continents and all available known Y-chromosomal lineages were arrayed horizontally in a template plate, in wells A2 to D12, the remaining wells were filled with the reference sample m19. The DNA samples come from different collections and were of varying quality and concentration. The greatest sensitivity of detection of heteroduplex by DHPLC occurs when the two samples being compared are present in equal, large amounts. Consequently the samples were diluted such that amplification gave roughly equal amounts of product from each well. Each PCR reaction was optimised so as to give specific amplification of a single product. In addition amplification on female genomic DNA was done to verify that the product was Y-specific. PCR was performed in 96-well plates with a multichannel

pipette used to transfer template DNA from the template plate described above. A fraction of the PCR product from each well was run on a gel to verify that each sample had amplified to give roughly the same amount of product.

After PCR, rows E to H of the PCR plate were aliquoted into rows A to D respectively and the plate subjected to an annealing program so as to allow heteroduplex formation. This gives 48 aliquots to test, one of which is a homoduplex control (in well A1). The other 47 aliquots are amplicons from each of the diverse DNA samples annealed to the same amplicon from the reference sample. Because each sample was annealed to the same reference sample, m19, a direct comparison of each sample with every other one is not required as they are compared indirectly through the same reference sample.

The annealing program to allow heteroduplex formation was 95° for 3 minutes followed by 95°C for 1 minutes with the temperature decreasing 1°C every minute for 30 minutes and then rapid cooling to room temperature. This annealing program was perfomed by an MJR PTC-200 PCR engine.

Heteroduplex detection

The HPLC apparatus comprised a biocompatible Waters 625LC pump and controller and 490E UV detector. This apparatus was controlled using the Maxima 825 software (Millipore Waters) run on a 286 PC. All wetted surfaces are either titanium or PEEK. To this apparatus was added a Spark Mistral peltier cooled column oven capable to holding temperatures to a 0.1°C accuracy within the range 5-90°C. The flow path was routed from the pump, through the column oven, to the UV detector. The column used for heteroduplex detection was the DNASep column (Transgenomic). A schematic diagram of the apparatus is shown in figure 3.1.

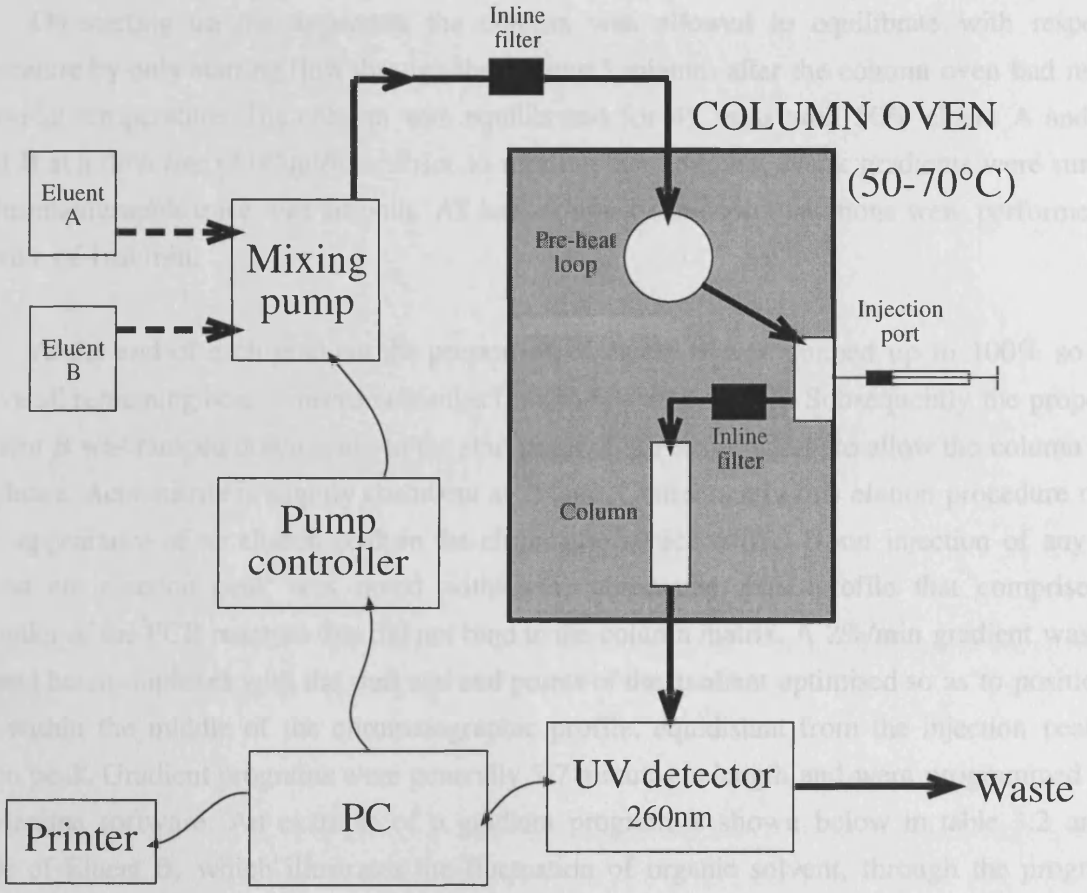


Figure 3.1- A schematic diagram of the DHPLC apparatus used in the screen for SNPs on the Y chromosome. Bold straight lines indicate the flow path, with dashed lines being low pressure and full lines high pressure. Curved lines indicate the electrical connections required to control the apparatus and output the chromatograms.

A 70cm-long eluent pre-heat line was located within the column oven to allow the eluent to reach the correct temperature before reaching the column. This was positioned prior to the injector in the flow path because the sample loop of the injector mechanism was within the oven itself, thus the sample was already at the correct temperature. Two DNA frits (Upchurch) containing 0.5µm PEEK filters were placed within the flow path as per the column manufacturer's instructions so as to protect the column from particulate matter. DNA was detected by setting the UV detector to a wavelength of 260nm.

On starting up the apparatus the column was allowed to equilibrate with respect to temperature by only starting flow through the column 5 minutes after the column oven had reached the correct temperature. The column was equilibrated for 45 mins with 50% eluent A and 50% eluent B at a flow rate of 0.5ml/min. Prior to running any samples, blank gradients were run until the chromatographic trace was smooth. All heteroduplex detection operations were performed at a flow rate of 1ml/min.

At the end of each gradient the proportion of eluent B was ramped up to 100% so as to remove all remaining bound macromolecules from the column matrix. Subsequently the proportion of eluent B was ramped down again to the start point of the next gradient to allow the column to re-equilibrate. Acetonitrile is slightly absorbent at 260nm. Consequently this elution procedure results in the appearance of an elution peak in the chromatographic profile. Upon injection of any PCR reaction an injection peak was noted within the chromatographic profile that comprised the remainder of the PCR reaction that did not bind to the column matrix. A 2%/min gradient was used to detect heteroduplexes with the start and end points of the gradient optimised so as to position the peak within the middle of the chromatographic profile, equidistant from the injection peak and elution peak. Gradient programs were generally 5-7 minutes in length and were programmed using the Maxima software. An example of a gradient program is shown below in table 3.2 and the profile of Eluent B, which illustrates the fluctuation of organic solvent, through the program is shown in figure 3.2.

Time (min)	**Eluent A (%)**	**Eluent B (%)**
0	60	40
1	50	50
5	42	58
6	0	100
7	60	40

Table 3.2 - An example of a typical gradient for heteroduplex detection

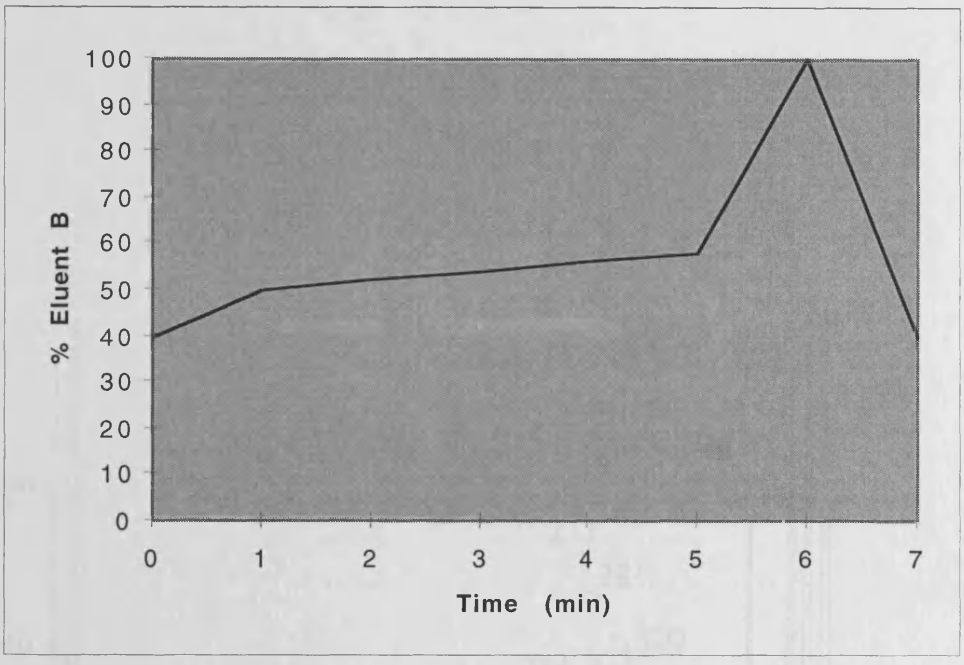


Figure 3.2 - Graph showing the fluctuation in amount of Eluent B passing through the column in the gradient program described in table 3.2

The optimisation of column temperature required to detect heteroduplexes was achieved in one of two ways. When the complete sequence of the amplicon being tested was known, a computer program 'DHPLC Melt' written by Peter Oefner and Nancy Hansen could be used to predict the correct temperature. This program is available on the web at 'http://insertion.stanford.edu'. The program uses a simple mathematical model of the DNA duplex to predict the melting temperatures of 10bp windows across the amplicon and thus suggest a temperature at which to screen for heteroduplexes. If the range of melting temperatures across the amplicon is wide then the program will recommend two temperatures at which to screen. When the entire sequence for the amplicon being tested was not known a second, empirical, method was applied. The elution position of the amplicon was measured on the same gradient over a range of temperatures. A dramatic drop in retention time is noted at the melting temperature of the amplicon. A screening temperature is chosen such that the retention time of the eluted product lies between the two end points of this dramatic shift in retention time.

Running a sample through the apparatus requires that an injection is made and that at the same time the gradient program and recording of the chromatographic profile are initiated. A Hamilton syringe was used to inject 5-10µl of the sample.

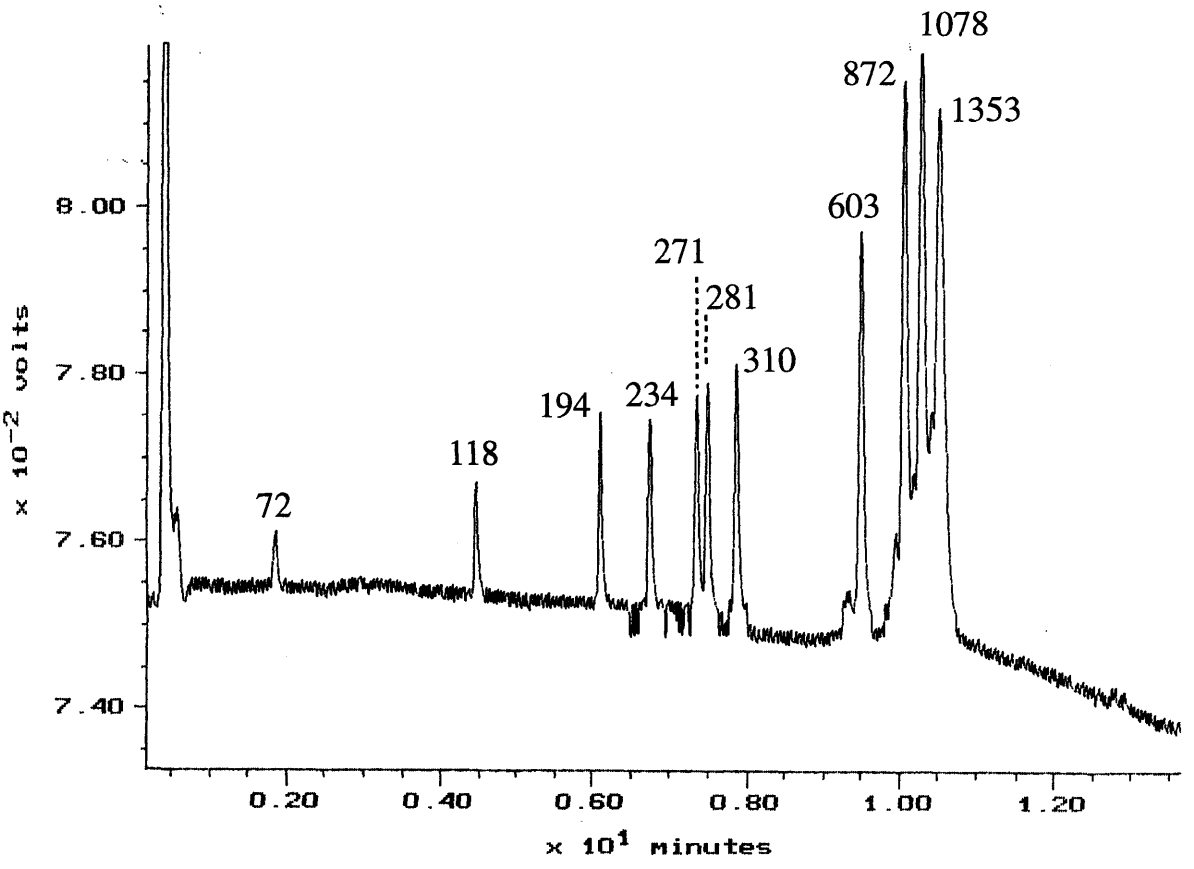


Figure 3.3 - Chromatogram of ϕx174 *Hae*III ladder run through the DHPLC column. Separation is achieved over 12 minutes and at a temperature of 50 degrees Celsius. Numbers above the peaks indicate the size of the relevant fragment in base pairs.

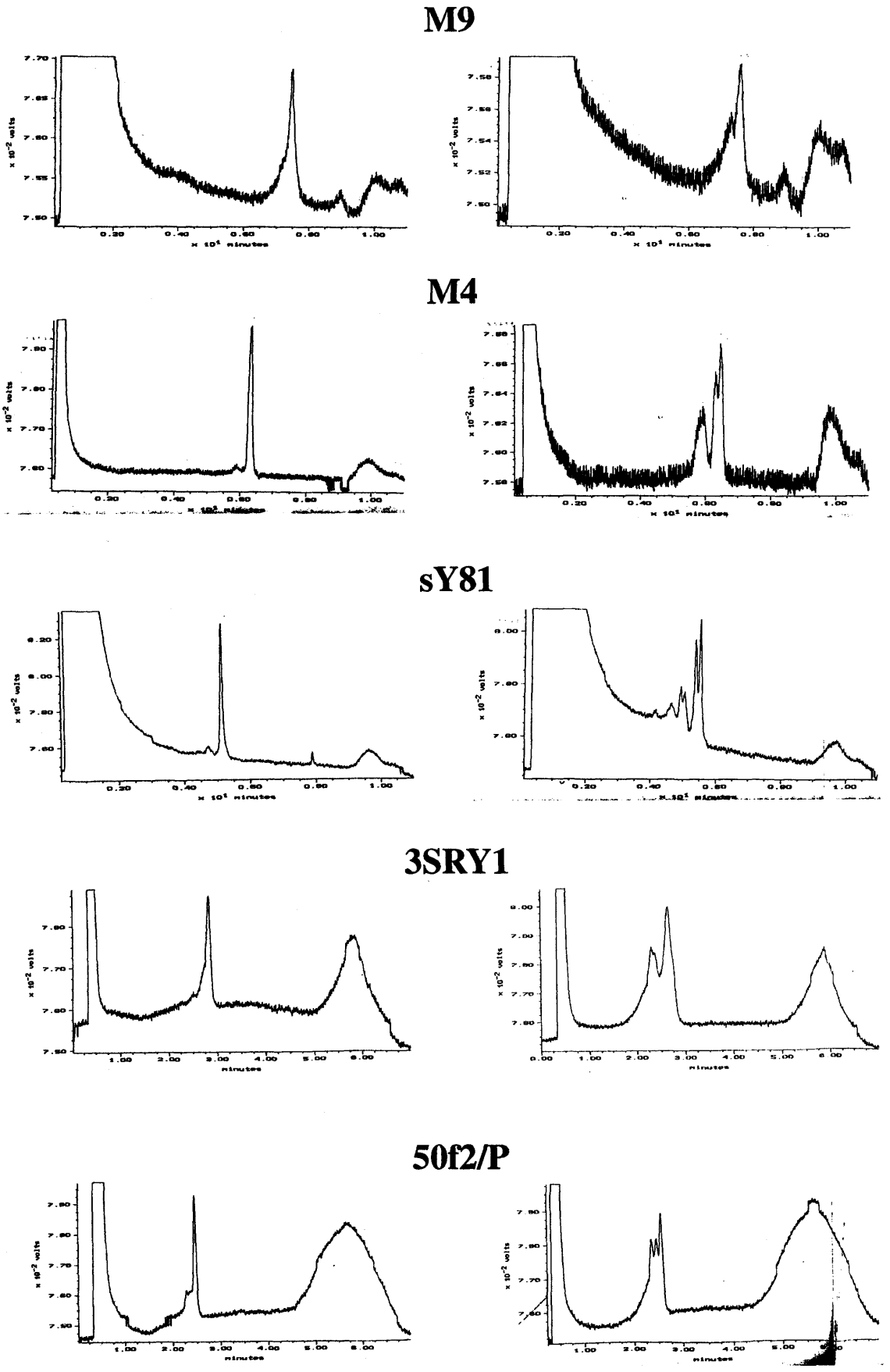


Figure 3.4 - Chromatograms of the five known polymorphisms that could be detected by the DHPLC apparatus. In each case the homoduplex control is shown on the left and the heteroduplex on the right.

Results

Verification of ability to detect single nucleotide polymorphisms using DHPLC

Prior to attempting heteroduplex detection, the ability of our equipment to achieve reasonable size-dependant resolution of DNA fragments was tested. If this could not be achieved then it was highly unlikely that heteroduplex detection would be feasible. Figure 3.3 below shows the results of running an optimised gradient for the separation of a *Hae*III digest of ϕx174 DNA. It can be seen that the column has clearly resolved all of the fragments within the digest, including the 271, 281bp doublet. This resolution is equivalent to that seen in publications and so it was decided to proceed with heteroduplex detection.

In order to verify the ability of the apparatus set up to detect heteroduplex DNA a panel of known base substitutions was assembled for which known positive and negative controls were available. These amplicons were tested in turn on the apparatus. In addition to testing heteroduplex detection those test polymorphisms having complete known sequence were also used to check the accuracy of the temperature prediction program.

Table 3.3 below shows the details of the amplicons that were tested, and whether detection was successful or not. Figure 3.4 shows the differences observed between homoduplex and heteroduplex chromatograms for those polymorphisms that were detectable by DHPLC.

Amplicon	Length	Mutation	Temp. screened	Detected?
M4	273bp	G to C	52°C	Yes
M9	341bp	C to G	54°C	Yes
sY81	209bp	A to G	56°C	Yes
50F2/P	750bp	C to G	59°C	Yes
50F2/I	850bp	C to T	56°C	No
3SRY1P	720bp	C to T	58°C	Yes
Tat	112bp	T to C	58°C	No

Table 3.3 - The amplicons containing known mutations that were used to verify the ability of the apparatus to detect SNPs

Clearly the apparatus is capable of detecting heteroduplex molecules resulting from single base pair mismatches. However two of the seven known polymorphisms tested could not be resolved by the equipment. Consequently it was decided to proceed with a screen for novel SNPs on the human Y chromosome.

For all those amplicons for which sequence was available it was confirmed that the DHPLC Melt program did give a temperature at which the polymorphism could be detected by the equipment. Some of those known polymorphisms that can be detected by the apparatus were used to investigate the temperature range in which a heteroduplex peaking pattern can be observed. It was found that this window of temperatures was usually 3°C wide, slightly narrower that that reported elsewhere (Society 1997).

During these preliminary attempts at heteroduplex detection a couple of improvements were made to the PCR protocol. Firstly, it had been noticed that those amplicons that used 11.1x PCR buffer (Jeffreys et al. 1990) gave significantly larger and more diffuse injection peaks than those using AB PCR buffer. A test of the components of the PCR buffer revealed that this was due to the presence of dNTPs at a four-fold greater concentration in the 11.1x buffer. Thus if all amplicons used the AB buffer, a shorter gradient profile could be used, and the throughput improved. Secondly it was noticed that at the partially denaturing temperature significant shouldering of the homoduplex peak could be observed. If two mismatched amplicons are not present in the same amounts the heteroduplex peak is smaller than the homoduplex peak. Consequently there is a possibility that shouldering of the homoduplex peak could obscure heteroduplex peaks. This shouldering only occurs at the partially denaturing temperature therefore it was hypothesised that this shouldering might be due to the presence of a population of mismatched molecules that result from the relative infidelity of the *Taq* polymerase. The number of cycles used in the preparatory PCR reactions were often quite large so as to insure that amplification from all samples was as equal as possible. When the proof-reading polymerase *Pfu* was added to the PCR reaction mix it was found that this shouldering effect was vastly reduced, thus confirming the hypothesis. *Pfu* was subsequently added to all preparatory PCR reactions, reducing the chance of missing heteroduplex peaks.

Results of a screen for single nucleotide polymorphisms on the Y chromosome

At the time in which this screen was conducted very little sequence information on the Y chromosome had been submitted into the Human Genome Project databases. Those sequences that were available had little information regarding whether they were Y-specific or indeed whether they were present in a single copy. Given that our requirement was for Y-specific, single copy sequences and that the Y chromosome is well known to be full of multicopy and X-Y homologous sequences it was decided not to use these sequences. Instead we focused on sequences that we knew were single copy and Y-specific, and that had not previously undergone systematic searches for polymorphism.

Details of the amplicons screened are shown below in table 3.4.

Amplicon	length	primers (5′ to 3′)	known sequence
3SRY1	715bp	aatcgggtaacattggctaca aggcttaaaagttaataggcca	Yes
3SRY2	793bp	aatggcaatctacttgtttcca gtgatcagatatgcggtgg	Yes
3SRY3	537bp	tctcctaagccaggttagc gagtagaattgtggctacc	Yes
3SRY4	675bp	atactcatctcttcacctgac agagcctgtgttctttattcag	Yes
GMGY6	1100bp	cgtaatcccgcgatcttatg gtagcgcaatgctggatctc	No
GMGY26	331bp	ttcccattactgctgccaaa cagaggacctaaaagccccc	Yes
GMGY34	520bp	tgtcctatcagaatgccctttt ccatcttgtctttcaccacca	No
50f2/P	750bp	caccaccatgtcccacagatttg gtgcatctattgactctttcatgg	No

Table 3.4 - The Y specific amplicons that were screened for mutations by DHPLC

The GMGY amplicons were identified by Carole Sargent in a screen for X-Y homologous sequences, however the amplicons themselves are Y-specific.

The SRY amplicons cover part of the region between the 3′ end of the SRY gene and the pseudoautosomal boundary between PAR I and the non-recombining region of the Y chromosome. These four amplicons are overlapping, so as to screen a contiguous sequence, with primers

designed to avoid the numerous repeated sequences within this region, including *Alu* elements and endogenous retroviral sequences. Although the SRY regions has been extensively screened for infertility-causing mutations (Hawkins et al. 1992; Kwok et al. 1996; Veitia et al. 1997) these studies have focused exclusively on the coding region and 5' flanking regions of the gene.

The 50f2 amplicon was developed by Turi King when converting an SNP within this region from a Southern hybridisation assay (Jobling 1994) into a PCR assay (T.E. King and M.A. Jobling, unpublished observations). This polymorphism had been originally identified by hybridising Y-specific probes to blots of *Taq*I digests of genomic DNA. Thus the sequence within this amplicon had not previously been systematically screened for SNPs.

Two of these amplicons contained known polymorphisms (SRY1: (Veitia et al. 1997); 50f2: (T.E. King and M.A. Jobling, unpublished observations)), these could be used as controls to check both the continuing ability of the column to resolve heteroduplex molecules from homoduplex ones, and that the correct conditions were being used for mutation detection, both in terms of the gradient and the oven temperature.

These amplicons were screened on the set of 47 diverse genomic DNA samples. Any anomalous chromatograms that exhibited signs of heteroduplex peaks were further tested to check the reproducibility of the observations. The remaining portions of the same PCR amplification were run and if these remained anomalous then repeat individual PCRs were done, to verify the reproducibility of these putative heteroduplex patterns. Once this reproducibility had been confirmed the amplicons were submitted for automated DNA sequencing.

Amongst the amplicons screened two produced patterns that seemed suitably reproducible to be submitted for sequencing. Only one of these was subsequently shown to result from a polymorphism. The other showed no sequence variation. It has been suggested that multiple peaking patterns can occur in samples not exhibiting sequence variation, causing false positives, if two differing conformations of equal energy can be adopted by an amplicon under partially denaturing conditions (Mike Hammer, personal communication), although such amplicons are likely to be rare.

The 50f2(P) amplicon contains a known polymorphism that creates a *Taq*I site. This was used to ensure that the amplicon was being screened at the correct temperature. During screening a multi-peak pattern was noticed that was different from either the homoduplex pattern or the peaking pattern of the heteroduplex produced by the known SNP. The three peaking patterns are shown in

figure 3.5. The putative SNP occurred in a chromosome from haplogroup 7, which has only been found in the San population of southern Africa (Hammer et al. 1998).

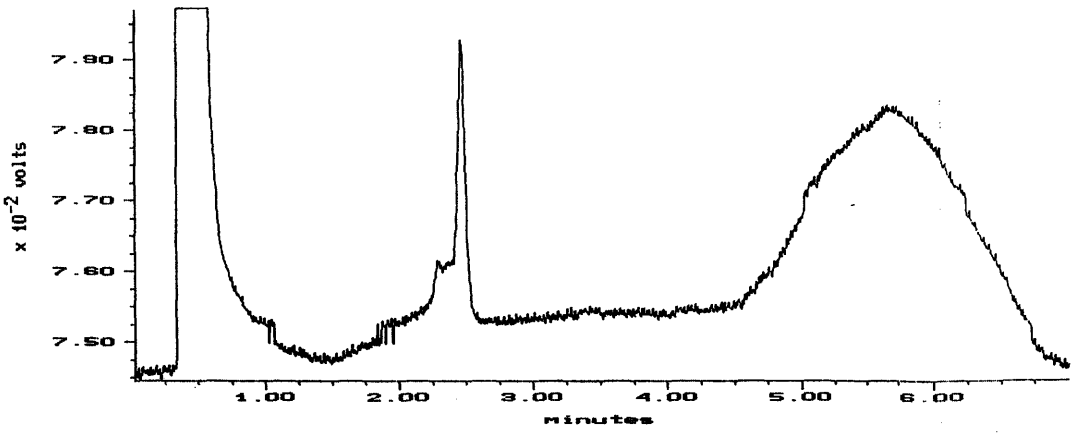
After sequencing from both ends in all three types of amplicon; reference, known SNP and putative SNP, it was shown that the putative SNP was in fact a C to G transversion. The full sequence can be found in appendix A. This polymorphism was named MEH1. The lineage defined by MEH1 was named haplogroup 27.

Investigation of the single nucleotide polymorphism MEH1

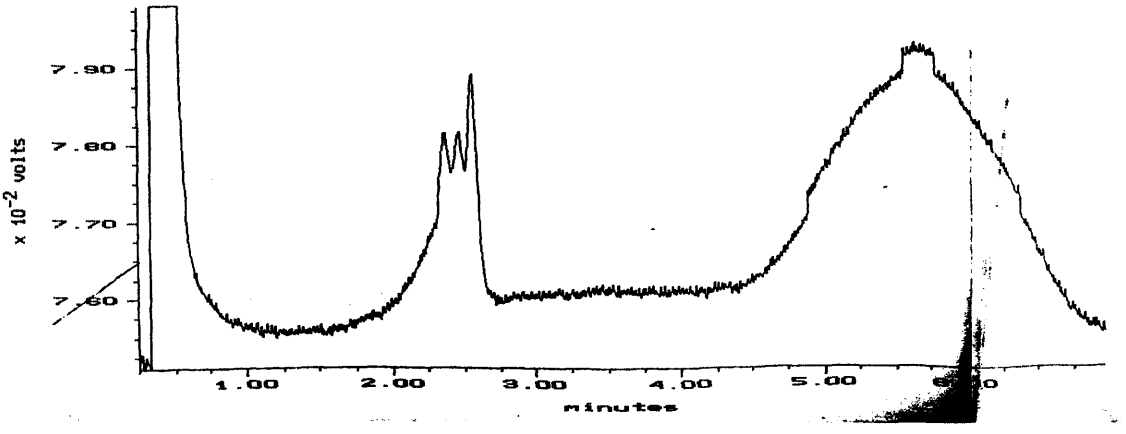
The SNP MEH1 neither creates nor destroys a known restriction enzyme site, therefore primers were designed that could type the SNP by allele specific amplification. One of the primers used for the original amplification of the 50F2/P amplicon, TEK B, was used as a universal primer and two primers specific for the different states of the polymorphism were designed, the sequence of these primers is given below.

MEHA - 5′ CTC AGA ACT GTG AAA CAT GAT CTG 3′
MEHB - 5′ CTC AGA ACT GTG AAA CAT GAT CTC 3′

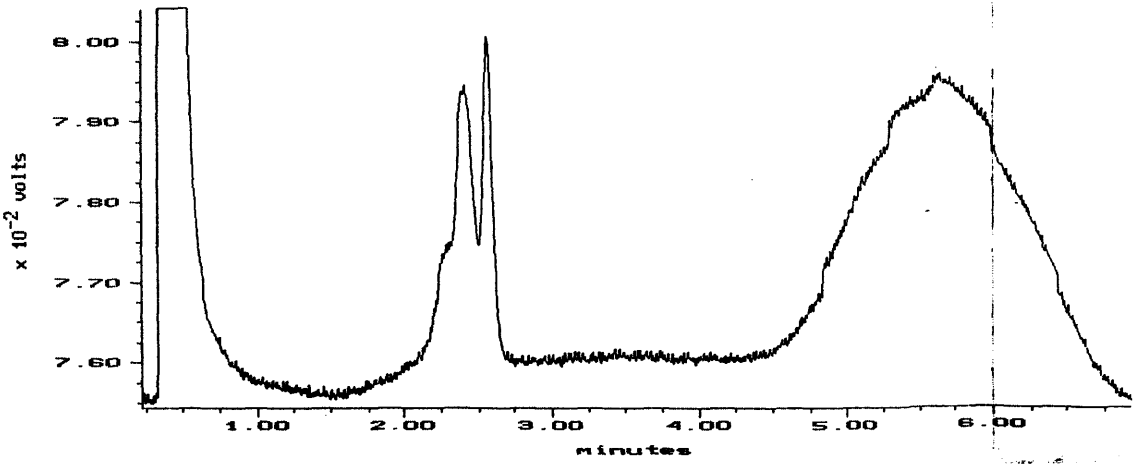
The annealing temperature of this amplification was titrated such that a product was only obtained when the specific primer was included in the reaction. The optimum annealing temperature was found to be 64.5°C. The assay was typed on two chimp samples and a single gorilla sample, to determine the ancestral state of the polymorphism. A PCR reaction designed in one of humans, chimps and gorillas can often be be used in the others, due to the high sequence homology between these three species. The gel showing this information is shown below in figure 3.6.



Homoduplex



Known Heteroduplex



Putative mutation

Figure 3.5 - Chromatograms showing the differences between the peaking pattern produced by the putative mutation compared with those of both the known mutation and the homoduplex control. The separation was achieved on a 7 minute gradient at 59 degrees Celsius.

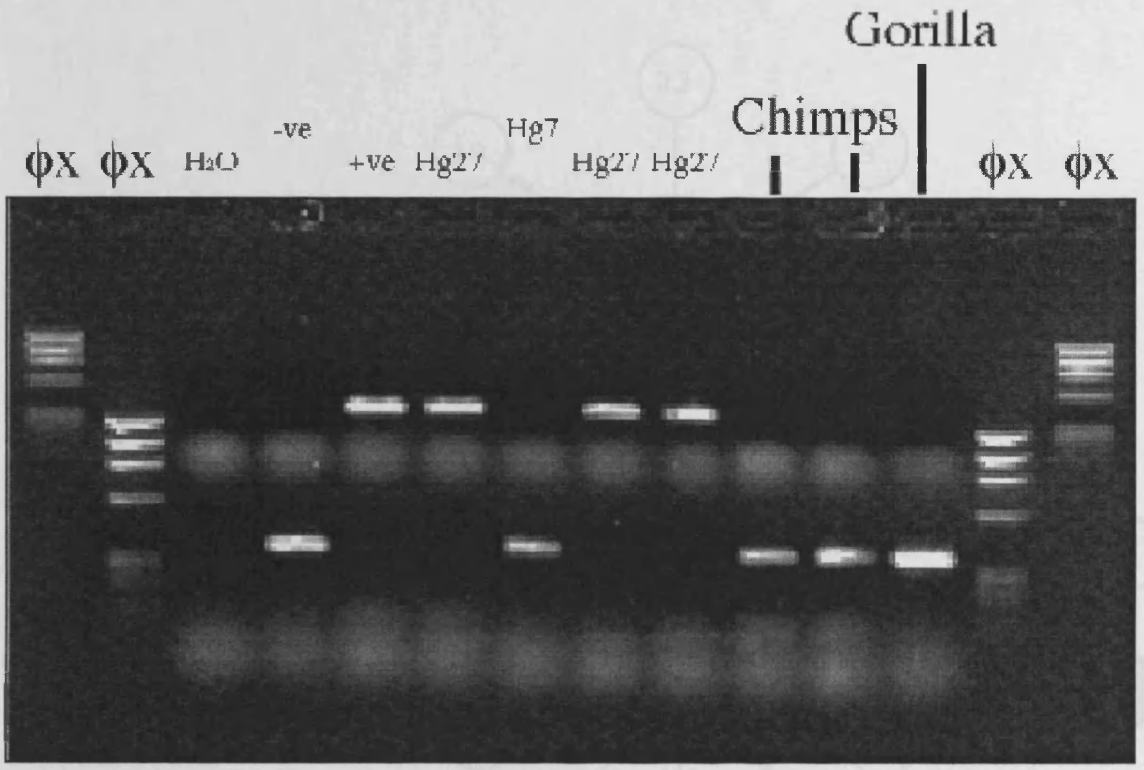


Figure 3.6 - Agarose gel showing the ancestral state of the MEH1 polymorphism. Two allele specific amplifications were performed on each sample using a primer specific for each allele of the polymorphism. All products from the amplification specific for the C allele were loaded first, followed 10 minutes later by the loading of products specific for the G allele, with the same samples being loaded in the same wells. Thus samples which show only the lower band indicate the presence of the C allele whereas those sample with only an upper band indicate the presence of the G allele. The symbol ϕx indicates the loading of a ϕx174 *Hae*III marker ladder. The inner ladders were loaded with the first batch of products whilst the outer ladders were loaded with the second batch of products.

All three primate samples share the same allele, the C allele, indicating that this is the ancestral state of this SNP. Thus haplogroup 27 must be derived from haplogroup 7 and not vice versa. The precise position of the lineage defined by the MEH1 SNP could be ascertained because all other subgroups of haplogroup 7 contain the ancestral state of this SNP. The single known chromosome belonging to haplogroup 19 was present in the original screening panel and did not exhibit a heteroduplex peaking pattern. The position of the lineage defined by this new polymorphism within the tree of Y-chromosomal haplogroups is shown below in figure 3.7.

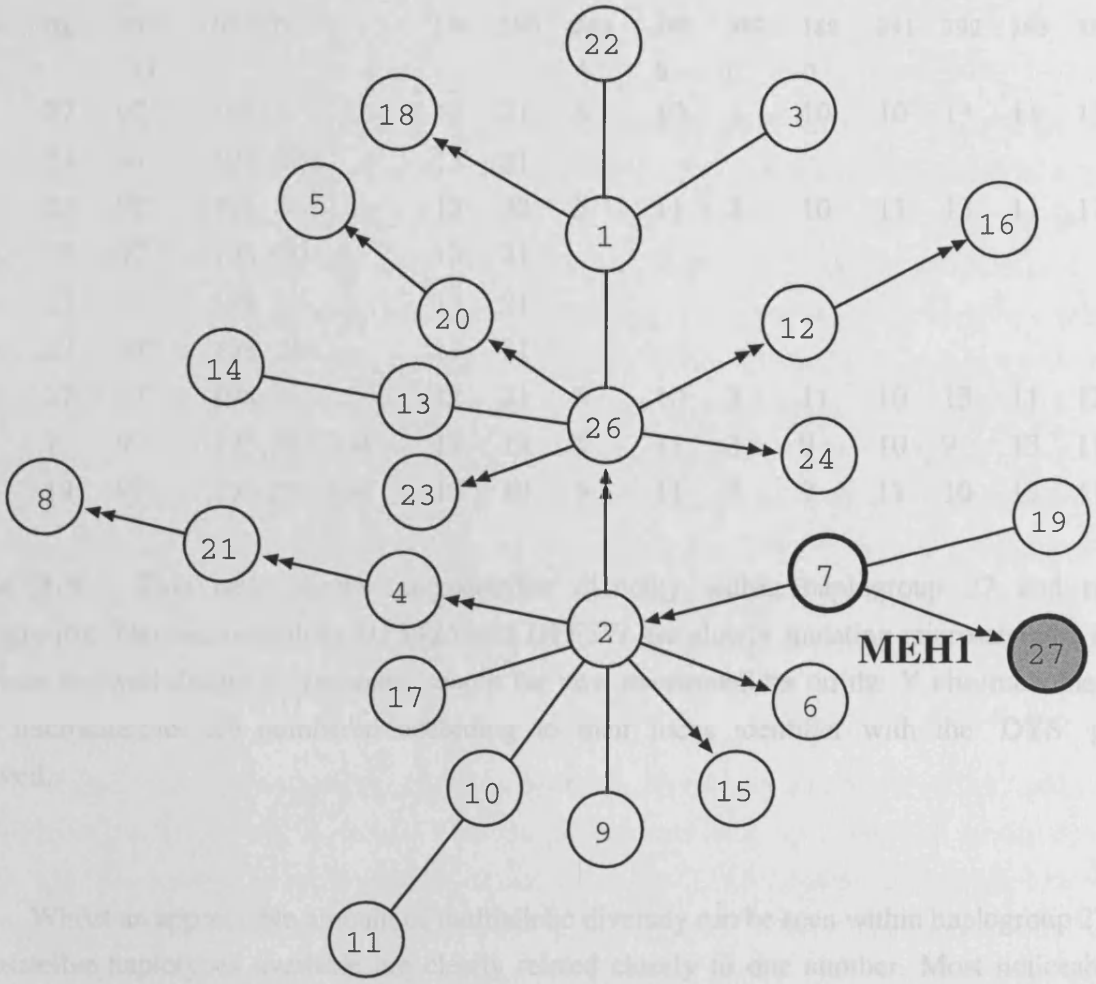


Figure 3.7 - A maximum parsimony tree of Y chromosome haplogroups. Circles are haplogroups and the lines separating them are single polymorphisms. The direction of the arrows indicates the ancestral state of each polymorphism where known. The new lineage defined by the MEH1 base substitution, haplogroup 27, is shaded.

All seven chromosomes previously classified as haplogroup 7 were typed for the MEH1 SNP and it was found that all but one share the derived allele of the MEH1 polymorphism and therefore belong to haplogroup 27. There was additional information available about many of these chromosomes in the form of microsatellite haplotypes from Neale Fretwell and Peter de Knijff. The typing and diversity information is summarised in table 3.5. All chromosomes detailed here are from the San population of southern Africa. Microsatellite data about the single known haplogroup 19 chromosome is included in this table for comparison.

Sample	Hg	*DYS 425*	*DYF371*	19	390	389 A	389 B	389 C	389 D	391	392	393	388
m82	27	97	198	12	21	5	10	3	10	10	13	11	12
m717	27	97	198, 201	13	21								
m718	27	97	198	13	22	5	11	3	10	11	13	11	12
JR046	27	97	198, 201	13	21								
JR060	27	97	198	13	21								
JR065	27	97	198, 201	13	21								
JR323	27	97	198	12	21	5	10	3	11	10	13	11	12
m708	7	97	195,201,204	17	19	5	11	3	9	10	9	13	11
m720	19	97	195,201,204	17	19	5	11	3	9	11	10	13	11

Table 3.5 - This table shows microsatellite diversity within haplogroup 27 and related haplogroups. The microsatellites *DYS425* and *DYF371* are slowly mutating microsatellites found by Neale Fretwell during a systematic search for new microsatellites on the Y chromosome. The other microsatellites are numbered according to their locus identifier with the 'DYS' prefix removed.

Whilst an appreciable amount of multiallelic diversity can be seen within haplogroup 27, the microsatellite haplotypes available are clearly related closely to one another. Most noticeably all haplogroup 27 chromosomes share unusually short alleles at the *DYS19* locus. This would seem to indicate that this lineage is not particularly ancient. There are not sufficient numbers of these chromosomes to allow us to do calculate any meaningful statistics on these lineages to further investigate population history.

Discussion

In this study 5.4kb of Y-specific sequence were screened by DHPLC in 47 samples. Allowing for the failure of some samples to amplify, this gives a total amount of sequence screened for SNPs in the region of 230kb. In 1995 a paper was published in Science in which 729bp of the first *ZFY* intron was sequenced in 38 individuals. No sequence variation was found (Dorit et al. 1995). Over eight times that amount of sequence has been screened here, indicating the high throughput nature of this novel mutation detection method. Such a high throughput method is particularly applicable to the Y chromosome due to its low levels of sequence polymorphism.

A rough figure for the efficiency of mutation detection of the DHPLC apparatus can be estimated from the screening of amplicons containing known SNPs. Five out of seven control mutations were found using DHPLC giving an efficiency in the region of 71%. This figure is lower than those published in recent literature for this mutation detection technique. There are many possible reasons for this discrepancy.

Firstly the equipment being used to perform the screening was relatively old and for the previous few years had been in storage. This resulted in number of problems. The sensitivity of the UV detector was lower than more modern equipment, at low DNA concentrations peak broadening occurred due to background noise in the chromatographic trace. Consequently the number of cycles of PCR had to be pushed to a level where misincorporation of nucleotides, even in the presence of *Pfu*, caused shouldering of the homoduplex peak. The pumping system of the HPLC apparatus was also prone to inconsistency causing variation in the column pressure which again affected the resolution of the chromatogram.

Secondly the screening of the Tat and 50f2/I SNPs was performed before the addition of *Pfu* to the PCR mix, therefore peak shouldering may have obscured the heteroduplex peaks. There was not sufficient time to repeat the screening under the new PCR conditions.

Finally the screening of the known SNPs was performed on a different column to that which was used for the actual mutation screen. This was because the first column, on which the known SNPs were screened, became unusable after only 300 injections as a sufficient flow rate could not be obtained without exceeding the pressure limits of the system. These columns are meant to last for at least 4000 injections. The second column is still in operation having performed more than 1500 injections. This suggests either a problem with the first column or that the HPLC

equipment was initially leaching material that clogged the first column irreparably. It was only after the failure of the first column that the tetrasodium EDTA was added to the eluents. The inclusion of this chelating agent is considered optional in the column manufacturer's instructions. It may have been that metal ion contamination shortened the life of the first column. Certainly there was a rapid deterioration in the attainable resolution with this column prior to becoming completely unusable. This drop in resolution may have affected the ability to detect heteroduplexes. All the known SNPs that were screened on the second column could be detected, suggesting that the above figure for efficiency of mutation detection may well be an underestimate for the screening for unknown SNPs conducted on the second column.

Other studies of sequence diversity on the Y chromosome have found mutations at an average of one every 1200bp (Underhill et al. 1997), thus we might have expected to find of the order of four mutations within the sequences screened here. Therefore to have found only a single new SNP is a disappointment. It is entirely possible that there is only a single new polymorphism within the sequences screened here. Alternatively, as with all mutation detection techniques, it is possible that SNPs were missed by the screen. Despite best efforts, equal amplification from all genomic DNA templates was not possible, so large was the variation in quality of the DNAs. This might well have lead to some heteroduplex molecules being missed due to non-stoichiometric amounts of the two amplicons. Were this screen to be repeated it might be wise to choose DNA samples more similar in quality, even if this sacrificed some of the diversity within the screening set.

Many of the problems associated with the equipment used in this study would be ameliorated by the use of HPLC apparatus specially designed for DHPLC mutation detection. Such apparatus has recently arrived on the market and offers integrated function of the modular systems, greater resolution and comes with software specifically designed for this application. In addition extra modules such as automated sample injection allow for a much greater throughput of screening. The future of this technology seems reasonably assured with its recent adoption and good performance in an important paper from an influential laboratory (Cargill et al. 1999).

MEH1 defines a new Y-chromosomal lineage, haplogroup 27, that is specific to the San population of southern Africa. Haplogroup 27 shows reasonable levels of diversity in terms of multiallelic variation though is obviously not an ancient lineage. Haplogroup 7 is ancestral to haplogroup 27 and exhibits a similar population specificity (Hammer et al. 1998). The vast majority of chromosomes within our sample collection previously thought to belong to haplogroup 7 actually belong to haplogroup 27 suggesting that a small long term effective population size or

relatively recent population bottleneck has occurred in this population so as to cause the derived allele to drift to higher frequency than the ancestral form. Both scenarios are consistent with the known history of this population. The San once occupied a much larger region than they presently do. The reduction in this area and size of the population was caused by the agriculture-driven expansion of Bantu-speaking peoples throughout Africa within the past 6000 years (Cavalli-Sforza et al. 1994). The severity of the population bottleneck that may have accompanied these changes is unknown. The present day size of this population is thought to be 55,000. The San have always been hunter-gatherers and so have not been through the population expansion associated with adopting agriculture (Cavalli-Sforza et al. 1994), consequently it is likely that they have had a relatively small constant long term effective population size. The small sample sizes within this survey means that it would be unwise to extrapolate too far from these results. However the population history of this group deserves greater investigation.

Chapter 4: Lineage analysis: how should Y-chromosomal diversity be analysed?

Introduction

Population-based comparisons using summary statistics developed into the major analytical tool for allele frequency data but are not the most informative way to analyse non-recombining portions of the human genome (Richards et al. 1997). Within such a genomic region, an allele at one polymorphic site can only be released from linkage to an allele at another polymorphic site by mutation. Consequently if a polymorphism occurs at a sufficiently low frequency that it can be considered to be unique, then it is forever linked to all other such sites within the non-recombining region (Jobling and Tyler-Smith 1995). Each unique marker defines a lineage and combining such markers into compound haplotypes produces a unique phylogeny of lineages (Jobling and Tyler-Smith 1995). Each lineage is founded by a single chromosome and thus is initially linked to a single haplotype of markers, composed of both unique and recurrent polymorphisms.

It has been increasingly recognised that the best way to analyse a non-recombining region is to take an explicitly genealogical rather than population-based approach (Avise et al. 1987; Richards et al. 1997; Barbujani 1999). Thus the lineage substitutes the population as the unit of investigation. A phylogeny of unique intra-specific lineages is more in keeping with traditional phylogenies of species than is a population tree, due to the lack of horizontal transmision (gene-flow) between taxa. Lineages and species can reasonably be expected to have unique origins whereas to produce a tree of poorly-understood populations is to impose a fission process that may not be appropriate in many cases.

As outlined in the introduction, Y-chromosomal polymorphisms are generally used in one of two ways to investigate human evolution, either to focus on a single lineage and its distribution and diversity throughout the world or to focus on a certain subset of populations and investigate all the lineages within them. Both approaches are explicitly genealogical. The emerging nature of Y-chromosomal markers has meant that the former approach was adopted in early papers (Seielstad et al. 1994; Underhill et al. 1996; Zerjal et al. 1997) but that as more markers become known the latter approach can be expected to predominate (Hammer et al. 1998). The finding of a new, unique, biallelic marker on the Y chromosome no longer guarantees a publication.

If the ancestor of modern humans was many millions of years old and the effective population size much larger, SNPs on the Y chromosome mutating at a rate of roughly 10^{-7} per site per generation would cease to be 'unique'. This would reveal itself through homoplasies on a tree constructed from such markers. This reveals one of the strengths of the genealogical approach; that phylogeny construction allows inferences on the information content of the underlying polymorphisms (Jobling and Tyler-Smith 1995). Phylogenies constructed from seemingly unique markers can cope with a limited amount of homoplasy as reticulations can be resolved through determining the ancestral state of the markers in question and by incorporating more markers into the analysis (Hammer et al. 1998).

Whilst the genealogical approach is applicable to any non-recombining region of the genome, including mtDNA and autosomal haplotypes, the wealth of polymorphic markers of different mutation rates and dynamics on the Y chromosome means that much information can be gleaned about a single lineage (Jobling and Tyler-Smith 1995).

Any genealogical approach to Y-chromosomal diversity, whether it focuses on a lineage or populations, relies on the same basic analytical tools to make inferences from lineages. This chapter is concerned with the development of these tools and their correct application. It uses two studies, each focusing on a single lineage, to show what sorts of inferences can be made and to illustrate the power and anthropological relevance of these inferences.

Identifying a lineage

Whilst delineation of lineages has relied mainly on 'unique' markers, even the fastest, most recurrent polymorphisms on the Y chromosome contain useful phylogenetic information.

The label 'unique' really needs to be considered within an explicit temporal setting. For example; the time depth within a small pedigree is so recent as to allow a single microsatellite mutation to be considered as being 'unique' within that pedigree and therefore capable of defining two distinct lineages within it (Foster et al. 1999). In addition, base substitution markers on the Y chromosome can not be considered to be 'unique' in an absolute sense, for the following reason. Multiplying the mutation rate of each nucleotide by the number of males in the present global male population reveals that each nucleotide on the Y chromosome must be multiply substituted in the

sample of all extant males. Thus the label 'unique' is a relative term, as it contains a consideration of the extent of sampling of the entire intraspecific diversity. It is worth noting that the precise insertion of *Alu* elements between any two specified nucleotides occurs at such a low rate that these mutations really can be considered to be unique within human evolution, with no recourse to considerations of sampling (Hammer 1994).

Unless otherwise indicated, 'unique', as used throughout this thesis, is defined pragmatically as a marker that, given the level of sampling apparent in studies of human populations, can reasonably be expected to define a monophyletic lineage within the entire time depth since the most recent common ancestor of all extant human Y chromosomes. Markers on the Y chromosome that can be used to define lineages in this way are base substitutions and some, though not all, indels (Hammer 1994; Jobling et al. 1996). Such markers define the same subset of chromosomes whether used in conjunction with other markers to define haplotypes or not.

The limited present resolution of lineages achievable using SNPs and indels alone has resulted in a number of studies attempting to use recurrent markers to define sublineages within lineages defined by other markers e.g. (Santos et al. 1995). At its simplest this involves subdividing a monophyletic lineage defined by unique markers into sublineages delineated by single microsatellite alleles (Ciminelli et al. 1995). This is highly contentious given that the high mutation rate of microsatellites means that few, if any, microsatellite alleles can be considered as unique events within a lineage, though obviously it depends on the time depth of a lineage. Although a lineage is little more than a very large pedigree, the time depth at which a lineage can truly be considered as such (with reference to the use of a 'unique' microsatellite mutation within a pedigree given above) is likely to be very short. Most lineages defined in published studies include chromosomes with microsatellite alleles shared by identity by state in appreciable proportions (Peter de Knijff, personal communication). Consequently any attempt to analyse sublineages defined by single microsatellite alleles is likely to introduce significant 'noise' through the confusion of identity by descent with identity by state (Cooper et al. 1996).

Compound haplotypes of microsatellite alleles contain far more phylogenetic information than do individual alleles (Cooper et al. 1996; Roewer et al. 1996) and are more likely to be identical by descent than identical by state. However diversity within most human populations is so high that to define sublineages on the basis of individual haplotypes would be meaningless, with too many different haplotypes, often private to individual populations, to be of any use in population comparisons. Consequently a number of recent studies have attempted to define sublineages on the basis of a grouping of related compound microsatellite haplotypes (Thomas et

al. 1998; Malaspina et al. 1998). At present these groupings are quantitatively-deficient, often being performed by inspecting networks of haplotypes within a lineage defined by unique markers and defining groups of haplotypes that seem more closely related to one another than to other haplotypes outside the grouping.

This approach has been used to define a lineage within a population of Jewish priests, a profession that is inherited patrilineally, by identifying a 'modal cluster' of haplotypes one mutational step from a predominant root haplotype (Thomas et al. 1998). A related approach has been used to subdivide lineages within Europe defined using 'unique markers' into groups representing 'networks of adjacent haplotypes' (Malaspina et al. 1998). These latter networks also contain multiply-represented, central, haplotypes, but are not restricted to haplotypes that are only a single step from the central haplotype. Both of these studies subsequently sought to use the diversity within the sublineages to date their origin, with little attempt to address the circularity problem; namely using diversity to date a lineage previously defined by its diversity, or rather lack of it.

Assaying diversity within a lineage

Multiallelic markers with high mutation rates can be used to investigate diversity within a lineage (Jobling and Tyler-Smith 1995). The markers most often used for this work are microsatellites; however, the minisatellite MSY1 provides an independent locus which despite its being the most hypervariable locus on the Y chromosome has as yet been little used for this purpose (Jobling et al. 1998).

There are a number of confounding factors that contribute to the degree of multiallelic diversity within a lineage. Whilst the age of the lineage obviously directly affects the apparent diversity, the confounding effects of population history, and specifically demographic history, must also be considered. The effects of population expansions and bottlenecks on lineage diversity have been investigated in greatest depth (Rogers and Harpending 1992; Martinson et al. 1993; Harpending et al. 1998).

Multiallelic diversity within lineages has generally been analysed in two complementary ways, firstly by constructing a graphical display of the apparent diversity in an attempt to reveal any structure within the lineage, and secondly by quantifying the diversity within a lineage (Jobling and

Tyler-Smith 1995). This has often been directly related to the age of the lineage by a variety of means.

These two types of analyses are considered separately below.

Displaying diversity graphically

Attempts to construct graphical displays of diversity are driven by a desire to tease patterns out of the data that may not be obvious on simple inspection. The major problem with constructing graphical displays of such diversity is that each haplotype can only be metrically positioned relative to all others in multidimensional space. The human mind is apparently incapable of comprehending more than three dimensions. Consequently a method is needed to reduce this multidimensional space. Such methods can be broadly be classified into two classes: those that seek to identify the evolutionary important links between haplotypes, and those that summarise the information in a small number of dimensions (usually 2 or 3) with the minimum loss of information (Sokal and Rohlf 1994). The former class of methods imposes a bifurcating pattern of connections between haplotypes, whereas the latter does not.

The attempt to construct phylogenetic trees of multiallelic haplotypes falls into the first class of methods. A number of different methods of tree construction have been tried (Bergen et al. 1999) . A common finding is that there are a large number of equally parsimonious trees. The study by Roewer and others (1995) found a huge number (10^{27}) of equally parsimonious trees describing the microsatellite diversity at seven loci within two closely related populations. The likelihood that any single tree of this set of equally parsimonious trees represents the real evolutionary relationship of these haplotypes is very low. Consequently much attention has focused on the use of networks which include reticulations to represent a set of equally parsimonious trees (Bandelt et al. 1995; Roewer et al. 1996; Santos and Tyler-Smith 1996). A number of different algorithms have been applied to network construction, varying in complexity from the computationally intense to those that can be done by hand.

Probably the simplest method of network construction used has been that of joining haplotypes connected by single mutational steps (Cooper et al. 1996; Malaspina et al. 1998). The central haplotype for such a network is usually that which is most represented in the dataset. This network method has been applied to datasets solely consisting of microsatellite data (Cooper et al.

1996) and to those that incorporate unique markers by only joining microsatellite compound haplotypes belonging to the same lineage (Malaspina et al. 1998). This method usually results in some of the haplotypes not being represented in the final networks. The study that did not incorporate lineage information utilised data from five microsatellites to construct a single network that contained only 79% of the haplotypes (Cooper et al. 1996). The study that used a number of unique markers in addition to four dinucleotide microsatellites constructed 6 networks that contained 95% of all haplotypes (Malaspina et al. 1998).

Another method of network construction is the minimum-spanning (M-S) network. This is based on Kruskal's algorithm for network construction which initially connects all haplotypes linked by steps of unit distance before proceeding to steps of progressively greater distances, connecting any haplotypes not previously incorporated into the network, until all haplotypes are included in the network and all equally parsimonious steps are included (Zerjal et al. 1997). The assumptions of the single stepwise mutational mechanism are implicit in this method. Only observed haplotypes can be used as nodes in this method of network tree construction. However, given this assumption the network includes all most parsimonious trees. Although programs exist to calculate the links of an M-S network, network construction can be done by hand once a matrix of pairwise mutational step distances has been calculated between all haplotypes.

Recently a new method of network construction has been introduced which produces networks containing all most parsimonious trees of a given dataset, without recourse to the 'observed-haplotypes-as-nodes' assumption (Bandelt et al. 1995; Bandelt et al. 1999). Dropping this assumption produces trees more parsimonious than does M-S network construction. These Median-Joining (M-J) networks use a branch of graph theory known as median vectors to identify unobserved nodes that shorten the overall length of the tree. Performed by a freely available computer program this analysis also provides a number of criteria that can be invoked to reduce the number of reticulations within the network, thus making it more tree-like (Bandelt et al. 1999). These include weighting loci differently according to their mutation rate, and therefore likelihood of a recurrence, and including frequency criteria to identify more likely connections within a reticulation. Varying certain parameters can control the degree to which these networks can be reduced. Reduced M-J networks have been used in a number of mtDNA studies to cope with the homoplasies inherent in sequence data from the hypervariable sequences of the human mitochondrial genome (Richards et al. 1996; Forster et al. 1996).

There are two types of multivariate analyses that can be used with multiallelic haplotypic data, and which have been commonly used to represent the relationships of populations in a non-

bifurcating manner e.g. (Poloni et al. 1997). Both seek to reduce the dimensions of the data but utilise different methods for minimising information loss. Principal Components Analysis (PCA) seeks to find a number of independent axes within the multidimensional space of the observed data, which contain maximal variance of the original data set (Cavalli-Sforza et al. 1994; Sokal and Rohlf 1994). The first axis chosen contains the most variance and the second axis the next most and so on. These axes can then be plotted orthogonally against one another to form 2D or 3D representations of the data set. PCA imposes certain restrictions on the data and a related analysis of Principal Co-ordinates Analysis is more often applied to genetic data (Sokal and Rohlf 1994). PCA has long been used to compare populations but has only more recently been used to relate haplotypes (Alan Redd, personal communication). Data from PCA can also be used to feed into other analyses that require reduced-dimensions summaries of the multivariate information within the data set (Cavalli-Sforza et al. 1994). An example of this is given in the next chapter when considering genetic landscapes.

Multidimensional Scaling (MDS) is applicable to a wider variety of research problems than is PCA by virtue of imposing fewer restrictions on the input data (Sokal and Rohlf 1994). MDS seeks to rearrange the haplotypes in a dimensional space imposed by the operator in such a way that best approximates to the observed distances between the haplotypes. A 'stress' measure is used to numerically compare different arrangements of haplotypes with the observed data and the arrangement with minimal stress is chosen as the final output.

The superimposition of additional information upon graphical displays of multiallelic diversity can be used to identify relationships that, after further testing for significance, can allow additional inferences to be fed into further analysis (Zerjal et al. 1997). Superimposing population affinity on such displays can allow population structuring of haplotypes to be easily seen which can then be confirmed statistically.

Dating lineages

Although this section is explicitly concerned with the use of intra-allelic diversity to date the age of a lineage, this is not the only method available to achieve this. Coalescent analyses have also been used to date the ages of alleles at many loci (Harding et al. 1997), including the Y chromosome (Hammer et al. 1998), without any intra-allelic diversity data. Although the tree structure is taken into account, essentially the frequency of an allele determines its age to a high

degree. Such coalescent estimates often come with very wide confidence limits as a result of the stochastic nature of the simulation process; the lower the effective population size of the population, the greater the stochastic effect and the wider the confidence limits. Coalescent analyses can also use intra-allelic diversity to calculate the age of a lineage (Wilson and Balding 1998) and this will be considered later in this chapter. Little, if any, work has been done to compare dates calculated from intra-allelic diversity with coalescent dates for alleles based on frequency.

A number of different statistical methods have been used to date lineages from intra-allelic diversity; these depend to a large extent on modelling the manner in which various statistics calculated from multiallelic data can be expected to increase over time, assuming a single step mutation model. Thus simulation experiments have been particularly useful in this regard. The large confidence limits around most age estimates are due largely to the uncertainty in the mutation rates of the loci being assayed, although some studies have also sought to investigate uncertainty due to different demographic scenarios (Thomas et al. 1998).

Perhaps the simplest dating method rests on the assumption that the mean number of mutations from a root haplotype is equal to the product of the mutation rate and the time since the first appearance of the ancestral haplotypes. This relationship was developed from that first hypothesised by Luria and Delbruck in their study of bacterial mutation in the 1940's (Luria and Delbruck 1943). More recently it was applied to microsatellite haplotypes when dating the Cystic Fibrosis ΔF508 mutation (Bertranpetit and Calafell 1996). This method requires that a root haplotype is identified and that the mean number of mutations is calculated averaged over all loci. This proportion is then divided by the mutation rate to give the age in generations of the root haplotype, in other words the time since the founding of the lineage (Bertranpetit and Calafell 1996). This method has been used most notably in Y chromosome research to date the origin of a widely distributed central Asian lineage to a relatively recent origin (Zerjal et al. 1997).

Another method that assumes linearity of a statistic with respect to time since the founding of a lineage is Average Squared Distance (ASD) dating. The linearity of this statistic with respect to time has been demonstrated by its proponent, David Goldstein, through simulation work on microsatellite evolution (Goldstein et al. 1995). The ASD statistic is simply the squared mutational distance between a root haplotype and any another haplotype within the lineage averaged over all loci and all haplotypes. Again dividing this statistic by the mutation rate gives the age to the MRCA of the lineage in generations. This method has been used to date the origin of a lineage specific to Jewish priests (Thomas et al. 1998). The linearity of ASD with respect to time has also been

investigated for dating population splits (Goldstein et al. 1995) . In this case it was found that a related statistic performed better. This statistic, known as $\delta\mu^2$, is a version of ASD corrected for intra-population variance. In this case the $\delta\mu^2$ statistic must be divided by two times the mutation rate to give the age of the MRCA in generations (Goldstein et al. 1995). This discrepancy of a factor of two can be appreciated by considering figure 4.1 below which shows that the 'root versus haplotypes' (ASD) comparison only involves travelling one way through time whereas the population v population ($\delta\mu^2$) comparison requires that the distance to the MRCA is considered twice.

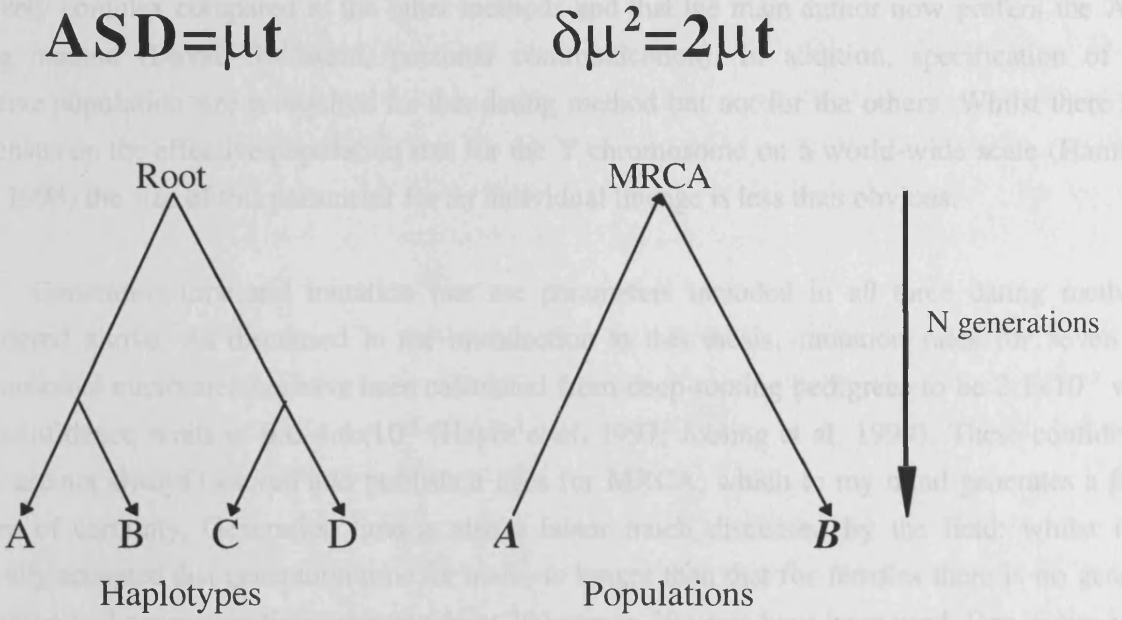


Figure 4.1: Illustration of how the comparisons 'root v haplotypes' and 'Population A v Population B' relate to time since MRCA.

Both of the above methods of dating lineages require that a root haplotype be assigned. This has generally been accomplished in one of two ways. The first method takes the most highly represented haplotype as the root, and relies on the assumption from models of population genetics that the root haplotype will be that most represented in the sample (Thomas et al. 1998). This is a feature of both single-step and infinite-sites models of mutation. The second method requires the use of diversity data from the ancestral lineage to find the most parsimonious link between the two lineages, thus identifying the root haplotype in the network of the derived lineage (Zerjal et al. 1997).

A third method for calculating lineage age comes from the modelling of how the variance at microsatellite loci increases with time and how this increase depends on various population parameters such as effective population size (Goldstein et al. 1996). This method explicitly includes equations for 95% confidence limits based on the stochasticity of the evolutionary process. This variance method for dating was used on a dataset of world-wide diversity at 5 Y-chromosomal microsatellites to show that no bottleneck relating to a selective sweep of the Y chromosome has occurred within the resolution of this analysis, thought by the authors to be of the order of 74000 years (Goldstein et al. 1996). At present there is only a single published application of this dating method to an individual lineage (Zerjal et al. 1997); this is probably due to the fact that it is relatively complex compared to the other methods and that the main author now prefers the ASD dating method (David Goldstein, personal communication). In addition, specification of the effective population size is required for this dating method but not for the others. Whilst there is a consensus on the effective population size for the Y chromosome on a world-wide scale (Hammer et al. 1998) the size of this parameter for an individual lineage is less than obvious.

Generation time and mutation rate are parameters included in all three dating methods considered above. As discussed in the introduction to this thesis, mutation rates for seven Y-chromosomal microsatellites have been calculated from deep-rooting pedigrees to be 2.1×10^{-3} with 95% confidence limits of $0.6\text{-}4.6 \times 10^{-3}$ (Heyer et al. 1997; Jobling et al. 1999). These confidence limits are not always factored into published ages for MRCA, which to my mind generates a false picture of certainty. Generation time is also a factor much discussed by the field: whilst it is generally accepted that generation time for males is longer than that for females there is no general consensus and generation times varying from 20 years to 30 years have been used. One group uses the figure of 27 years for male generation time (Cavalli-Sforza et al. 1994; Underhill et al. 1996) which comes from anthropological studies of hunter-gather societies (Weiss 1973).

Once the age of a lineage has been calculated the question arises, what is the anthropological relevance of this age? One approach that has been much debated is the correlation of lineage age with population age. This assumption has been adopted in a number of mtDNA studies (Richards et al. 1996) but has been widely attacked as failing to appreciate the differences between a population MRCA and a lineage MRCA (Barbujani 1999). The age of the MRCA of a single allele is susceptible to stochastic effects and is a poor estimator of population age. In general alleles are thought to be older than the populations in which they are found (Barbujani 1999). This debate has unfortunately eclipsed the fact that providing a temporal aspect to patterns of lineage sharing can inform in other anthropologically useful ways.

Population subdivision calculations

Apart from dating calculations, there are other numerical approaches that can be used to investigate the structure of diversity within a lineage, specifically with respect to the populations amongst which the lineage is shared (Hudson et al. 1992; Slatkin 1995; Petit et al. 1998). Most simply this can be a comparison of the relative diversities of the lineage in different populations. A number of different measures of diversity have been used to this end (Nei 1987; de Knijff et al. 1997; Perez-Lezaun et al. 1997). Nei's estimate of diversity is widely used. The measures considered here only take into account population affiliation and not geographical distances between samples. These geographically-informed statistics are considered in the next chapter.

F_{ST} is the measure of population subdivision devised by Sewall Wright in 1931 (Wright 1931; Wright 1969). It is estimated as a normalised form of allele frequency variance across subdivisions (normally subpopulations). F_{ST} values at a neutral locus result from the opposing effects of genetic drift and gene flow (Kimura 1983; Slatkin 1995). Permutation tests can be used to calculate whether any given F_{ST} value represents a significant subdivision of diversity. F_{ST} can be used to determine which level of population subdivision summarises the most variance within the entire sample (Barbujani et al. 1997). This approach was used by Lewontin with allele frequency data to show that racial groupings have no genetical basis and that the vast majority of the worlds diversity is found within rather than between populations (Lewontin 1972). This work was subsequently extended by Barbujani *et al.* (Barbujani et al. 1997).

F_{ST} does not take account of the mutational distance between different alleles, or in the case here, microsatellite haplotypes. The Analysis of Molecular Variance (AMOVA) is a hierarchical test of population subdivision that does include this information and as such is more appropriate for this type of molecular data (Michalakis and Excoffier 1996; Schneider et al. 1997). AMOVA has been used to show that even small differences between two closely related populations could be shown to be significant (Roewer et al. 1996). The ratio of the molecular variance between populations to the total variation in the data set is known as ϕST and can be considered to be an analogue of F_{ST} informed with a consideration of molecular distance between haplotypes (Schneider et al. 1997). There are a number of other population distance statistics that take account of mutational distance (Hudson et al. 1992; Slatkin 1995; Petit et al. 1998).

Y-chromosomal lineage studies

Having considered the major tools for lineage analysis I shall now consider a couple of examples of Y-chromosomal lineage studies to illustrate how these tools have been used in combination to generate meaningful anthropological conclusions.

The study by Zerjal and others (Zerjal et al. 1997) identified a novel lineage defined by a T to C base substitution. This lineage is called haplogroup 16 by the authors. The distribution of this lineage was investigated by typing over 1000 samples world-wide. It was found to be present in a subset of populations of Central Asia and North-eastern Europe. Microsatellite diversity at 10 loci was assayed to further investigate this lineage. Substantial population subdivision of this intra-lineage diversity was seen. A minimum-spanning network of this diversity was constructed which clearly showed the population subdivision. A network was also constructed for the microsatellite diversity apparent within the lineage ancestral to haplogroup 16, called haplogroup 12. A root was proposed for haplogroup 16 based on the most parsimonious linkage between these networks. This root was used in one of the dating methods, that based on the mean number of mutations within the lineage. This dating method, together with that based on microsatellite variance, showed that this lineage was relatively young in origin, being 2-4,000 years old. This study did not present confidence limits on the date for the origin of the lineage based on mutation rate uncertainty as a consequence of it being published prior to the establishment of confidence limits for this parameter. Diversity considerations were invoked to propose a central Asian origin for the mutation. This lineage was found at high frequency in the Finnish population which speaks a language belonging to the Uralic family of languages which are spoken mostly in Central Asia. Studies of Finnish mtDNA had failed to show any convincing differences between Finns and other Europeans (Sajantila et al. 1995). Consequently this study proposed the paternal co-inheritance of this lineage and the Uralic language in this population.

The study by Thomas and others (Thomas et al. 1998) sought to investigate the origins of Jewish Priests, the Cohanim. Patrilineal inheritance of this profession has occurred since the time of the Temple, 2-3,000 years ago. Six unique markers and six microsatellites were assayed to construct compound haplotypes. A common 'modal' haplotype was found shared between Ashkenazic and Sephardic Cohanim yet was at low frequency in other Jews from these communities. These two communities have been relatively isolated from one another over the past 500 years, and therefore the sharing of this common haplotype is good evidence for the common origin of the Cohanim. A sublineage within one of the groups delineated by the unique markers was defined as containing the modal haplotype and all other haplotypes within one microsatellite

mutational step of it. The age of this lineage was calculated using the ASD method of dating to be 2650 years old using a generation time of 25 years. Confidence limits around this figure were calculated on the basis of evolutionary sampling, assuming population expansion indicated by the 'star' phylogeny of this lineage, to be 2100-3250 years, which fits nicely with the timing of the Temple period in Jewish history. However, factoring in the confidence limits in the mutation rate gives the much wider range of 850-11375 years. Not addressed in this study was the circularity of defining a lineage on the basis of microsatellite diversity and then subsequently using the same data to date the lineage. It was not made clear whether there were any haplotypes which were two mutational steps away from the modal haplotype, in which case the selection of haplotypes a single mutational step away from the modal haplotype seems quite arbitrary.

In this chapter I intend to apply the above methods in two studies of individual lineages, one in Bulgarian Gypsies and the other found mainly in Iberia. The latter sudy resulted in a publication (Hurles et al. 1999) which is included in appendix C. By necessity I shall detail some anthropological background to these populations in order to illustrate what we can gain from Y-chromosomal lineage analysis. In addition I hope to extend some of the analyses to make them suitable for application to intra-allelic diversity of MSY1. Given that there are competing methods for doing similar analyses and that the field has not reached a consensus as to how such data should be analysed I hope to explore which methods offer the best opportunity to make valid anthropological inferences from the analysis of Y-chromosomal lineages.

Materials and Methods

Data

In the study of Gypsy populations: the samples were donated by Luba Kalaydjeva, mtDNA typing and analysis was performed by Francesc Calafell, Y-chromosomal microsatellites were typed by Peter de Knijff, MSY1 was typed by Mark Jobling and Y-chromosomal lineages were identified by Zoë Rosser. My contribution to this study was the analysis of the Y-chromosomal data and the comparison of the analyses from the Y-chromosomal data and the mtDNA data.

In the study of the haplogroup 22 lineage, samples were provided by: Manuel Armenteros, Eduardo Arroyo, Anne Cambon-Thomsen, Lalji Singh, Manfred Kayser, Yuri Dubrova, Santos Alonso, Carlos Polanco, John Armour, John Mitchell, Marisol Rodriguez-Calvo, Jaume Bertranpetit, Chris Tyler-Smith and Mark Jobling. Microsatellite typing was done by Reiner Veitia, Anna Pérez-Lezaun. MSY1 typing was performed by Paul Taylor and Mark Jobling. DNA purification was performed by Maria Shlumukova and myself. Typing of the SRY-2627 polymorphism was done by Fabricio Santos, Arpita Pandya and myself. I found the vast majority of chromosomes belonging to haplogroup 22. Arne Röhl provided the Median-Joining Network program 'Network 1.1'. Ian Wilson calculated the age of the lineage using a coalescent approach. I performed all other analyses on the data set.

Software

The program Microsat used to calculate R_{ST} and ASD was written by Eric Minch and obtained from the web site, http://lotka.stanford.edu

Multidimensional Scaling (MDS) was performed by Yuri Dubrova using the MDS program in the Statistica Package.

Median-Joining Networks were calculated using the program 'Network 1.1' from Arne Röhl.

The Microsoft Excel spreadsheet used to calculate dates for the lineages and to provide the matrix of pairwise differences for the drawing by hand of minimum-spanning networks was based on a spreadsheet designed to perform only the latter function, developed and kindly supplied by Fabricio Santos.

Permutation testing of pairwise F_{ST} comparisons was done using the Arlequin v1.1 software available from the web site http://anthropologie.unige.ch/arlequin

DNA purification by the silica method

The silica was prepared by the following protocol (Boom et al. 1990):

12g of silica particles (Sigma) were placed in a clean glass 100ml measuring cylinder and water added to 100ml. The cylinder was then sealed, well mixed and the solution left to settle for 24 hours.

The upper 86ml of water was then removed and the cylinder filled up to 100ml with fresh water, sealed, mixed and allowed to settle for 5 hours before the removal of the upper 88ml of water.

120µl of 10M HCl was added, the silica fully resuspended by mixing and 500µl aliquots put into 1.5ml Eppendorf tubes. These were stored away from light, covered in aluminium foil.

The buffers L2 and L6 were made by the following protocols before being stored away from light in aluminium foil:

L6

24g of Guanidinium thiocyanate (GuSCN, Sigma) was added to 20ml of 0.1M Tris-HCl, pH 7.4 and dissolved at 60°C in a waterbath

1.8ml of 0.5M of EDTA, pH 8.0 was added to this solution

0.5ml of Triton X100 (Sigma) was added and this final solution mixed by inverting

L2

24g of GuSCN was added to 20ml of 0.1M Tris-HCl, pH 7.4 and dissolved at 60°C in a waterbath

New Wash - 50% ethanol, 0.1MNaCl (Sigma),10mM Tris (pH7.5)

DNA purification

6-7 volumes of buffer L6 were added to the material from which DNA was to be purified (blood, agarose plug), and incubated at 50°C for 15 minutes in a 1.5ml Eppendorf tube.

The tubes were cooled to room temperature and 20µl of vortexed silica suspension added.

The tubes were then incubated at room temperature for at least 30 minutes on a slow vertical rotor to maintain the silica in suspension.

The tubes were spun in a microfuge at maximum speed for 2 minutes and the supernatant removed.

5 volumes of L2 buffer were then added to the silica pellet and the tube vortexed briefly to resuspend the pellet.

The pellet was again spun down and the supernatant removed; this process was repeated three times with the adding of ice cold New Wash solution instead of the L2 buffer.

After the final wash the tube was spun twice to insure the removal of all supernatant.

The silica pellet was air dried at room temperature of 37°C to remove all residual ethanol.

The DNA was eluted by adding 50-60µl of TE (pH 8.0), resuspending the pellet and incubating at 50°C for 15-20 minutes with occasional agitation.

The pellet was then spun down as before and the DNA containing supernatant pipetted off and stored at -20°C.

The elution step was repeated to obtain a lower concentration DNA solution.

Results

A comparison of different methods for representation of lineage diversity

An obvious first step to approaching lineage analysis would seem to be to test some of the competing methods available for different analyses against one another using some external criteria for deciding between them. However for some analyses it is of interest to apply multiple methods because there is no obvious criterion for deciding between the methods and thus no 'right' answer. Dating is a good example of such an analysis, with there being no available data set of intra-allelic diversity for a lineage, of independently-defined age.

By contrast we can test the performance of different methods for displaying intra-allelic diversity by their ability to resolve unique sublineages that we know exist but the evidence for which we exclude from the analysis. Although this empirical test relies on the subjective interpretation of the observer as to which method best resolves the sublineages, this is exactly the same mechanism by which inferences are made from these analyses in the first place. In other words; to what extent can we trust the patterns of clustering within these graphical displays of diversity to reveal groups of phylogenetic relevance?

To these ends three different methods of displaying intra-allelic diversity were applied to the same data set. The three methods applied were a Median-Joining network, a Minimum-Spanning network and a 3-dimensional multi-dimensional scaling (MDS) analysis. The data set consisted of seven-locus microsatellite haplotypes for 35 chromosomes belonging to a lineage characterised by the presence of the SRY-1532 derived allele and ancestral alleles at the 92R7, SRY+465, Tat and Yap markers. The lineage could be further subdivided into sublineages by three unique markers, one a base substitution found by Peter Underhill, M9, and the other two MSY1 modular structures found only within a single lineage in a world-wide sample set, symbolised by the nomenclature (...4,0,4) and (3,1,3+,4-) (see chapter 6). Each sublineage was mapped onto both networks and the MDS analysis separately and the extent to which these sublineages formed clusters compared between the analyses.

Figure 4.2 shows the results of this process on the M-J network, Figure 4.3 shows the result on the M-S network and Figure 4.4 the result on the MDS analysis.

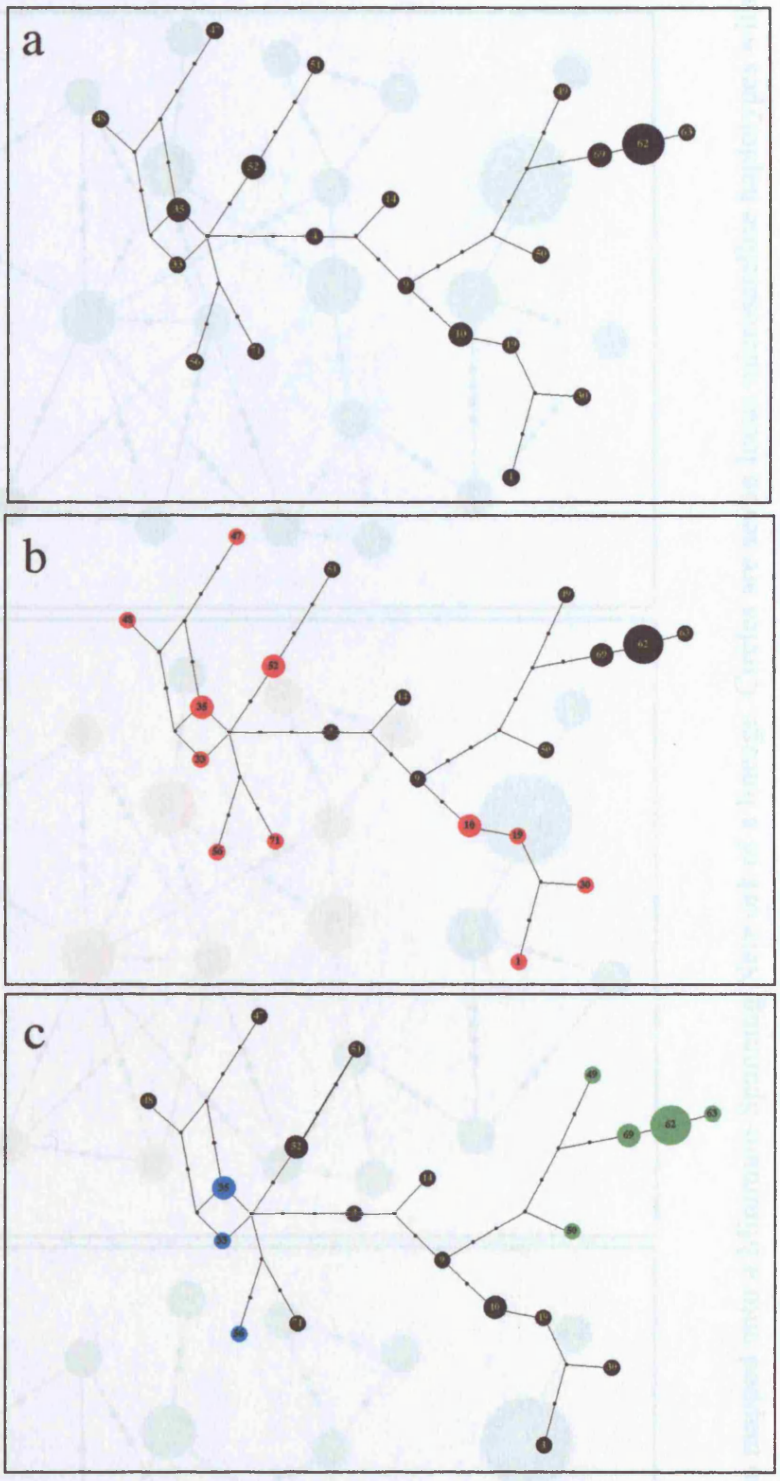


Figure 4.2: Unique sublineages mapped onto a Median-Joining Network of a lineage. Circles are seven-locus microsatellite haplotypes with area proportional to frequency Numbers are haplotype identifiers. Small circles represent unobserved intermediate haplotypes. (a) the network of the lineage, (b) the sublineage defined by the unique marker M9 mapped onto the lineage in red and (c) the sublineages defined by the two MSY1 modular structures mapped onto the lineage (...4.0,4) in green and (3,1,3+,4-) in blue

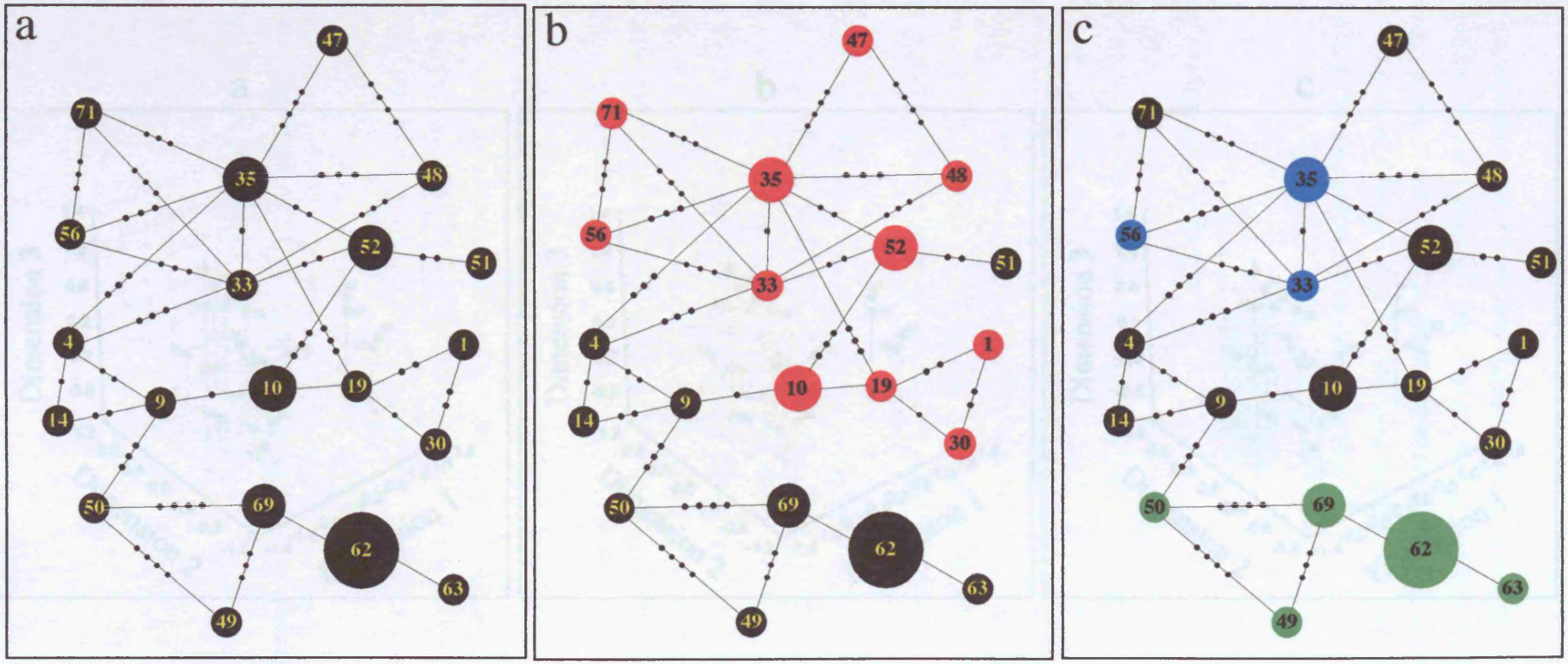


Figure 4.3: Unique sublineages mapped onto a Minimum-Spanning Network of a lineage. Circles are seven-locus microsatellite haplotypes with area proportional to frequency. Numbers are haplotype identifiers. Small circles represent unobserved intermediate haplotypes. (a) the network of the lineage, (b) the sublineage defined by the unique marker M9 mapped onto the lineage in red and (c) the sublineages defined by the two MSY1 modular structures mapped onto the lineage (...4.0,4) in green and (3,1,3+,4-) in blue.

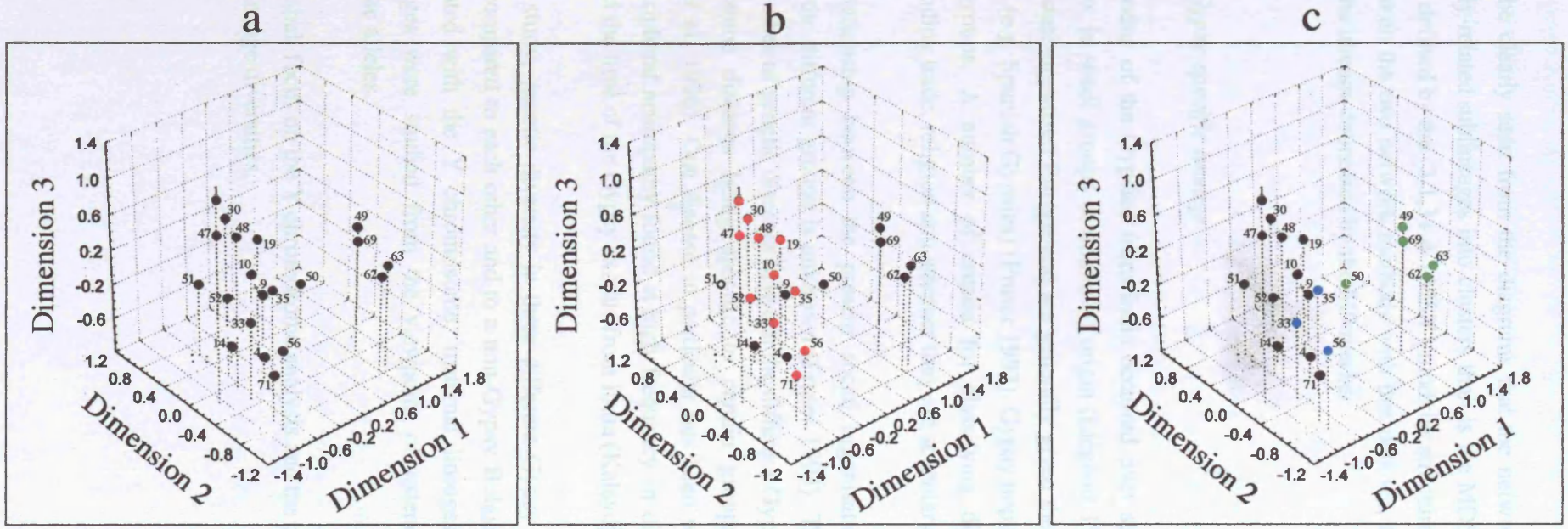


Figure 4.4: Unique sublineages mapped onto a Multidimensional Scaling representation of intra-lineage diversity. Circles are seven-locus microsatellite haplotypes. Numbers are haplotype identifiers. (a) the entire lineage, (b) the sublineage defined by the unique marker M9 mapped onto the lineage in red and (c) the sublineages defined by the two MSY1 modular structures mapped onto the lineage (...4.0,4) ingreen and (3,1,3+,4-) in blue.

It can be clearly seen from the diagrams that the networks are better at resolving these phylogenetically-related sublineages into clusters than is the MDS analysis, especially with regard to the lineages defined by the (3,1,3+,4-) MSY1 modular structure and the SNP M9. There is little to choose between the two network methods with this data set, though perhaps the M-S network better defines the lineage delineated by the M9 marker.

Analysis of a Gypsy-specific lineage

The exodus of the Gypsies from India occurred over several centuries, starting roughly 1000 years ago, in small groups of unknown origin (Liegeois 1994). The Gypsies are presently spread across state-organised Europe and are generally given the descriptor of the state in which they are found (e.g. Spanish Gypsies) (Fraser 1993). Gypsy populations tend to be organised into endogamous groups. A number of criteria for classifying different gypsy tribes have been proposed, including trade, religion and whether they are sedentary or Nomadic (Liegeois 1994).

The relationship between the present social organisation of Gypsies and the genetic relatedness of the different groups is unknown (Fraser 1993). The epidemiological pattern of the sharing of a number of genetic diseases between the different Gypsy groups is inconclusive in this regard, with some diseases being specific to certain groups and others shared by many (Kalaydjieva et al. 1996). One disease in particular has been well studied. A mutation causing HMSNL, a peripheral neuropathy found at high frequency in different Gypsy groups, has been dated to around the time of the Gypsy exodus from India (Kalaydjieva et al. 1996).

In this study, genetic diversity in three different Gypsy populations from Bulgaria were analysed and compared to each other and to a non-Gypsy Bulgarian population. Paternal lineages were investigated with the Y chromosome, maternal lineages with mtDNA and biparentally inherited lineages were studied from the viewpoint of extended microsatellite haplotypes of HMSNL disease alleles.

Here I shall focus on the Y chromosome analysis and the comparisons between the paternal and maternal lineage diversities.

The Y-chromosomal markers typed on the four samples were eight biallelic markers, seven Y-specific microsatellites and the minisatellite MSY1. The typing information on the three Gypsy groups is summarised in table 4.1.

Forty-three out of the 48 Gypsy Y chromosomes belong to a single haplogroup, haplogroup 2. Forty of these 43 haplogroup 2 chromosomes also have the same MSY1 modular structure (3,1,3,4). Haplotypes could be constructed for 43 chromosomes consisting of both minisatellite and microsatellite loci, where each block of a single repeat type at MSY1 is considered to be analogous to a microsatellite. A Minimum-Spanning network was constructed, comprising the 35 compound haplotypes belonging to haplogroup 2 and having the (3,1,3,4) MSY1 modular structure, and is shown in figure 4.5. The group affiliations of the different haplotypes are mapped onto the network. Thirty of these 35 chromosomes cluster into a tight grouping which would seem to define a sublineage, named '2.1'. Two other potential sublineages are also defined. Only a single Y chromosome from the 18 non-Gypsy Bulgarian sample also belongs to this major sublineage. Five more Gypsy chromosomes could also be assigned to this 2.1 sublineage on the basis of their MSY1 codes, despite not having microsatellite data, taking the total to 35 Gypsy Y chromosomes belonging to this lineage. A survey of 147 MSY1 codes having the modular structure (3,1,3,4) in Y chromosomes belonging to haplogroup 2 failed to reveal a single code fewer than eight mutational differences from any of these 35 chromosomes. Therefore this cluster was considered to be a Gypsy-specific lineage.

The haplogroup distributions of the 3 Gypsy groups amalgamated into a single sample and the Bulgarian non-Gypsies show a pairwise F_{ST} value significant at the 5% level. Removal of the 2.1 lineage from the Gypsy sample gives a pairwise F_{ST} value between the amalgamated Gypsy sample and the Bulgarian sample that is no longer significant. This might indicate a low level of admixture between the Gypsies and the surrounding population. It is interesting that this level of probable admixture is considerably lower than for maternal lineages based on the mtDNA data (Francesc Calafell, personal communication).

There is a striking degree of code sharing within Gypsy groups; Nei's unbiased estimator of diversity gives values of 0.711, 0.867 and 0.814 in the Monteni, Kalderas and Lom respectively. Previous work on MSY1 code diversity in known isolates such as the Basques, Finns and Cook Islanders have revealed code diversities for these populations in excess of 0.96. This indicates a probable small effective population size for these groups.

Name	Group	hg	ht	19	389I	389II	390	391	392	393	MSY1	MSY1 code
1	W	2	3	15	16	14	22	10	11	12	2.1a	
2	W	2	12	15	16	14	23	10	11	12	2.1a	
3	W	2	3	15	16	14	22	10	11	12	2.1a	
4	W	2	3	15	16	14	22	10	11	12	2.1a	
8	W	2	3	15	16	14	22	10	11	12	2.1a	
28	W	2	3	15	16	14	22	10	11	12	2.1a	
29	W	2	3	15	16	14	22	10	11	12	2.1a	
35	W	2	-	nd	nd	nd	nd	nd	nd	nd	2.1a	
48	W	2	3	15	16	14	22	10	11	12	2.1a	
49	W	2	3	15	16	14	22	10	11	12	2.1a	
52	W	2	21	15	16	14	21	10	11	12	2.1a	
5	W	2	3	15	16	14	22	10	11	12	2.1b	
6	W	2	3	15	16	14	22	10	11	12	2.1c	
7	W	2	3	15	16	14	22	10	11	12	2.1d	
11	L	2	13	14	16	14	22	10	11	12	2.1e	
12	L	2	13	14	16	14	22	10	11	12	2.1e	
13	L	2	3	15	16	14	22	10	11	12	2.1e	
14	L	2	13	14	16	14	22	10	11	12	2.1e	
17	L	2	3	15	16	14	22	10	11	12	2.1e	
26	L	2	-	nd	nd	nd	nd	nd	nd	nd	2.1e	
31	K	2	3	15	16	14	22	10	11	12	2.1e	
32	K	2	3	15	16	14	22	10	11	12	2.1e	
33	K	2	3	15	16	14	22	10	11	12	2.1e	
41	K	2	12	15	16	14	23	10	11	12	2.1e	
47	K	2	3	15	16	14	22	10	11	12	2.1e	
16	L	2	13	14	16	14	22	10	11	12	2.1f	
19	L	2	13	14	16	14	22	10	11	12	2.1f	
20	L	2	13	14	16	14	22	10	11	12	2.1f	
22	L	2	-	nd	nd	nd	nd	nd	nd	nd	2.1f	
23	L	2	13	14	16	14	22	10	11	12	2.1f	
24	L	2	-	nd	nd	nd	nd	nd	nd	nd	2.1f	
43	L	2	-	nd	nd	nd	nd	nd	nd	nd	2.1f	
36	K	2	3	15	16	14	22	10	11	12	2.1g	
51	W	2	3	15	16	14	22	10	11	12	2.1g	
45	K	2	3	15	16	14	22	10	11	12	2.1h	
9	W	2	7	14	16	12	22	10	11	13	2.2a	
15	L	2	14	14	16	13	22	10	11	13	2.2b	
10	L	2	9	17	17	13	24	10	11	13	2.3a	
27	L	2	16	16	18	13	24	11	11	13	2.3b	
37	K	2	9	17	17	13	24	10	11	13	2.3c	
18	L	2	15	14	17	14	23	10	11	12	2.4a	
21	L	2	15	14	17	14	23	10	11	12	2.4b	
46	K	2	11	16	19	13	22	10	11	12	2.5	
30	L	1	17	15	16	13	24	11	13	13	1.1	
38	K	1	18	14	16	13	25	10	13	13	1.2	
42	W	21	2	13	17	13	24	10	11	13	21.1	
44	K	21	2	13	17	13	24	10	11	13	21.2	
50	W	3	20	17	16	13	24	10	11	13	3	

Table 4.1 Haplogroup (hg), microsatellite (ht) and MSY1 data on all Gypsy samples. Microsatellite alleles described by numbers of repeat units. MSY1 is coded as follows: ○ : type 1 repeat; ● : type 3 repeat; ◌ : type 4 repeat; ⊙ : null repeat. nd: not done

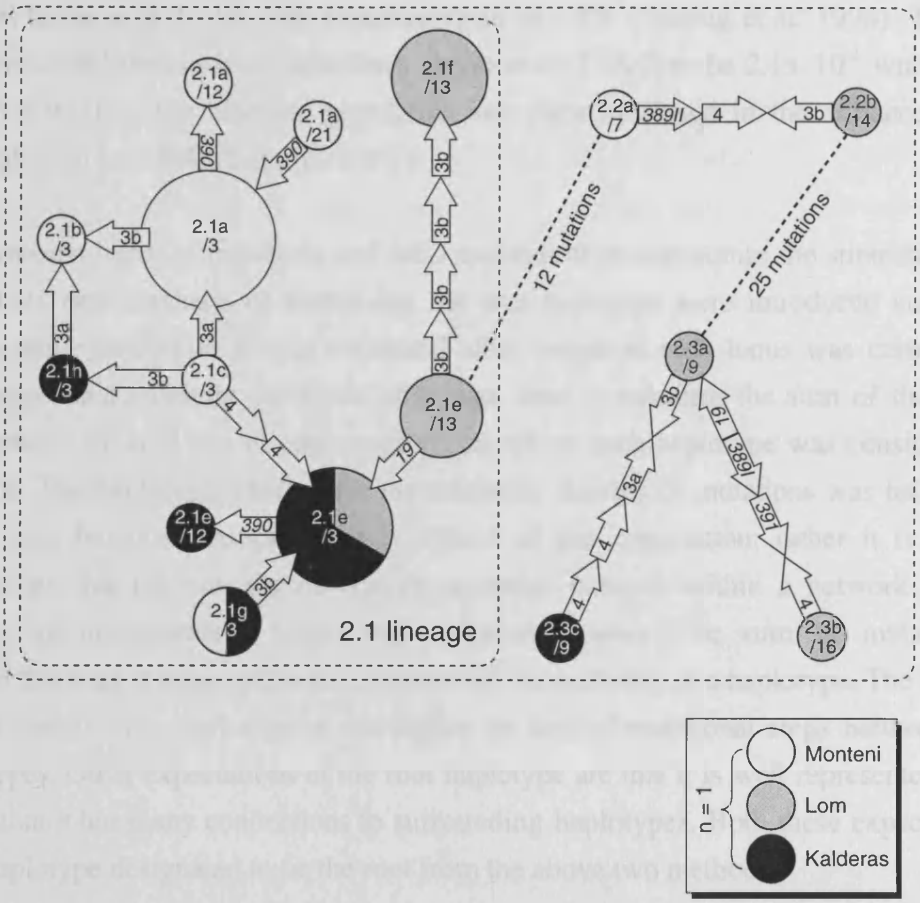


Figure 4.5: Minimum-Spanning Network of the 35 Gypsy Y chromosomes belonging to Haplogroup 2 having both microsatellite and MSY1 data. Circle area is proportional to frequency. Each arrow represents a single mutational step with the arrow head indicating the larger allele. The number within the arrow indicates the locus at which the mutation step occurs. MSY1 blocks are numbered from the 5′ end 3a,1,3b,4. The number within the circles indicates the MSY1 codes and microsatellite haplotypes described in Table 4.1.

An Excel worksheet was designed that was capable of calculating the age of the Gypsy lineage from microsatellite data by three different methods. This Excel table is shown in appendix B. The three methods were that of the mean number of mutations (Bertranpetit and Calafell 1996), the average variance (Goldstein et al. 1996) and the ASD method (Thomas et al. 1998). The latter method required the external calculation of the ASD value which was then fed into the worksheet while the other dating methods were performed automatically. A generation time of 20 years was used. This is shorter than is normally used for male generation time but is supported on the basis of ethnological studies in the Gypsies (Luba Kalaydjeva, personal communication). The mutation rate

of MSY1 was taken to be 2-11% with a median value of 6.5% (Jobling et al. 1998). The mutation rate of the microsatellite loci was taken from Heyer *et al.* (1997) to be 2.1×10^{-3} with confidence limits of $0.6\text{-}4.9 \times 10^{-3}$. The effective population size parameter used in the variance method of dating was taken to be 4,900 (Hammer 1995).

The mean number of mutations and ASD methods of dating require the stipulation of a root haplotype. Two new methods of identifying the root haplotype were introduced and they both identified the same haplotype. Firstly the modal allele length at each locus was combined into a single haplotype, and secondly the Excel table was used to calculate the sum of the number of mutations between the root and all other haplotypes, where each haplotype was considered in turn to be the root. The haplotype which gave the minimum number of mutations was taken to be the most likely root. No mathematical proof is offered of this expectation; rather it is an intuitive conclusion given that the root should occupy a central position within a network of diversity generated by an unconstrained single step mutation process. The sum of mutational steps considered in this way is expected to be a measure of the centrality of a haplotype. The further from the centre of a network a haplotype is, the higher the sum of mutational steps between it and all other haplotypes. Other expectations of the root haplotype are that it is well represented within the sample, and that it has many connections to surrounding haplotypes. Both these expectations were true of the haplotype designated to be the root from the above two methods.

In addition the dating methods were modified to allow dating from the code diversity of MSY1, by considering blocks of same-sequence repeats to be analogous to independent microsatellite loci. Initially all four blocks of repeat types at MSY1 were considered to be equally mutable, as are the microsatellite loci that comprise the seven-locus haplotype. This gave very young ages for the lineage (data not shown). It is obvious from inspecting the MSY1 codes of this and other lineages that the larger blocks of repeat types are more variable than are the smaller ones. In the absence of any direct information on mutation rates of different repeat types the simplest correction for this observation is to give each block of repeats a mutation rate proportional to the number of repeats in that block. These mutation rates were normalised such that the sum of the mutation rates of the four blocks equals the known mutation rate of the whole minisatellite. In effect this considers that each repeat unit is equally likely to mutate.

Table 4.2 shows that this correction gives ages for the '2.1' lineage much greater than those from microsatellite data. In addition there is poor agreement between the ages produced from the three methods. This is unusual given that, on other occasions that these different methods have been compared, they have given similar ages (see haplogroup 22, below). One possible reason for

this discrepancy was proposed after closer inspection of the network. All three methods assume that the loci involved obey the single step mutation model. The network shows a number of links in the network of multiple mutational steps, and all of these involve MSY1 alone. This non-observance of intermediate haplotypes may result from saltatory mutation or from genetic drift. Drift is likely to be high in these small endogamous groups. However the network provides additional evidence that the lack of intermediate haplotypes is due to saltatory mutation in that both such examples of multi-step links involve a single block of repeats either expanding or contracting. If this non-observance of intermediate haplotypes was due to drift then we might expect that many of these multi-step links would involve mutations within multiple blocks. Saltatory mutations might explain the divergent age estimates both between methods and compared to microsatellite-based ages.

A crude way to correct for saltatory mutation might be to consider those mutations of multiple repeats in the same block to be single mutation events, equal in rate to normal single step mutations. The ages for lineage 2.1 calculated making the above assumptions and using the three different methods are given in table 4.2. It can be seen that there is now much better agreement both between methods and between types of loci.

	microsatellites		**MSY1 single-step**		**MSY1 saltatory**	
Method	**age**	**95% CI**	**age**	**range**	**age**	**range**
(i) Mean mutations from root	454	194-1587	652	385-2118	335	198-1088
(ii) Accumulated variance	392	150-1874	1254	663-5054	239	169-1179
(iii) ASD	457	196-1600	1973	1166-6412	480	284-1560

Table 4.2: Dates for the origin of the 2.1 lineage.
Ages, and 95% confidence intervals (CI) in years are given for microsatellites, and ages and expected ranges for MSY1 under a single-step assumption, and also under the assumption that the four-repeat increase between sub-lineages 2.1e and 2.1f, and the two-repeat increase between 2.1c and 2.1e, are single mutational events ('saltatory').

The ages for the 2.1 lineage given in table 4.2 agree on an age of roughly 400 years for this lineage with upper confidence limits ranging up to 1900 years. Given that this lineage is shared between the three Gypsy groups it is likely that the origin of this lineage predates the fission of these three groups, therefore it is of interest to see that the age of this lineage agrees well with the dates of the Gypsy exodus from India given earlier.

It is obvious from the network that the distribution of diversity within the 2.1 lineage within the three Gypsy groups is significantly non-random. Despite their common origin there is a marked divergence between the different groups. The $\delta\mu^2$ population distance value can be used to generate an absolute date for the split between these populations of chromosomes or it can be used to arrive at an age relative to the age of the lineage as a whole, by comparing to twice the ASD calculated for the dating of the lineage. The dates of population fission estimated as a percentage of the lineage age are: Monteni-Lom - 88%, Kalderas-Lom - 40% and Monteni-Kalderas - 33%. The recent ages and the small sample size mean that these estimates will have been affected to a great deal by the stochasticity of the evolutionary process. However it seems likely from these estimates that these closely-related and geographically closely-positioned Gypsy groups have existed independently for centuries.

Analysis of an Iberian-specific lineage

The Basque population of Northern Spain speak a language not known to be related to any other language, extant or ancient (Collins 1986). All other Iberian populations, including Catalans, speak a language belonging to the Indo-European family of languages. The Basques exhibit unusual frequencies of many autosomal and mtDNA alleles compared to the other Iberian populations (Calafell and Bertranpetit 1994; Côrte-Real et al. 1996; Comas et al. 1998). A survey of barriers to gene flow in Europe using autosomal loci clearly distinguishes the Basques from surrounding populations (Barbujani and Sokal 1990). Consequently it is thought that the Basques may represent a Mesolithic, isolated, population, distinct from its neighbours.

A previously described T/C base substitution 2627bp upstream of the SRY gene was known to be polymorphic in French and in Iberian-derived populations of South America (Bianchi et al. 1997; Veitia et al. 1997). A collaborative study was initiated whose purpose was to survey the world-wide distribution of the Y-chromosomal lineage defined by the probable derived state of this polymorphism, haplogroup 22. The lineage was found at relatively high frequencies in a sample of

Basques and at low frequencies in some populations of Western Europe. Subsequently a large number of samples were obtained from Iberia to fine-map the distribution of this lineage within this region. Some of these genomic DNA samples were obtained from blood by the silica method of DNA purification. The results of the typing of 1191 chromosomes are given in Table 4.3. The world-wide distribution of haplogroup 22 is shown in figure 4.6. It can be seen that, considering only populations of sample size greater than 10, the lineage is most prevalent in the Basques (11%) and the Catalans (22%).

Sequencing of the chimp and gorilla homologues failed to allow unambiguous assignation of ancestral state; however, a number of indirect lines of evidence point strongly to the rare, C, allele which defines haplogroup 22, being the derived state (Hurles et al. 1999).

In order to investigate the nature of this lineage-sharing between two populations speaking radically different languages, seven-locus microsatellite haplotypes and MSY1 codes were generated for all available chromosomes belonging to haplogroup 22. In addition a number of chromosomes belonging to the presumed ancestral lineage, haplogroup 1, were typed for the same microsatellite markers. Tables 4.4 and 4.5 summarise these data.

Two different scenarios for the sharing of this lineage between Basques and Catalans can be envisaged; the first requires that gene flow has occurred over the linguistic barrier that currently exists between the two, whilst the second does not. Firstly haplogroup 22 may have had a recent Iberian origin and thus the rare non-Iberian examples indicate a low rate of Iberian emigration. Alternatively lineage sharing is explained by this lineage having an ancient origin outside of Iberia, and subsequent immigration into Iberia, occurring at a time when the two populations were not linguistically differentiated. Drift has subsequently raised the frequency of this lineage in both populations. The expectations of the two explanations for the apportionment of the intra-lineage diversity between different populations are radically different.

Median-Joining networks were constructed for haplogroup 22 microsatellite data, MSY1 codes belonging to the modular structure (1,3,4), and the haplogroup 1 microsatellite data. These networks are shown in figure 4.7.

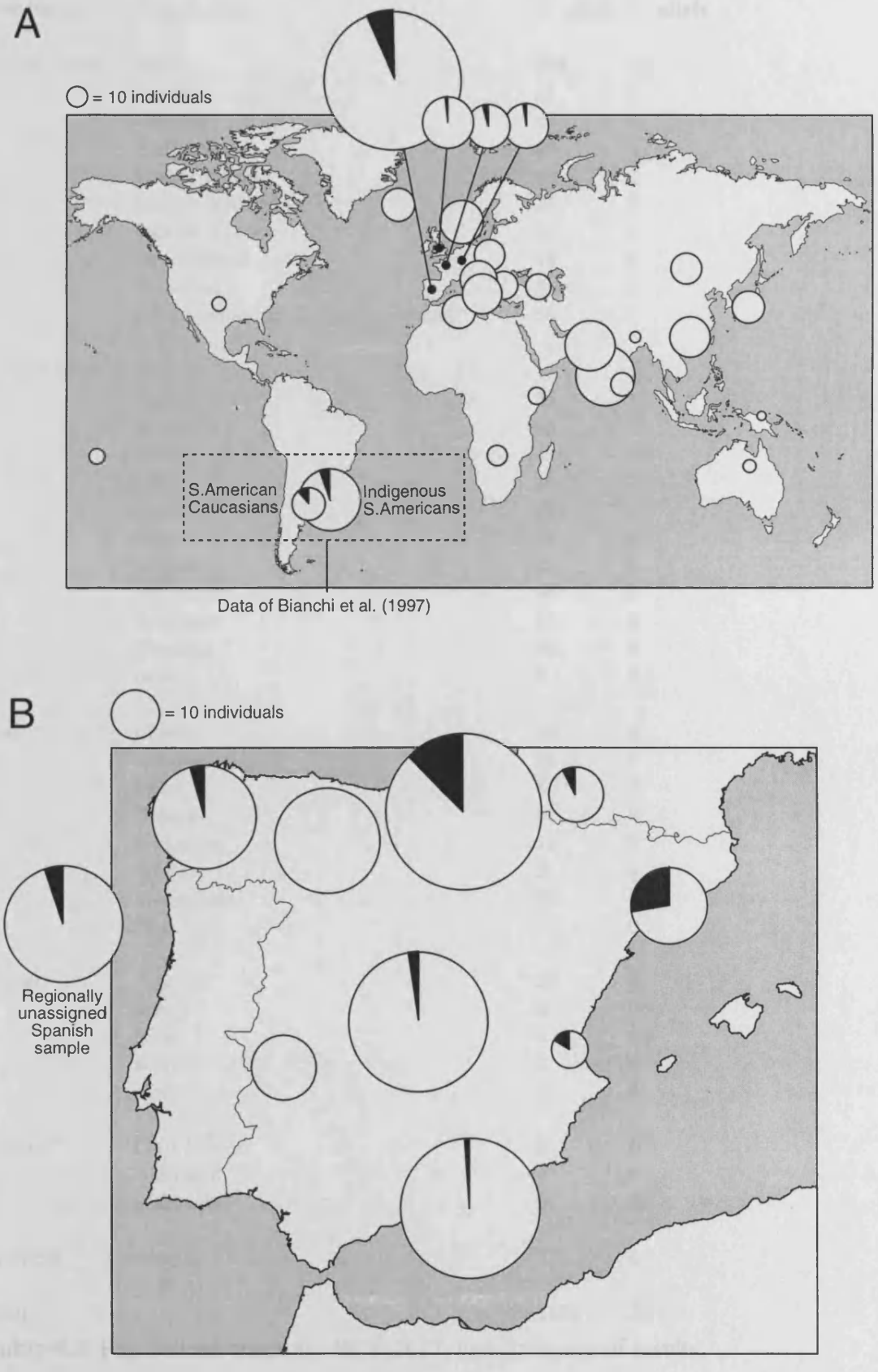


Figure 4.6: Maps showing the distribution of haplogroup 22: A - World-wide and B - in Iberia Circle area is proportional to sample size. The black segment represents haplogroup 22.

Continent	Population	C allele	T allele
Europe: Iberia	Basque	104	13
	Catalan	25	7
	Galician	46	2
	Andalucian	83	1
	Madrid (urban)	60	2
	Castilla la Mancha	23	0
	Castilla y León	47	0
	Extremaduran	18	0
	Valencian	5	1
	other Spanish (not Basque or Catalan)	58	3
Europe: other	Béarnais	13	1
	French	33	1
	British	63	1
	German	48	1
	Italian	39	0
	Greek	19	0
	Hungarian	34	0
	Icelandic	27	0
	Norwegian	46	0
	Belarusian	23	0
	Caucasus	16	0
	other	8	0
Asia	Chinese	40	0
	Japanese	25	0
	Indian	86	0
	Gujarati	61	0
	Sri Lankan	13	0
	Tibetan	3	0
	Mongolian	23	0
	other	5	0
Africa	Algerian	27	0
	San	6	0
	Biaka	4	0
	Kenyan	7	0
	other	3	0
Oceania	Cook Islander	6	0
	Australian	4	0
	Melanesian	2	0
Americas	various	5	0
Total		1158	33

Table 4.3. Populations tested for SRY-2627, and summary of results.

Male	Origin	*DYS19*[a]	*DYS389I*[a]	*DYS389II*[a]	*DYS390*[a]	*DYS391*[a]	*DYS392*[a]	*DYS393*[a]	µsat ht[b]	MSY1 code[c]	MSY1 ht
m337	Basque	14	10	27	24	10	13	13	**14**	$(1)_{16}(3)_{40}(4)_{18}$[g]	**10**
4301	Basque	14	10	27	24	10	13	13	**14**	$(1)_{16}(3)_{38}(4)_{19}$	**7**
46205	Basque	14	10	27	24	10	13	13	**14**	$(1)_{17}(3)_{38}(4)_{14}$	**18**
35v	Basque	14	10	27	24	10	13	13	**14**	$(1)_{16}(3)_{39}(4)_{19}$	**9**
67c	Catalan	14	10	27	24	10	13	13	**14**	$(1)_{15}(3)_{38}(4)_2(3)_1(4)_{15}$	**22**
m354	Basque	14	11	28	24	11	13	13	**32**	$(1)_{16}(3)_{38}(4)_{19}$[g]	**7**
m362	Basque	14	11	28	24	11	13	13	**32**	$(1)_{16}(3)_{38}(4)_{19}$[g]	**7**
m363	Basque	14	11	28	24	11	13	13	**32**	$(1)_{17}(3)_{37}(4)_{19}$[g]	**17**
ma19	Madrid	14	11	28	24	11	11	13	**30**	$(1)_{18}(3)_{39}(4)_{19}$	**20**
m62	British	14	10	28	24	11	13	13	**33**	$(1)_{15}(3)_{38}(4)_{20}$	**3**
m147[e]	French	14	10	29	24	10	13	13	**34**	$(1)_{16}(3)_{40}(4)_{19}$	**11**
m348	Basque	14	10	27	24	11	13	13	**21**	$(1)_{16}(3)_{41}(4)_{17}$[g]	**14**
m95	French	14	10	27	24	11	13	13	**21**	$(1)_{16}(3)_{39}(4)_{18}$[g]	**8**
ga29	Galician	14	10	27	24	11	13	13	**21**	$(1)_{14}(3)_{38}(4)_{18}$	**1**
7c	Catalan	14	10	27	24	11	13	13	**21**	$(1)_{16}(3)_{38}(4)_{18}$	**6**
m341	Basque	14	10	27	24	12	13	13	**35**	$(1)_{14}(3)_{39}(4)_{19}$[g]	**2**
CEPH201[f]	French	14	10	27	24	12	13	13	**35**	$(1)_{16}(3)_{38}(4)_{19}$	**7**
6201	Béarnais	14	10	26	24	10	13	13	**17**	$(1)_{16}(3)_{38}(4)_{17}$	**5**
8v	Basque	14	11	30	24	10	13	13	**36**	$(1)_{15}(3)_{39}(4)_{18}$	**4**
32v	Basque	14	10	26	24	11	13	13	**15**	$(1)_{16}(3)_{41}(4)_{16}$	**13**
98v	Basque	14	11	28	24	10	13	13	**37**	$(1)_{16}(3)_{38}(4)_{19}$	**7**
101v	Basque	14	11	28	24	11	11	13	**30**	$(1)_{16}(3)_{38}(4)_{19}$	**7**
41c	Catalan	14	10	27	23	10	13	13	**38**	$(1)_{15}(3)_{37}(4)_1(3)_2(4)_{16}$	**21**
45c	Catalan	14	10	27	24	10	11	13	**39**	$(1)_{16}(3)_{41}(4)_{15}$	**12**
56c	Catalan	14	10	27	24	11	11	13	**25**	$(1)_{16}(3)_{38}(4)_{18}$	**6**
81c	Catalan	15	10	28	24	11	13	13	**40**	$(1)_{16}(3)_{38}(4)_{18}$	**6**
85c	Catalan	14	10	27	24	10	11	14	**41**	$(1)_{16}(3)_{38}(4)_{17}$	**5**
ma20	Madrid	14	10	27	24	11	13	14	**42**	$(1)_{16}(3)_{44}(4)_{14}$	**15**
sp21	Spanish	15	10	26	24	11	13	13	**43**	$(1)_{16}(3)_{41}(4)_{16}$	**13**
ga22	Galician	14	9	26	24	11	13	13	**23**	$(1)_{18}(3)_{38}(4)_{17}$	**19**
sp77	Spanish	14	9	26	24	11	13	13	**23**	$(1)_{16}(3)_2(1)_1(3)_{40}(4)_1(3)_2(4)_{15}$	**23**
sp79	Spanish	15	10	27	24	10	13	13	**29**	$(1)_{17}(3)_{37}(4)_{19}$	**17**
sp123	Valencian	14	11	28	23	11	13	13	**44**	$(1)_{16}(3)_{39}(4)_{18}$	**8**
CEPH3501[f]	French	14	10	28	23	11	13	13	**45**	$(1)_{16}(3)_{38}(4)_{19}$	**7**
ge3202	German	14	10	29	24	11	13	13	**46**	$(1)_{16}(3)_{44}(4)_{16}$	**16**
alm1	Andalucian	14	11	28	24	11	12	13	**47**	n.d.[h]	-

Table 4.4. Microsatellite haplotypes and MSY1 codes of haplogroup 22 chromosomes.
[a] Alleles described by numbers of repeat units, as defined by Kayser *et al.* (1997). [b] 'µsat ht' - microsatellite haplotype. [c] e.g. '$(1)_{16}(3)_{40}(4)_{18}$' indicates 5′ - (16 type 1 repeats, 40 type 3 repeats, 18 type 4 repeats) - 3′ (Jobling et al. 1998). [d] 'MSY1 ht' - MSY1 haplotype; [e] Identified in a previous study (Veitia et al. 1997). [f] Identified in a previous study (Bianchi et al. 1997). [g] Codes determined by P.G. Taylor (Jobling et al. 1998). [h] 'n.d.': not done.

Male	Origin	*DYS19*[a]	*DYS389I*[a]	*DYS389II*[a]	*DYS390*[a]	*DYS391*[a]	*DYS392*[a]	*DYS393*[a]	µsat ht[b]
m419	Indian	14	11	27	23	10	10	14	**1**
m432	Indian	14	11	27	23	10	10	14	**1**
m456	Indian	15	11	27	23	10	10	14	**2**
m464	Indian	13	10	27	23	10	10	14	**3**
m467	Indian	14	9	27	23	11	10	14	**4**
m503	Indian	13	10	29	23	10	13	13	**5**
m507	Indian	13	11	27	22	10	16	13	**6**
m620	Indian	15	10	27	23	10	10	14	**7**
m245	Mongolian	14	10	29	23	11	16	13	**8**
m283	Mongolian	16	11	27	23	10	14	13	**9**
m418	Indian	13	10	27	22	10	16	13	**10**
m434	Indian	14	11	27	23	10	10	14	**1**
m437	Indian	15	10	28	23	10	13	13	**11**
m469	Indian	14	11	27	23	10	10	14	**1**
m470	Indian	13	10	27	23	10	10	14	**3**
2001	Basque	14	9	27	23	11	13	14	**12**
2401	Basque	14	11	26	24	10	13	13	**13**
3101	Basque	14	10	27	24	10	13	13	**14**
3301	Basque	14	10	26	24	11	13	13	**15**
5501	Basque	14	11	26	24	11	13	13	**16**
6103	Basque	14	10	26	24	11	13	13	**15**
2801	Béarnais	14	11	26	24	11	13	13	**16**
3001	Béarnais	14	10	26	24	10	13	13	**17**
3701	Béarnais	14	10	26	24	11	13	14	**18**
3901	Béarnais	14	10	26	24	10	13	13	**17**
4604	Béarnais	14	10	26	24	11	13	13	**15**
4901	Béarnais	14	11	26	24	11	13	14	**19**
m256	Irish	14	10	27	25	11	13	13	**20**
m285	Norwegian	14	10	27	24	10	13	13	**14**
m288	Norwegian	14	10	27	24	11	13	13	**21**
m293	Italian	14	10	27	24	11	13	13	**21**
m366	Italian	14	10	27	23	11	13	13	**22**
7v	Basque	14	9	27	24	11	13	13	**23**
16v	Basque	14	9	27	24	11	13	13	**23**
21v	Basque	13	10	27	24	10	13	13	**24**
23v	Basque	14	10	27	24	11	11	13	**25**
24v	Basque	14	10	27	24	11	13	13	**21**
25v	Basque	14	11	27	23	11	13	14	**26**
30v	Basque	14	10	27	25	10	13	12	**27**
41v	Basque	14	11	28	24	10	13	13	**28**
44v	Basque	14	10	27	24	10	13	13	**14**
8c	Catalan	14	10	27	24	10	13	13	**14**
15c	Catalan	14	10	27	24	11	13	13	**21**
17c	Catalan	14	10	27	24	11	13	13	**21**
20c	Catalan	14	10	27	24	11	11	13	**25**
26c	Catalan	14	10	27	24	11	11	13	**25**
30c	Catalan	14	10	27	24	11	13	13	**21**
32c	Catalan	15	10	27	24	10	13	13	**29**
36c	Catalan	14	11	27	24	11	11	13	**30**
43c	Catalan	14	10	27	23	11	14	14	**31**

Table 4.5. Microsatellite haplotypes of haplogroup 1 chromosomes.

[a] Alleles described by numbers of repeat units, as defined by Kayser et al. (1997). [b] 'µsat h microsatellite haplotype.

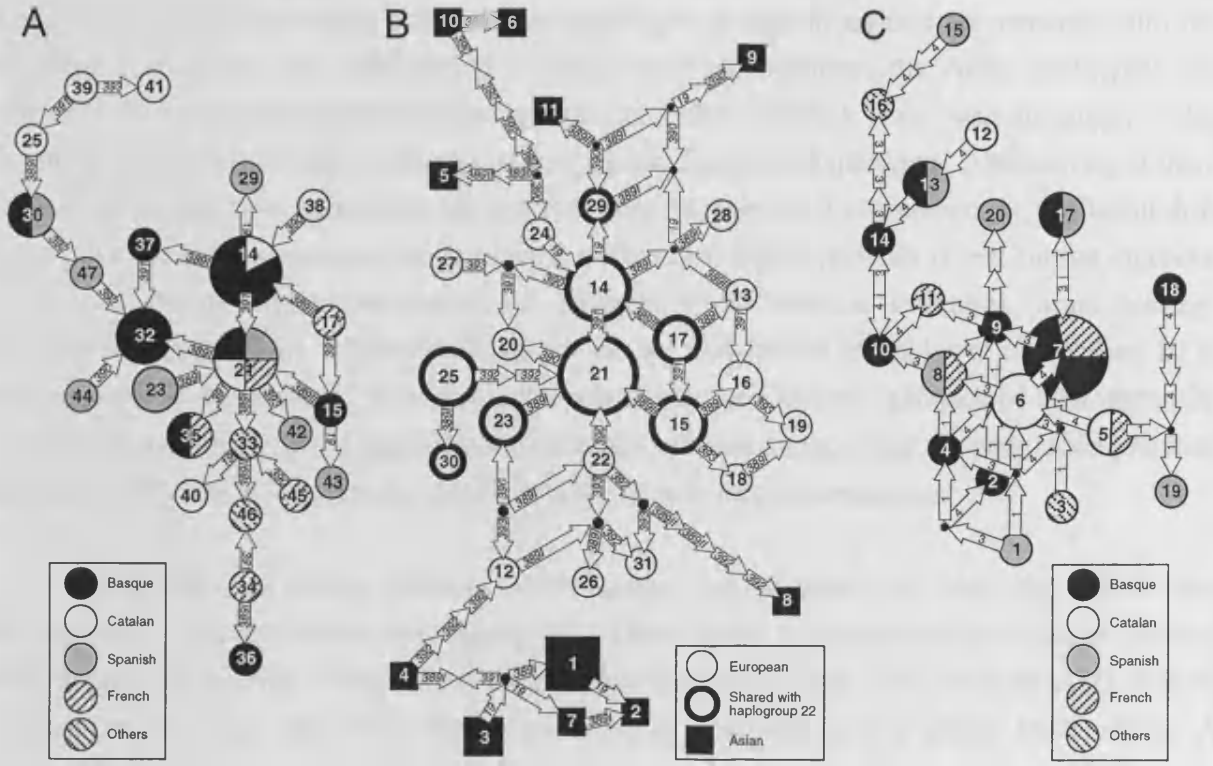


Figure 4.7: Median-Joining Networks of multiallelic diversity within Haplogroups 22 and 1. A - microsatellite diversity within Haplogroup 22, B - microsatellite diversity within Haplogroup 1, C - MSY1 diversity within haplogroup 22. A microsatellite haplotype or MSY1 code is represented by a circle with area proportional to frequency. The number within the circle refers to the haplotype numbers given in Tables 4.4 and 4.5. Arrows indicate single mutational steps, with the head pointing to the larger allele. The mutating locus is indicated on the arrow. A small, filled, circle represents an unobserved intemediate haplotype. Thickened edges to circles (B only) represent haplotypes shared between chromosomes from haplogroups 1 and 22.

It can be clearly seen that the haplogroup 22 microsatellite network is more compact and therefore likely to be younger, than the haplogroup 1 network. In addition many haplotypes are shared between the two lineages indicating insufficient time for substantial divergence of microsatellite haplotypes to occur by drift.

Haplogroup 1 is thought to have an Asian origin (Karafet et al. 1999) and in this haplogroup we find that the diversity of haplogroup 1 in Asians as measured by ASD is much greater than that of Europeans (1.762 vs 0.359). This greater diversity is apparent in the network, where the Asian examples are spread around the periphery (Figure 4.7B). Ideally, assuming an

Asian origin, we would expect to find Asian haplotypes spread throughout the network with the European haplotypes only representing a subset of them; however, the Asian haplotypes are outnumbered by the European haplotypes and are not widely sampled. If we were to sample wider we might expect to find Asian examples of the typically European haplotypes. Alternatively if there has been little gene flow between Asian and European haplogroup 1 chromosomes, sufficient drift might have occurred to generate the haplotype differences; this hypothesis is lent further credence by the two groups having characteristically different MSY1 modular structures (Mark Jobling, personal communication). In contrast we do not see the non-Iberian examples of haplogroup 22 at the periphery of the network. Comparing the non-Iberian and Iberian examples of haplogroup 22 we find that though there is a significance difference between them, using the population-pairwise comparison R_{ST}, there is no greater diversity in the non-Iberian chromosomes.

Three different dating methods were applied independently to both the MSY1 and microsatellite diversity within haplogroup 22. These were the proportion-of-mutants method (Bertranpetit and Calafell 1996), the variance method (Goldstein et al. 1996) and the ASD method (Thomas et al. 1998). The MSY1 dating was done as described for the Gypsy study above. A fourth, coalescence-based, dating method, assuming an expanding population, was devised and executed by Ian Wilson. All the calculated dates are given in Table 4.6.

	microsatellites		**MSY1**	
Method	**age**[a]	**95% CI**	**age**	**range**
Mean mutations from root[b]	2693	1154-9425	1107	604-3320
ASD[c]	3452	1480-12083	2632	1555-8554
Accumulated variance[d]	3116	1166-16001	2328	1217-10143
Coalescent[e]	1650	1044-8248	n.d.[f]	-

Table 4.6. Estimates of the age of the SRY-2627 mutation.
[a] Ages, and 95% confidence intervals (CI) are in years; [b] Bertranpetit and Calafell (1996); [c] Thomas et al. (1998); [d] Goldstein et al. (1995); [e] Wilson and Balding (1998); [f] n.d.: not done.

It can clearly be seen that there is good agreement between the different dating methods with five of the seven ages between 2300 and 3500 years old. As with the Gypsy lineage dating, the dates from MSY1 are generally younger than their microsatellite counterparts.

Putting these pieces together it seems likely that haplogroup 22 has a recent, Iberian origin. Linguistic prehistory of these two population groups is poorly understood and this, in combination with the wide confidence limits of the lineage age, makes it difficult to exclude the hypothesis that the lineage entered these two populations while they were speaking similar languages. Consequently it is of interest to see if we can date the split between the Basque and Catalan haplogroup 22 chromosomes. This can be done using the $\delta\mu^2$ measure as described in the study of the Gypsy-specific lineage. Thus the calculated value for $\delta\mu^2$ of 0.115 indicates that the age of the split between Basque and Catalan haplogroup 22 chromosomes is only about 20% of that of the lineage. This provides much stronger evidence that this gene flow occurred recently whilst there was indeed a language barrier between the two populations.

There does not appear to be a similar sharing of mtDNA lineages between Basques and Catalan populations; indeed these two populations are the most differentiated of all Iberian populations with respect to maternal lineages. This raises the interesting hypothesis that this gene flow may well have been specific to males.

Discussion

The comparison of different methods for displaying intra-allelic multiallelic diversity graphically clearly indicates the benefit of using network methods over multivariate analyses. The extension of this kind of comparison to other datasets could be used to further support this conclusion. It may be that there are peculiarities within this dataset which make networks better able to resolve the sublineages, and other datasets may not give the same impression. Having said that, this comparison still provides good support for considering network methods preferable to multivariate ones. My preferred method for displaying intra-lineage diversity would be the M-J network. In using M-J networks over M-S networks, not only are the summed equally parsimonious trees of shorter length but in using unobserved haplotypes as nodes M-J networks probably represent a closer approximation to the underlying evolutionary processes.

When constructing networks it is important to keep in mind the motive behind this kind of analysis, which to my mind is to provide a graphical display that might tease trends out of the data not visible on simple inspection of the haplotypes. These trends can then be used to guide statistical analysis. Networks are not an end in themselves. M-J networks can be further reduced by the inclusion of extra information such as differential-weighting of loci. This process reduces reticulations and makes the network more 'tree-like'. This approach is attractive to some researchers as they hope to arrive at a closer approximation to the real tree (Bandelt et al. 1995; Bergen et al. 1999); however the chance of obtaining the real tree is infinitesimally small and whilst this approach may exclude some of the less likely equally parsimonious trees from the network, the chance that the real tree has been excluded increases. What has been gained from reducing the network? Very little, yet important information may have been lost.

The only circumstances under which reducing a network might be sensible is when a network is so reticulated that it is difficult to make out any structure at all. This might occur when a large number of closely related haplotypes are being considered. However by taking the genealogical approach and only constructing networks within lineages the chance of this occurring is slight, especially given that the number of lineage-defining unique markers is expected to rise dramatically on publication of Peter Underhill's long-awaited DHPLC survey of Y-chromosomal diversity (Peter Underhill, personal communication). Certainly the need to reduce networks for this reason has not been apparent in any of the data sets on which I have worked, one of which included a network of 118 different MSY1 codes (see chapter 6).

Networks also provide insight into mutational mechanisms, especially with regard to networks of MSY1 diversity. Saltatory mutations can be easily spotted as successive mutations in the same block of repeats in the same direction. Networks can also be used as evidence for repeat-switching at boundaries between blocks of repeats of different types. Independent support for the existence of both saltatory mutations and boundary switching comes from studies of deep-rooting pedigrees (Jobling et al. 1999). This information can then be fed back into dating calculations as was done in the analysis of the Gypsy-specific lineage. The direct analysis of mutation events at MSY1 by small pool PCR should allow investigation of these mutational phenomena at far greater resolution than network analysis of extant diversity. However prior to the publication of this work, network analysis of this kind can inform dating analyses.

The utility of superimposing population affiliation upon networks has already been demonstrated (Zerjal et al. 1997) and is further validated by its use here. The network of the Gypsy-specific lineage clearly showed population subdivision whereas that of haplogroup 22 did not. Both of these observations were confirmed by statistical methods, specifically the correlation between the degree of apparent subdivision and the age of the population splits relative to the age of the lineages.

Within published studies there is an undefined line between using recurrent multiallelic markers to subdivide lineages identified by unique biallelic markers and using the same markers to quantify diversity within the self-same lineages (Thomas et al. 1998). To some extent the issue must focus on the mutation rate of the marker and thus how recurrent it is likely to be. It seems to me that using individual microsatellite alleles to subdivide lineages as a general approach is irresponsible given the known extent of recurrence for such alleles. Although some microsatellite alleles result from saltatory mutational events that can be considered to be unique (Forster et al. 1998), in general microsatellites are best used for quantifying diversity as haplotypes. MSY1 code diversity within a given modular structure can also inform diversity quantification. We shall see in chapter 6 that MSY1 modular structures can define sublineages, within lineages previously defined by unique markers.

Having said that individual microsatellite alleles are not useful for defining lineages, what about compound haplotypes of microsatellites? Whilst individual compound haplotypes are unlikely to be informative, related groups of such haplotypes may be. Sometimes the resolution of lineage detection attainable with unique markers is not sufficient. Networks of intra-lineage diversity, as in the Gypsy study, are capable of revealing substantial clustering of related haplotypes indicating the

presence of a monophyletic sublineage. This method for lineage detection is supported by the subsequent finding of a unique marker which defines the same set of chromosomes (Luba Kalaydjieva, personal communication). The use of microsatellite data and MSY1 data in the same network improves the chance of resolving between such sublineages.

Whilst this looks like a promising approach, the assumption that a lineage can be defined as comprising haplotypes a certain distance away from a modal haplotype is obviously a step in the wrong direction, which can only lead to biased estimates of ages of lineages defined in such ways. Note that in the study of haplogroup 22, about the same age as the putative Cohanim lineage (Thomas et al. 1998), roughly half of the haplotypes differ by more than one microsatellite mutational step from the modal haplotype. Although the Cohanim study uses only six microsatellites rather than the seven in the Iberian study this would seem to indicate that the age of the Cohanim lineage has been underestimated.

Although lineage definition in the Gypsy study could be done by eye, a formalised, mathematical method of recognising when a cluster can be considered a monophyletic lineage needs to be developed. Geographical criteria can not be invoked to support a lineage because invariably the lineage is being defined with a view to investigating its distribution, in which case the argument becomes circular. What is needed is a set of statistical criteria that in no way biases the age of the resulting lineage. One way of achieving lineage definition might be to compare alternative groupings of haplotypes and compare some statistic of intra-group and inter-group distances, the lineage being defined as the grouping that maximises the ratio of intra-group to inter-group statistics. Empirical testing similar to that used to compare different graphical methods might be useful in this regard.

MSY1 provides an independent locus with which to test conclusions from microsatellite analysis, or combined with it, to provide greater resolution. Dating calculations on one type of locus can be verified against those from the other. Alternatively MSY1 diversity can be combined with microsatellite data, as in the networks on the Gypsy study, to provide greater resolution. In the absence of information from unique markers MSY1 modular structural diversity has also been used to quantify diversity (Jobling et al. 1998), but given the poor understanding and relative rarity of modular changes this approach is unlikely to be accurate and should not be used when lineages can be identified from other markers in the same samples.

In general there is striking agreement between age estimates for lineages provided by the three very different methods investigated here. In addition the development of lineage dating using

MSY1 in this chapter indicates the agreement between different types of multiallelic loci. This congruence can provide extra confidence in the mind of the investigator in the accuracy of age estimates.

The specification of a root haplotype is required to date lineages by two of the three methods considered here. In this chapter I have introduced two new criteria upon which to chose the root haplotype; that of combining the modal allele lengths and that of the choosing the haplotype that gives the minimum number of summed mutational steps to all other haplotypes. In both the lineage studies described here, these criteria define the same root haplotype. In general I would expect these criteria, to be more accurate than the defining of a root haplotype on the basis of connections between networks of derived and ancestral lineages (Zerjal et al. 1997). Such connections are likely to be highly dependent on sample size, the expectation being that if larger sample sizes were being considered, there would be multiple equally parsimonious links between the two networks. In summary the consideration of multiple criteria for root haplotype designation that do not rely explicitly on network connections, including the expectations of high frequency and multiple connections, is probably the most accurate way to define a single root haplotype. If multiple potential root haplotypes are defined by the different criteria they can be considered together within the framework of wider confidence limits on age estimates.

In practice, the wide confidence limits on Y-chromosomal lineage age estimates often fail to discriminate between the competing anthropological scenarios being investigated. This, together with the confusion of lineage ages with population ages, has led to a growing feeling against lineage dating in the field (Cavalli-Sforza and Minch 1997; Barbujani 1999). However, the utility of providing a temporal aspect to the pattern of lineage sharing should not be underestimated, especially when an age for a lineage is found that is younger than the supposed split between two populations in which it is found. As more mutational data are compiled on multiallelic loci the confidence limits around their mutation rates will decrease and correspondingly tighter confidence limits around age estimates will result. This does not of course take into account other ingredients of uncertainty, such as demographic history. An appreciation of the effect of demography on the estimates of various population parameters needs to be incorporated into lineage analysis. For example the effect of expanding, contracting or constant population sizes on the extant multiallelic diversity of a lineage needs to be considered. This is likely to have the effect of increasing still further the confidence limits around age estimates. There are few published examples of attempts to achieve this, most are based on simulations, some invoking the coalescent process (Thomas et al. 1998).

The study of the Gypsy-specific lineage provides a striking example of the biological impact of social tradition. Whilst the three Gypsy groups studied here have a recent common origin, the substantial differences between them indicate a significant barrier to mixing between these endogamous groups. The finding of greater admixture between Gypsies and non-Gypsies of maternal rather than paternal lineages is supported by the higher tolerance to female outsiders marrying into Gypsy groups observed by ethnologists (Piccolo et al. 1996).

The Basques are considered by many to be a Mesolithic isolate (Collins 1986); the sharing of the haplogroup 22 lineage indicates that this can not be so. Either the Basque population maybe Mesolithic and recent gene flow explains the lineage distribution, or it is an isolate and a very recent common origin with Catalans explains the present situation. We prefer the former explanation.

The finding in the next chapter of a genetic barrier around the Basques is not necessarily contradictory in this respect. Firstly a Catalan population was not used in the European study and secondly barriers are defined relative to the rest of the landscape whereas the term 'isolate' implies an absolute lack of gene flow.

The fact that the Y chromosome carries a wealth of different polymorphic markers within a single, large, non-recombining locus means that in principle we can determine more about each lineage, having identified it, than we can with lineages at other, analogous, loci. Attempts to use intra-allelic diversity at autosomal loci often rely on one or two closely positioned microsatellites and certainly cannot call upon the same number of microsatellites as are considered here. Dating of mtDNA lineages has been attempted using base-substitutional diversity within the hypervariable sequences of the D-loop (Richards et al. 1996; Richards et al. 1998) .

Both of the lineage studies detailed here involve comparisons between Y-chromosomal and mtDNA diversity to develop hypotheses about relative differences between the sexes. This is becoming a fast-growing area of publication e.g. (Perez-Lezaun et al. 1999) although some cautionary opinions as to the validity of such conclusions have been expressed (Barbujani 1999). It must not be forgotten that the Y chromosome and mtDNA are single loci and therefore highly likely to show different evolutionary pictures purely through the stochastic nature of evolution. So far the field has not explicitly addressed the issue of how to tell the difference between real sex-specific differences and stochastic between-loci variation. It is not easy to conceive of tests to discriminate between these alternatives, as both loci represent the sole uniparentally-inherited loci in the human genome, though perhaps the X chromosome, which passes through twice as many female meioses as male, could be used to investigate this (Kaessmann et al. 1999).

A related cautionary tale lies in the fact that by having such small effective population sizes, in conjunction with the effect of sampling inherent in such studies, rare mtDNA and Y-chromosomal lineages can exhibit unusual world-wide distributions that may not have any basis in population history. A good example of this is the sharing of haplogroup X of mtDNA at low frequency between Amerindians and Europeans, but not Asians (Brown et al. 1998).

What further developments can be envisaged to improve the resolution of the type of lineage analysis considered here? At present, analysis of MSY1 diversity is confined to single modular structures and thus to relatively young lineages. Older lineages can reasonably be expected to exhibit greater MSY1 modular structural diversity. The extension of both network and dating analyses to multiple, related, modular structures depends to some extent on a better appreciation of the mutational dynamics of modular structural changes. Due to the assumed low mutation rate of such changes this might well be beyond the resolution of small pool (SP) PCR experiments. Consequently initial attempts to use multiple modular structures might have to make several explicit assumptions that can subsequently be refined. However, as the resolution of lineages attainable using biallelic markers increases these types of analyses will become less necessary.

As the human genome project reaches its conclusion more sequence data from the Y chromosome will become available for analysis. A rapid increase in the identification of useful multiallelic markers can be expected (Chris Tyler-Smith, personal communication). This will allow us to gain even more quantitative information on individual lineages and identify more accurately sublineages based on haplotype clustering.

An issue that has not been addressed here is how to sum information between multiple shared lineages between the populations of interest. This is a relevant concern for studies based on entire populations rather than single lineages. Obviously disregarding lineage information is not an option, given the multiple advantages a genealogical approach bestows, as was advocated in the introduction to this thesis. If two lineages have the same history in the populations of interest then we can reasonably expect them to give the same quantitative estimates of parameters such as the age of a population split. This is because it is populations which move and undergo gene-flow, and not individual lineages. In this case averaging answers from multiple lineages provides the best way of estimating these parameters. Discrepancies between estimates from multiple lineages may well reveal the magnitude of the effect of the stochastic nature of the evolutionary process on the confidence limits of such estimates. The major disadvantage with this method of analysis is that by reducing each population to a number of lineages the sample sizes can be dramatically reduced and

correspondingly the sampling effect magnified which will result in greater confidence limits around quantitative estimates. Consequently one aim of future studies in this field should be a substantial increase in typical sample sizes.

Having said that, it can be envisaged that certain lineages shared between two populations have different histories, in which case divergent estimates of quantitative parameters may result. The example of such a scenario is if only one of the two populations receives admixture from a third population that contains a lineage that was already shared between the first two populations. Hopefully, the adoption of a genealogical approach allows the researcher to distinguish these divergent lineage histories prior to parameter estimation. Were a genealogical approach not adopted and admixture not accounted for, then parameter estimation on the basis of entire populations might well be severely biased by the inclusion of data from anomalous lineages. These issues will be further discussed in chapter 6.

One issue that has not been addressed here is that of calculating gene-flow as opposed to population splits. Both calculations derive from estimates of population distances and involve assumptions about the nature of population evolution. They are usually treated as being mutually exclusive alternatives. In this chapter I have dealt solely with dating population splits. The reality of population evolution is that both the age to a common ancestral population and the amount of subsequent gene flow contribute to the amount of divergence between two populations. Models are needed that can take account of both, within a framework of lineage analysis.

Chapter 5: Spatial analysis of genetic diversity: detecting barriers to gene flow in Europe

Introduction

The Y chromosome evolves in a genealogical fashion, consequently the lineage is the correct unit of *inquiry*, yet often the unit of *interest* is the population, or in the case of regional studies, groups of populations (Avise 1989). This chapter is concerned with attempts to reconcile this apparent disharmony.

Investigation of the distribution of diversity within space is a goal common to many genetic studies. It can provide inferences on population histories and structures, and provide insight into evolutionary processes themselves (Barbujani 1999).

As with all molecular evolutionary study, the field of spatial analysis of genetic diversity has had to adapt to the changing nature of the underlying data. The transition from protein allozyme polymorphisms to DNA-based data required that analyses be developed that could take account of the new information about the mutational distance between different alleles (Barbujani 1999). A recent further development has seen phylogenies themselves become the centre of attention, founding the new field of phylogeography (Avise 1989; Avise 1998). It is this field that most directly addresses the potential disharmony between lineages and populations. Having said that, I shall argue later in this chapter that in some analyses it may be better to ignore the underlying phylogenetic information.

The incorporation of geographic information into analyses of diversity is a wider field than evolutionary genetics. As a result, when seeking new analytical methods, spatial analysis can often usefully borrow from related fields in other disciplines, specifically geostatistics, where problems common to both fields have been being considered for longer.

Genetic spatial analysis is, by its very nature, an integrative discipline, relying on information from other more established disciplines. Apart from genetics and geography, information on demography, ethnology and history can also be incorporated into spatial analyses.

In anthropological studies language and ethnology provide additional sources of information that can be incorporated into a common analytical framework.

The concept of phylogeography

Phylogeography is a relatively recent concept. The word itself was first coined by John Avise in 1987 (Avise et al. 1987). Since then, the number of published articles using this concept has increased almost exponentially (Avise 1998). As with other rapidly-adopted neologisms, phylogeography provided a label for a pre-existing, but as yet undefined, field; namely the interface between population genetics and phylogenetic systematics, especially in intraspecific studies. Phylogeography is effectively a geographically-informed extension of the analyses, raised in the previous chapter, that consider how genetic lineages are apportioned between groups and whether any differences between populations are significant.

Phylogeography was initially defined in relation to the geographical distribution of a mtDNA phylogeny (Avise et al. 1987). This molecule has provided the mainstay of phylogeographic work for reasons outlined in the Introduction to this thesis. The definition of phylogeography, *sensu strictu,* relies on genetic data in the form of phylogenies, but its use has since expanded to include work on other loci that have known mutational distance between alleles but do not necessarily form phylogenies, such as unlinked microsatellites.

Applications of spatial analysis

The vast majority of studies that seek to explain the spatial distribution of genetic diversity have investigated species other than *Homo Sapiens*. To a large extent the human and ecological research communities have remained distinct despite their obvious common interests, though they often use similar analytical techniques.

Genetic spatial analysis has been used extensively in both the Plant and Animal kingdoms to inform conservation strategies, breeding strategies and other ecological studies as well as the evolutionary scenarios of individual species. Chloroplast DNA (cpDNA) has proved to be a useful tool, analogous to mtDNA, for the study of plant phylogeography (Dumolin-Lapegue et al. 1997). Much of the non-human work has focused on either endangered species or those of agricultural importance (Petit et al. 1998).

Spatial analysis of genetic diversity can provide insight into the evolutionary and ecological consequences of certain behavioural traits. Analysis of maternally-, paternally- and biparentally-inherited markers can reveal sex-specific differences in variation, which can be related to mating strategies and other sex-specific behavioural traits (Piertney et al. 1998).

A recent development has been the comparison of the phylogeography of different species that co-exist within the same ecological and evolutionary framework (Bermingham and Moritz 1998). Commonalities between multiple species can provide information on the evolution of the landscape. For example the present distribution of European Oak and Elder diversity, showing that both species expanded out from common glacial refugia, indicates how vegetation responded to ice ages (Dumolin-Lapegue et al. 1997). Dubbed 'comparative phylogeography' this approach has recently been applied to the human colonisation of the Pacific. In this case the commonalities between different species does not indicate the effect of landscape on evolution but rather the effect of man on his environment. The lizard, *Lipina noctua,* and the rat, *Rattus exulans*, were both first brought to the outer Pacific islands in the canoes of the Polynesian colonists and consequently all three species exhibit a common spatial apportionment of their phylogenies which strengthens support for a specific model of human colonisation (Sykes et al. 1995; MatisooSmith et al. 1998; Austin 1999). This will be discussed in greater depth in chapter 6.

Different spatial analyses

Phylogeographic conclusions can be attempted by mere casual observation of the overlaying of a genealogy upon the geography in which it is found. Such methods do not inspire confidence and are impossible when multiple lineages are shared between many well-separated sample sites. Consequently a battery of analytical techniques, some, but not all, dependent on phylogenies, have been developed to bring a much needed statistical rigour (Barbujani 1999).

Different models of population genetics provide different expectations of how genetic diversity should be spatially distributed. An example of this will be shown in the next section. Although such models require many assumptions and oversimplifications of real landscapes, their expectations still provide useful null hypotheses for testing with real data.

As outlined in the previous chapter the most important null hypothesis to investigate is that diversity is distributed randomly, as a result of panmixia. Although few populations, especially human ones, conform to panmixic expectations, the resolution apparent in the data may not be sufficient to reject panmixia. Consequently the rejection of this hypothesis is a useful first start to any analysis (Slatkin 1995). Often the distribution of the statistic being investigated is unknown under the null hypothesis, consequently many different analytical techniques use Monte Carlo sampling, also known as permutation testing, to generate this unknown distribution (Manly 1991). This procedure involves calculating the statistic multiple times from independent resamplings of the observed data by assigning random localities to individual sequences or haplotypes. These permutation tests can vary according to how the data are permuted (Manly 1991).

Some analyses seek only to test a null hypothesis, whereas others seek, in addition, to provide reasons for why such a null hypothesis does not apply to the population in question. One weakness of the former class of analyses is that having rejected a null hypothesis the investigator is then open to explain the non-random distribution of diversity with a hypothesis, either from a population genetics model or from a complementary discipline, that seems compatible (Templeton 1997). This approach has been criticised as being open to biased interpretations and not being sufficiently question-driven (Templeton 1998). Ideally if a hypothesis is being considered it should be tested and not simply deemed compatible.

There are many reasons why lineages may exhibit geographical association, in other words, a non-random distribution within the range of the species. Such an association may be due to population structure as a result of the phenomenon of isolation-by-distance (IBD) or to population history through such events as colonisations or fragmentations (Templeton 1998; Barbujani 1999). In addition selective forces can often generate strongly non-random lineage distributions and a number of examples of this occurrence, that relate to infective disease, are well known (Silvestroni and Bianco 1975). Disentangling the relative influences of these different factors within the dimensions of time, space and phylogeny is the aim of spatial analysis. There is no one analysis that can address all these issues, rather a mix of competing and complementary analyses that go some way to achieving this aim (Barbujani 1999).

One analysis that explicitly seeks to test competing hypotheses for the non-random distribution of phylogenetic diversity is Templeton's Nested Cladistic Analysis (NCA). It is claimed that this analysis is capable of distinguishing between different scenarios for the non-random distribution of clades (Templeton et al. 1995). NCA calculates geographical distance statistics for specific clades to test the null hypothesis that each clade within the phylogeny in

question is randomly distributed. Once this null hypothesis has been rejected a number of these distance-related statistics are analysed in conjunction with an inference key to identify whether population structure or population history is to account for the observed distribution (Templeton et al. 1995). NCA has been used recently in studies of the human Y chromosome to support the theory of a substantial back migration to Africa (Hammer et al. 1998).

One strength of this method is that it independently tests all clades within the phylogeny by virtue of its nested design. Consequently one region of the phylogeny may be randomly distributed whereas another is non-randomly distributed due to population structure and a third exhibits the effects of population history. Different clades can tell different stories. A weakness is that it requires a certain degree of confidence in the phylogeny, a confidence that is often lacking in mtDNA phylogenies due to homoplasy from recurrent mutation. The validity of this method has also been questioned on the basis of its use of an inference key. Templeton has gone to great lengths to try and validate the assumptions within this inference key (Templeton 1998), yet it remains intuitively unsatisfactory to many people.

In common with many other phylogeographic analyses, NCA uses permutation as the basis for testing the null hypothesis of no geographical association (Roff and Bentzen 1989). The test used permutes the observed data in such a way as to preserve both the overall frequency of a lineage within the entire dataset and the sample size at each locality sampled. Consequently the power of such an analysis is often compromised by large variations in sample size (Templeton 1998).

Spatial autocorrelation is a phenomenon common to many spatially-distributed variables, whereby samples within a region are not independent of each other, but are related through geography (Bertorelle and Barbujani 1995). The analysis of spatial autocorrelation uses levels of genetic resemblance calculated between pairs of localities within arbitrary distance classes to investigate the nature of this relationship (Barbujani 1999). The degree of spatial autocorrelation found within different distance classes can be displayed on a correlogram. Measures of resemblance have been modified to take account of mutational distance between haplotypes (Bertorelle and Barbujani 1995). A permutation test can be used to reject the null hypothesis of no spatial autocorrelation.

It is claimed by its advocates that the nature of the spatial autocorrelation between samples within a region can be used to distinguish between competing explanations for the non-random association of genetics and geography (Barbujani et al. 1994; Barbujani et al. 1994; Chikhi et al.

1998). The shape of the correlogram is meant to be indicative of the cause of autocorrelation. Simulation work has been used to back up this hypothesis (Barbujani et al. 1995) and much work has focused on the contentious issue of whether genetic spatial autocorrelation is due simply to the effect of isolation by distance (population structure) or results from the existence of a cline related to prehistoric population movements (Chikhi et al. 1998). Barbujani and others have used spatial autocorrelation analysis in European populations to support the existence of clines that were created by the demic diffusion process accompanying the Neolithic transition (Barbujani et al. 1995).

The Mantel procedure is a test of matrix correspondence widely applied in fields as diverse as geography, psychometrics and population biology (Mantel 1967). It tests for correspondence between two matrices of distances by calculating the sum of cross products. In the case of spatial analysis these matrices are of genetic distances and geographic distances between population pairs. The significance of an association can again be tested using Monte Carlo randomisation, whereby one matrix is held constant and the other is randomly permuted multiple times and the association of the new arrangement determined. Measures of genetic distance that incorporate mutational distance between haplotypes can be used to make this analysis suitable for DNA data.

More recently this test has been modified to make it more suitable for genetic analysis (Smouse et al. 1986) . It has also been extended to allow the examination of partial correlations of multiple non-independent matrices. These multiple matrices (often three) may all be strongly associated to one another, but partial correlation allows the two matrices with the strongest association to be identified by controlling for the indirect effect of the third matrix. This allows the ascertainment of the relative effect on population evolution of different factors.

Human studies using this approach have often focused on the conflated effects of language and geography on genetic distance (Poloni et al. 1997). Geography and language can not be considered to be independent as the two are themselves often highly correlated. This analytical method can be used to show that languages and genes correspond independent of geography (Lum and Cann 1998). This approach has also been used to investigate whether male or female lineages correspond better with linguistic distances (Poloni et al. 1997). The results indicate that on a global scale language has been inherited more strongly through paternal rather than maternal lineages, contrasting markedly with the idea of 'the mother tongue' (Barbujani 1997).

An interesting application of these partial Mantel tests has sought to use a database of ethnohistorical population movements within Europe, initially to show the correspondence with

genetic distances (Sokal et al. 1993; Sokal et al. 1996) but more recently to investigate whether history or geography is a better predictor of cancer mortality throughout Europe (Sokal et al. 1997).

Interpolation

Given a matrix of values of a variable at regularly spaced sites a contour map in 2D, or a surface map in 3D, of the landscape of that variable can be constructed (Lam 1983). The most familiar of these are the contour lines on an ordnance survey map that indicate the variation of altitude throughout a landscape. Such maps provide a visual way of making inferences about natural selection gradients, local gene flow and historical migration patterns (Piertney et al. 1998; Barbujani 1999). In addition to their value in constructing intelligible maps, these matrices can also be used for analyses that seek to further investigate this landscape (Barbujani et al. 1989).

It is rare that genetic data, especially those concerning humans, are sampled at regular intervals. This is a problem common to many fields of geography. However spatial interpolation provides a means to generate the matrices discussed above and as such is the mainstay of geostatistics (Lam 1983). An algorithm is used to estimate the value of the variable at regular spaced grid points, throughout a region defined by the extremities of the sample sites. There are a wide variety of different algorithms available for use, varying in their complexity (Lam 1983). These algorithms relate the previously unknown value of the variable at the grid point to the known values of the variable at surrounding sample sites, by means of a mathematical model. In addition all interpolation procedures include a searching method to identify the sample sites from which to interpolate. Searching functions can be simple, finding the closest N sites, or all sites within a certain radius, or they can be complex, searching each quadrant for sample sites and weighting each quadrant differently (anisotropic).

It is commonly recognised that there is no one interpolation algorithm which provides the best results for all data sets (Lam 1983). Other criteria also come into play when choosing which algorithm to use, not least the computational effort required (Barbujani et al. 1989; Fortin and Drapeau 1995). Algorithms can be classified in a number of different ways. A common subdivision is between those algorithms that retain the values at the sample sites within the surface (so-called exact interpolation methods) and those that do not (approximate interpolation methods) (Lam 1983). In addition some methods are only capable of assigning values that lie within the range observed at the sample sites: these are considered to be smoothing algorithms. Many people are

dismissive of such algorithms as they are unable to reconstruct important features of the landscape, such as maxima and minima, unless they happen to fall at one of the sample points.

Here I shall consider two of the most widely used algorithms, at opposite ends of the complexity spectrum.

Distance-weighted algorithms are simple to apply and are widely used in many fields. The principle is that an algebraic expression is used to assign more weight to nearby sample sites by using a weighting function inversely related to distance. The most commonly used weighting function is that of the inverse-distance squared. The weighting function at distance = 0 is ∞, and consequently inverse-distance weighting methods are exact. Inverse-distance weighting is effectively an averaging process as only values that lie within the range of those found at actual sample sites can be obtained. Therefore inverse distance weighting can be considered a 'smoothing' interpolation method (Lam 1983).

Kriging is a very much more complex interpolation procedure and was first derived to improve the estimation of reserves in mining (Matheron 1971). It takes account of the spatial autocorrelation between samples that decays as the distance increases. Initially a semi-variogram is constructed that relates the difference between sample values and the distance between sample sites. This semivariogram provides certain parameter values that are then fed into interpolation (Lam 1983). The interpolation algorithm uses a linear combination of weighted sample values, the weights being dependent on the semi-variogram.

Kriging is also an exact interpolation method but is not a smoothing function (Lam 1983). It can be shown mathematically that the resulting estimate for the variable is the best linear unbiased estimate. As well as providing sample values at grid points kriging can also provide error values for each unknown variable estimate and so confidence limits can be placed on features within the landscape. Unsurprisingly, in light of the complexity of the method, kriging is computationally intense.

One problem with interpolation is that only a single variable can be considered at a time. It is impossible to infer the history of an entire locus from a single allele, especially when the number of alleles is large (Barbujani 1999). Different alleles can tell different stories and it is difficult to see how to sum this information. Consequently many studies have sought to use multivariate statistics to maximise the amount of information present in a single variable. Typically Principal Components

Analysis (PCA) is applied to genetic data, and the first principal component, which often summarises most of the variance within a data set is used as the variable to be interpolated (Piertney et al. 1998). It has been suggested that successive principal components can also be interpolated to identify population histories of diminishing importance. This procedure has been followed in the extensive work by Cavalli-Sforza, Mennozi and Piazza, published in 1994, which used allele frequency data to generate interpolated maps of principal components throughout all regions of the world (Cavalli-Sforza et al. 1994). These maps were subsequently interpreted in the light of the other disciplines that inform our view of population movements in prehistory; archaeology and linguistics.

Barrier detection

Once a landscape of a given variable has been constructed, a number of questions often arise: How variable is the landscape? Are there regions which can be considered to be homogenous? Are there regions of sharp variation? There are two statistical approaches to such work: the reduction of the landscape into regions of clustered, spatially-adjacent, sites of sufficient similarity or the delineation of boundaries by the use of edge-detection algorithms (Fortin and Drapeau 1995). As with other biological fields, 'lumpers' and 'splitters' can be found. Here I shall only consider boundary delineation, or barrier detection, as it is more often known.

Boundaries within a landscape of an ecological variable are referred to as ecotones and often correspond to ecological important zones of change. For example changes in soil type can cause a rapid change in types of vegetation (Fortin and Drapeau 1995). On a larger scale, regions of sharp change within a continental landscape of allele frequencies may result from the presence of a geographical barrier that causes a decrease in gene flow across it (Barbujani et al. 1989). A region of reduced gene flow is known as a genetic barrier. Apart from geography there are other potential underlying causes of the formation of genetic barriers (Barbujani et al. 1989). Within the field of human prehistory lies the possibility that cultural differences may underpin genetic barrier formation (Barbujani and Sokal 1991). The investigation of this possibility with respect to the human Y chromosome is the focus of this chapter.

There are a number of analyses focusing on barrier detection within a landscape, although not all of them require prior interpolation (Fortin and Drapeau 1995). The main requirement of these techniques is that they allow information from different alleles to be summed to present a

composite picture. The strength of techniques that rely on interpolation is that they may sum information from surfaces which, though they cover the same region of interest, were constructed from sample sites that differ between the surfaces. Consequently independent data sets can be combined into a single analysis.

As mentioned earlier, selection is known to play a role in generating non-random spatial distributions of allele frequencies, for example haemoglobin alleles that protect against malaria (Silvestroni and Bianco 1975). A rapid change in the relevant selective environmental factor could therefore result in a rapid change in gene frequency of the locus under selection (Barbujani 1999). Consequently there is potential conflict as to whether a barrier in an individual surface results from selection or some facet of population history. These two causes can be disentangled by considering multiple surfaces from unlinked loci (Barbujani 1999). Whereas selection can be expected to affect a single locus, an impediment to gene flow can be expected to reveal itself in barriers in multiple surfaces. Therefore when barriers are found within an analysis that sums information from multiple surfaces, individual surfaces can be inspected separately to discriminate between these two explanations.

Womble proposed a method for barrier detection in 1951 that formalised the time-honoured approach of identifying, by eye, regions of closely spaced contour lines within a landscape as a barrier (Womble 1951). If the derivative of such a surface were calculated and the surface of the derivative plotted, regions of sharp gradient would be represented by peaks, regions of gradual change in the original surface would become non-zero plateaux in the derived surface, and regions of homogeneity would become plateaux at zero. Figure 5.1 illustrates this and indicates that the latter two scenarios are precisely those expected under the population genetic models of isolation by distance and panmixia respectively. Womble suggested that such derived surfaces could be averaged together to form a "systemic function" that sums the information from all surfaces. This barrier detection technique has become known as 'wombling' (Barbujani et al. 1989).

A number of approaches have used this general methodology, and consequently wombling analyses have been further subdivided (Fortin and Drapeau 1995). When interpolated surfaces are used, as described above, it is known as 'lattice-wombling'. When sample sites are reasonably well distributed an alternative, less computationally intense, wombling approach has been applied which does not require prior interpolation. These techniques rely on a procedure from geostatistics known as Delaunay triangulation, in which three sample sites are joined to form apices of a triangle when there is no sample site that lies within that triangle. A convex hull of localities can be reduced to a planar surface by this procedure. In 'triangulation-wombling' the Delaunay triangles can be used as

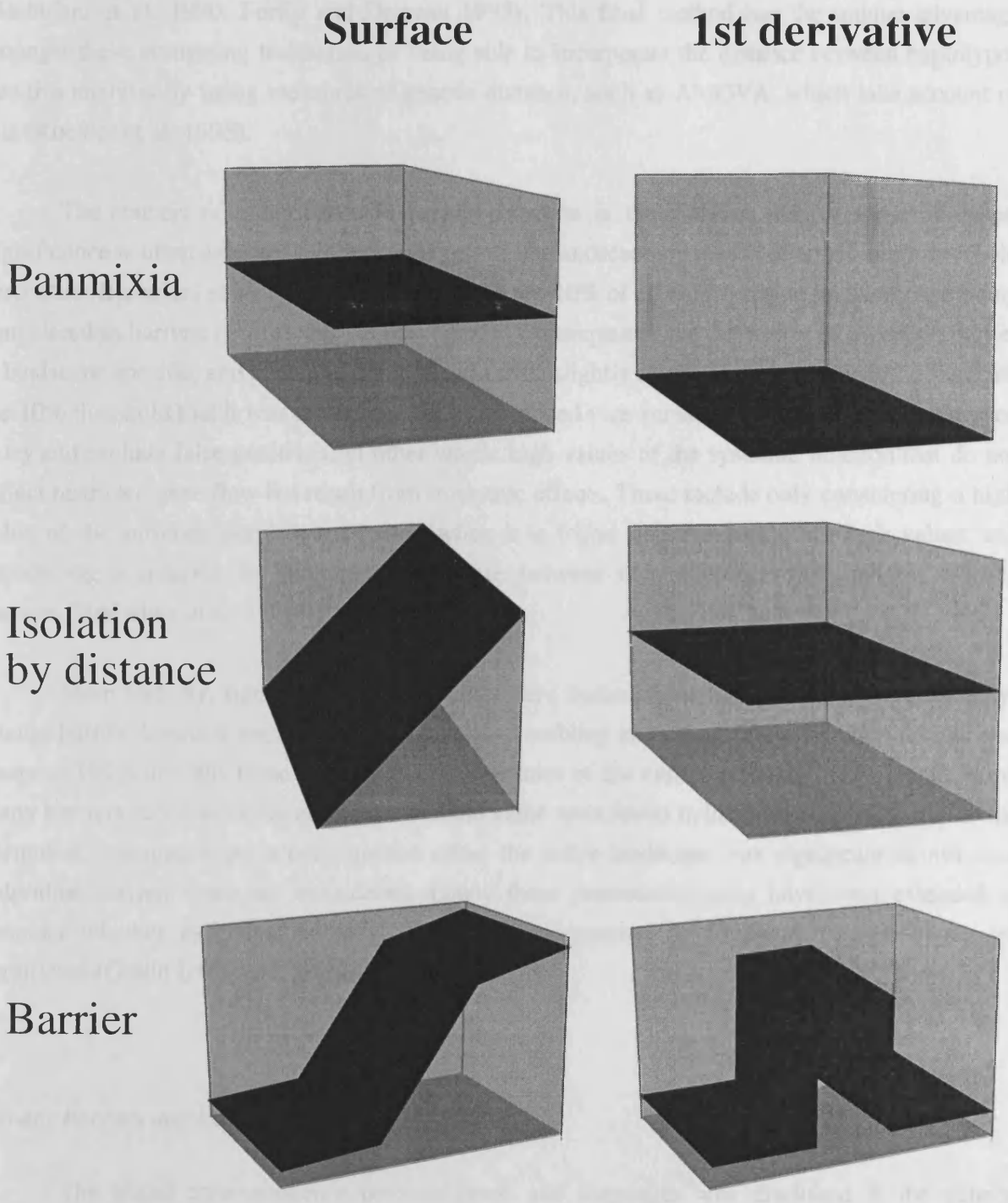


Figure 5.1: Surfaces of allele frequencies expected under different models of the spatial distribution of genetic diversity and the corresponding first derivatives of those surfaces.

the input for barrier detection by calculating and summing the rates of change at the centroid of the triangle for each allele frequency (Fortin and Drapeau 1995). In contrast 'categorical-wombling' utilises the edges of the Delaunay triangles as connections for comparisons of genetic distances (Barbujani et al. 1990; Fortin and Drapeau 1995). This final method has the unique advantage amongst these competing techniques of being able to incorporate the distance between haplotypes into this analysis by using measures of genetic distance, such as AMOVA, which take account of this (Roewer et al. 1996).

The concept of significance in barrier detection is the weakest part of these analyses. Significance is often assessed relative to the rest of the landscape by means of an arbitrary threshold percentile (Barbujani et al. 1989), for example, the top 10% of all values within the landscape being considered as barriers (Fortin and Drapeau 1995). Consequently the definition of a genetic barrier is landscape specific, and changing the landscape even slightly might cause a barrier to fall beneath the 10% threshold that it was previously above, or indeed vice versa. Other criteria have been used to try and exclude false positives, in other words high values of the systemic function that do not reflect restricted gene flow but result from stochastic effects. These include only considering a high value of the summed derivative a barrier when it is found neighbouring other high values and introducing a criterion for directional coherence between neighbouring values before defining barriers (Barbujani et al. 1989).

More recently, rigorous permutation tests have been defined for the less computationally-intense barrier detection methods of triangulation-wombling and categorical-wombling (Fortin and Drapeau 1995). Initially these only tested characteristics of the entire landscape, for example, how many barriers above a certain absolute threshold value were found in the observed, compared to the permuted, landscapes. As a consequence either the entire landscape was significant or not, and individual barriers were not considered. Lately these permutation tests have been extended to consider whether individual barriers between sites connected by Delaunay triangle edges are significant (Guido Barbujani, personal communication).

Genetic barriers and languages

The global correspondence between genes and languages was discussed in the general introduction and here I want to consider the relationship on a more intimate, regional scale. This is

a scale that linguists are more comfortable working with, in the realms of well characterised language families rather than contentious connections between language super-families.

Most of the studies that have sought to detect genetic barriers between human populations have focused on the relative contributions of cultural barriers and geographical barriers to the inception and maintenance of these barriers (Barbujani and Sokal 1991; Barbujani et al. 1992). Language boundaries are reasonably stable in space and represent the cultural differences most often hypothesised to represent barriers to gene flow (Barbujani and Sokal 1990), although religious beliefs and social attitudes have also been investigated (Barbujani and Sokal 1991).

Correspondence between genetic barriers and language boundaries could have arisen by two mechanisms. Firstly the processes that lead to language differentiation may have also affected genetic differentiation. For example two populations previously separated by a reasonable distance resulting in their having separate evolutionary and language histories, brought into recent juxtaposition such that insufficient time had passed to allow gene flow to negate their genetic differences, might appear to have a genetic barrier between them. Alternatively linguistic differences may represent a reproductive barrier allowing two spatially close populations to drift apart genetically (Barbujani and Sokal 1990). These two scenarios are not mutually exclusive and are analogous in some ways to allopatric and sympatric speciation respectively.

Initial attempts to relate genetic barriers to languages have focused on Europe. There are two good reasons for this; firstly Europe is the continent most densely sampled genetically (Cavalli-Sforza et al. 1994) and secondly the phylogenies of European languages and Indo-European in particularly are well characterised and agreed upon (Ruhlen 1991). The geographical locations of the various language families and subfamilies within Europe is shown in figure 5.2. Lattice-wombling provides a good way of combining all the available genetic information on allele frequencies at multiple loci sampled at different locations throughout the continent (Barbujani and Sokal 1990).

Barbujani and Sokal showed that the genetic barriers in Europe, revealed by lattice-wombling of allele frequencies at 19 different loci at 3119 European sample sites, correspond closely to linguistic boundaries (Barbujani and Sokal 1990). Thirty-three genetic barriers were recognised, of these thirty-one corresponded to linguistic boundaries, nine of which did not correlate to geographical barriers, thus implicating the importance of linguistic differences alone as a cause of reproductive isolation.

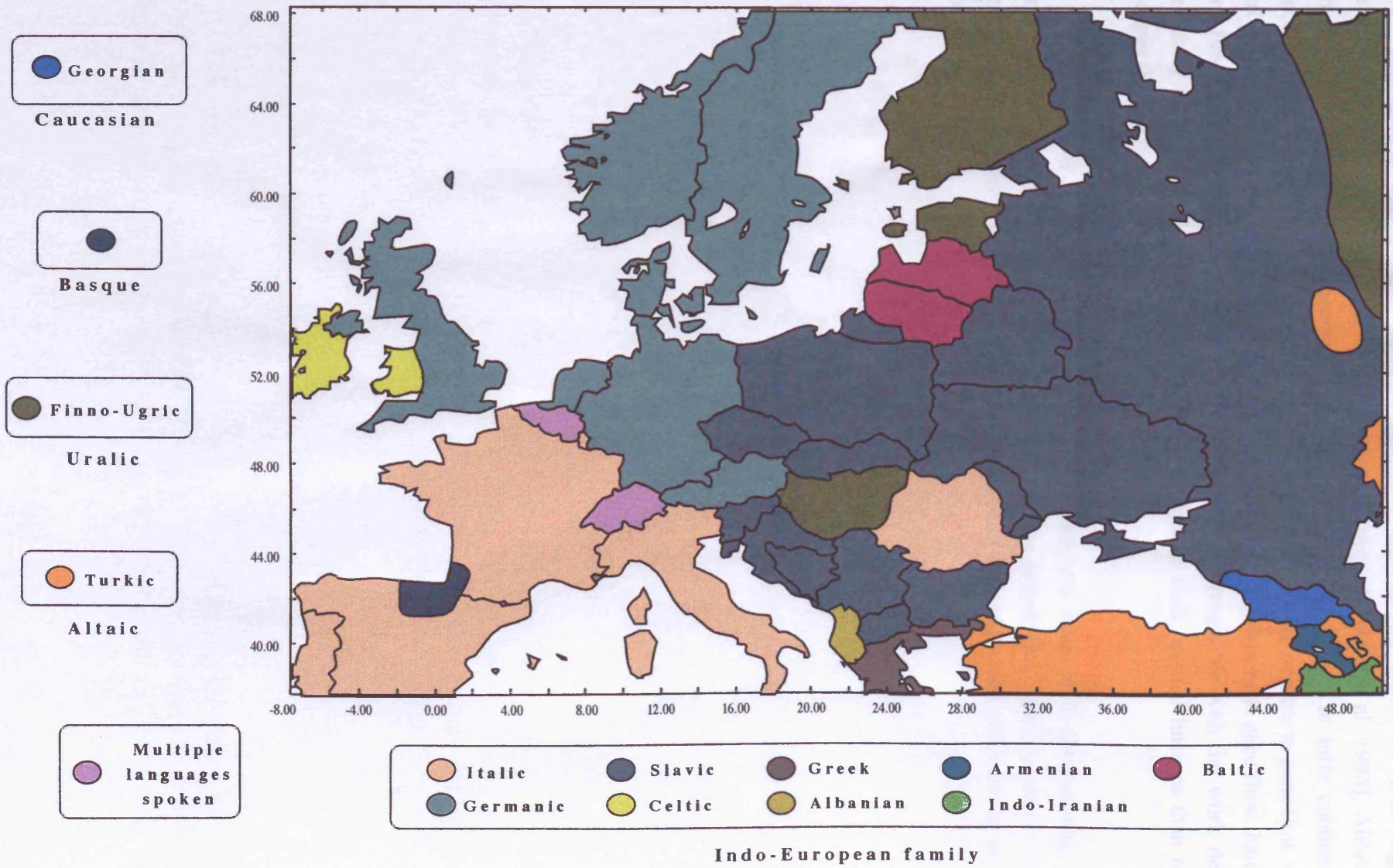


Figure 5.2: Map showing the distribution of different language families and subfamilies throughout Europe

A recent study on the European diversity of mtDNA failed to find significant correspondence of maternal lineages with languages in Europe (Sajantila et al. 1995). Although the sampling in this study was so unrepresentative as to make the attempt to infer continent-wide conclusions highly questionable, it raises an interesting issue. Are the barriers to gene flow noted in Europe from biparentally autosomal markers more due to a reduction in male gene flow than female gene flow across linguistic barriers? This would seem to be compatible with the work described previously that found a greater correspondence of languages with paternal lineages than maternal lineages (Poloni et al. 1997).

This chapter describes work performed to address the issue of Y-chromosomal genetic barriers within Europe and their congruence with linguistic boundaries. In the process I present a computer program written to perform a permutation test of individual barriers identified by lattice-wombling.

Materials and methods

Hardware

All writing and running of programs was performed on a G3 PowerMac with a 233MHz processor and 64Mb of RAM. In order for the program to run successfuly, virtual memory had to be raised to twice the size of the RAM.

Software

All three programs, EuroBarrierSig, DisplayBarrierSig and EuroDelaunay were written in Interactive Data Language 5.1 (IDL) from Research Systems Inc. IDL is an array-based higher order language containing many subroutines useful for geographical analyses. The runtime for the major program 'EuroBarrierSig' was 4 hours. The modular designs of the two programs that combine to form the wombling analysis are shown in figure 5.3. The source code for the programs is given in appendix D.

Data

The data used for the European analysis came from a collaborative study of haplogroup distributions in different European populations. Eight different lineages were assayed in 43 populations. The vast majority of the data was generated by Zoë Rosser in the dept. of Genetics, University of Leicester and Tatiana Zerjal in the dept. of Biochemistry, Oxford University.

Designing and writing the programs

IDL was chosen as the language in which to write the programs because of its flexibility with respect to geographical analysis. IDL contains many map manipulation routines and in addition holds within it low resolution CIA maps that can be called upon to display with data. IDL

is used commonly in the fields of Earth Observation Science (EOS) and geostatistics, fields which have similar requirements to this analysis.

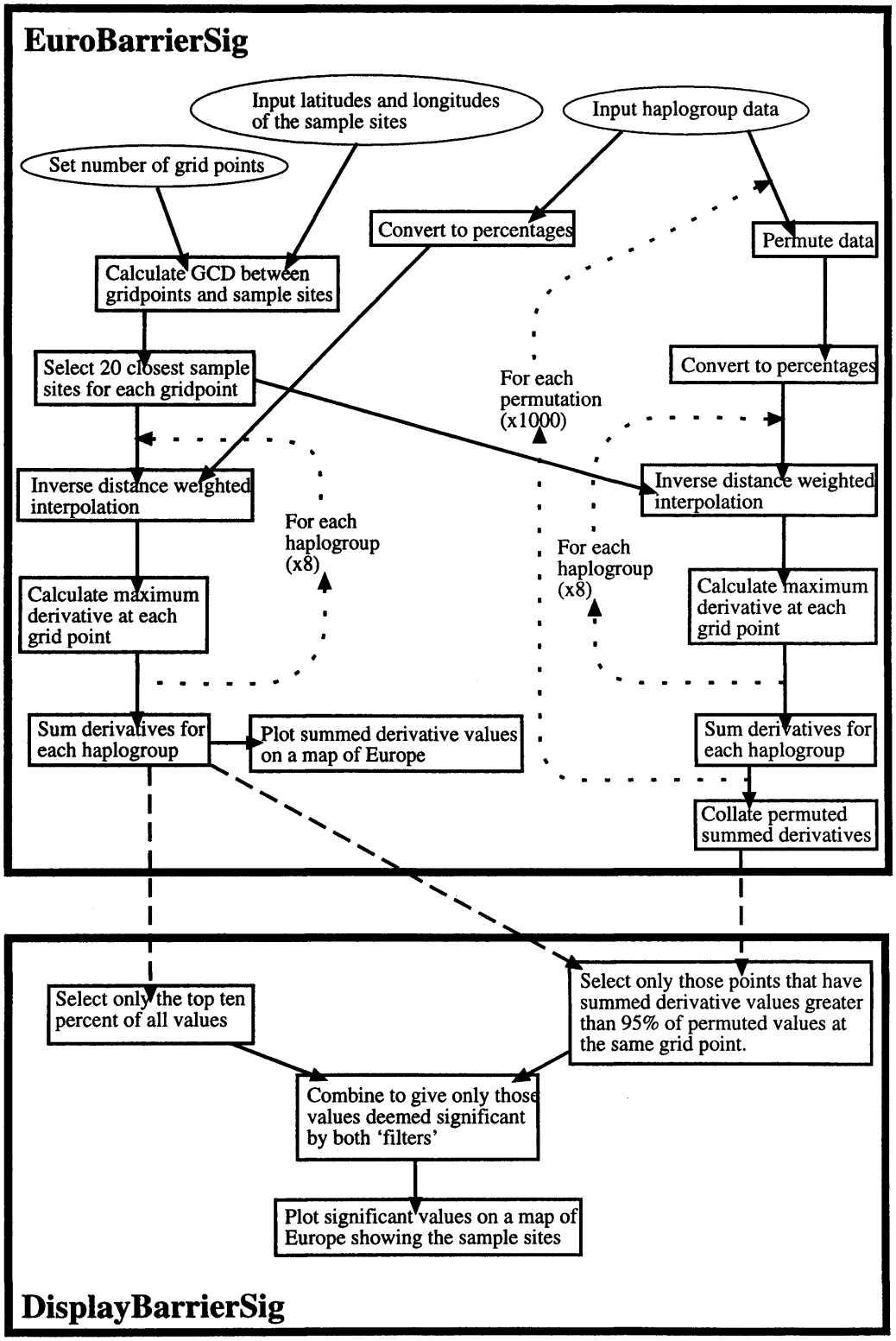


Figure 5.3: The modular design of the wombling programs.

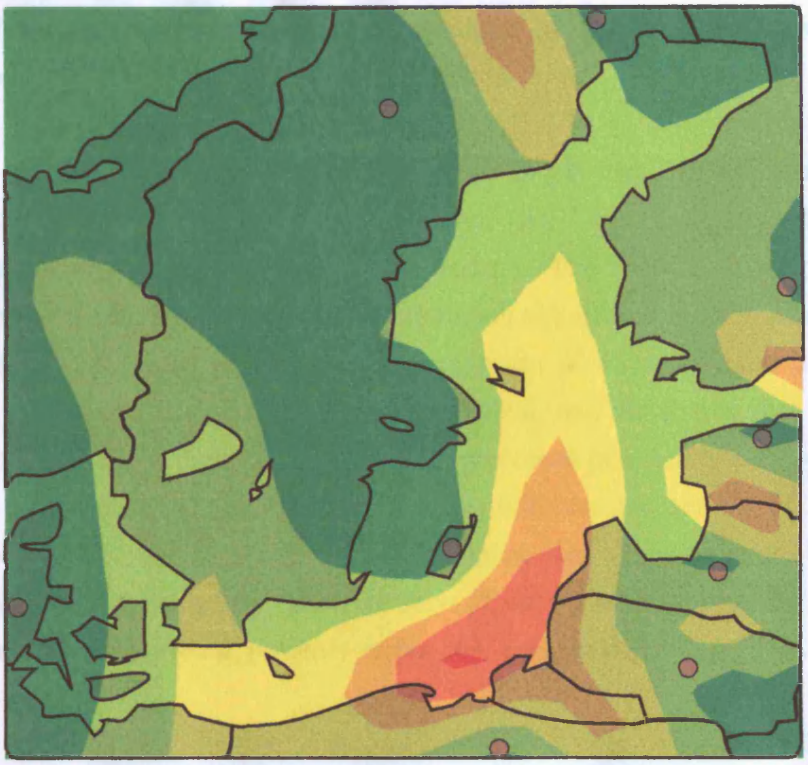
In essence, the EuroBarrierSig (EBS) and DisplayBarrierSig (DBS) programs use two significance 'filters' to exclude non-significant barriers. The first filter is the threshold value,

commonly used in the literature, of the top ten percent of all values in the landscape. The second filter represents 95% confidence limits for each individual grid point as assessed by a permutation test. Thus barriers are significant with respect to both the landscape in which they are found and to the null hypothesis of a random distribution of lineages.

This modular design gives flexibility in that once the replicates have been calculated in the EuroBarrierSig program, the DisplayBarrierSig program can draw upon them in such a way that the levels of significance of both significance filters can be varied to investigate the effect on the delineation of barriers.

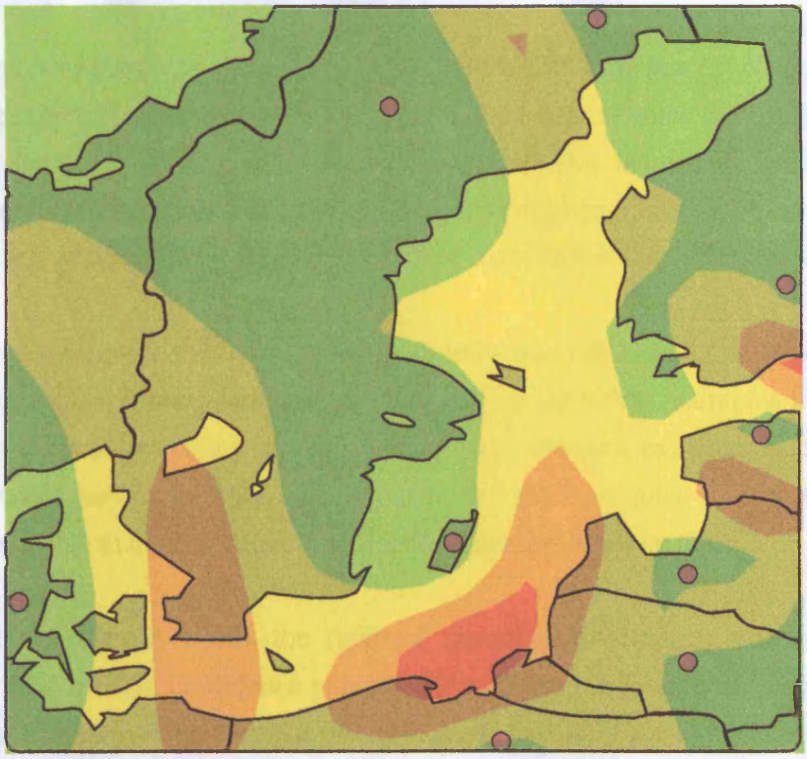
Lattice-wombling was chosen as the method for barrier detection. Although it does not take account of the known mutational distance between haplogroups I believe doing so could introduce error into the analysis for the following reason. By inspecting the congruence of linguistic boundaries with genetic barriers we are effectively limiting ourselves to the prehistory of the past few thousand years. Not only is this the deepest resolution attainable through linguistic comparison in this part of the world but given the known vast population movements over this continent in prehistory we can only reasonably expect barriers to reflect population history over this short time scale. The age of all the major lineages assayed here, from both coalescent and diversity analyses, is thought to be older than this time scale (Hammer et al. 1998). In addition, the geographical origin of some of these haplogroups is expected to be outside of Europe (Hammer et al. 1998). Taking account of the mutational distance between different lineages only makes sense when these lineages were founded within the temporal or spatial confines of the phenomena being studied. If this is not so then the lineage frequencies can be considered to be equivalent to allele frequencies and thus appropriate for the same analytical techniques.

Summarising the information from multiple surfaces can be achieved in one of two ways: either the derivatives of surfaces of each haplogroup can be summed, or the derivatives of surfaces of principal components calculated from the haplogroup frequencies can be summed. This latter approach is of interest for applications of this analysis to other types of data, for example in using microsatellite data. Figure 5.4 shows the results of a comparison of these two approaches when applied to a regional subset of the European data set around the Baltic. These two approaches were compared by using the Surface III interpoation software. Somewhat surprisingly there is a substantial difference between the two approaches, both in the numbers and size of barriers identified by the threshold criteria. This would seem to indicate that the latter approach, reliant on PCA, introduces substantial noise into the analysis perhaps in the form of differential weighting for



Summed derivative of interpolated maps of all 8 haplogroups

● =sample site



Summed derivative of interpolated maps of the first three principal components, representing 96% of the variance within the data.

Figure 5.4: Comparing the systemic functions of Wombling using Principal Components or Haplogroup frequencies. Successively more red contours represent higher barriers

barriers from different lineages. Consequently the method of summing derivatives of haplogroup surfaces was chosen for the lattice-wombling presented here.

Distance-weighting interpolation algorithms are much less computationally intense than alternatives. Consequently an inverse distance squared (IDS) weighting algorithm was used in order to minimise the size and running time of the program. The Surface III interpolation software was used to investigate whether the choice of interpolation algorithm makes a substantial difference to the barriers identified. Figure 5.5 shows the comparison of barriers found using Kriging and IDS algorithms, as defined by the 10% threshold criteria, and illustrates that though there are obvious differences, there is little fundamental disagreement between the two. In addition, as described in the Introduction to this chapter, Kriging involves the setting of certain parameter values which require the user to specify a model fitting the semi-variogram in question. The automated setting of these parameter values for each of the replications required in a permutation test would require many additional, possibly erroneous, assumptions.

The IDS algorithm was written to incorporate Great Circle Distances (GCD) to correct for the Earth's curvature. Other methods are available to achieve this, but are often computationally intense as they involve the mapping of a sphere onto a planar surface.

The sample site searching strategy was simply to identify the nearest N sample sites; N was chosen to be large (N=20) so as to control for any potential anisotropies that might result from the sample sites not being regularly distributed. Having said that the sample sites are reasonably regularly distributed throughout the landscape being investigated, certainly in comparison to other studies claiming to represent the entire continent (Barbujani and Sokal 1990; Sajantila et al. 1995).

The permutation algorithm chosen for this study was the Monte Carlo sampling algorithm of Roff and Bentzen (1995) that is also used in NCA. This algorithm maintains the sample sizes at each sample site. This is important as *a priori* the most obvious explanation of falsely positive genetic barriers is that they lie between closely positioned small samples such that sampling effects alone might account for sharp differences in allele frequencies between them.

It is well recognised within the field of permutation statistics that in order to make inferences at the 95% level of confidence at least 1000 replications are required (Efron 1982; Manly 1991). Consequently this was the number of replications used in the program.

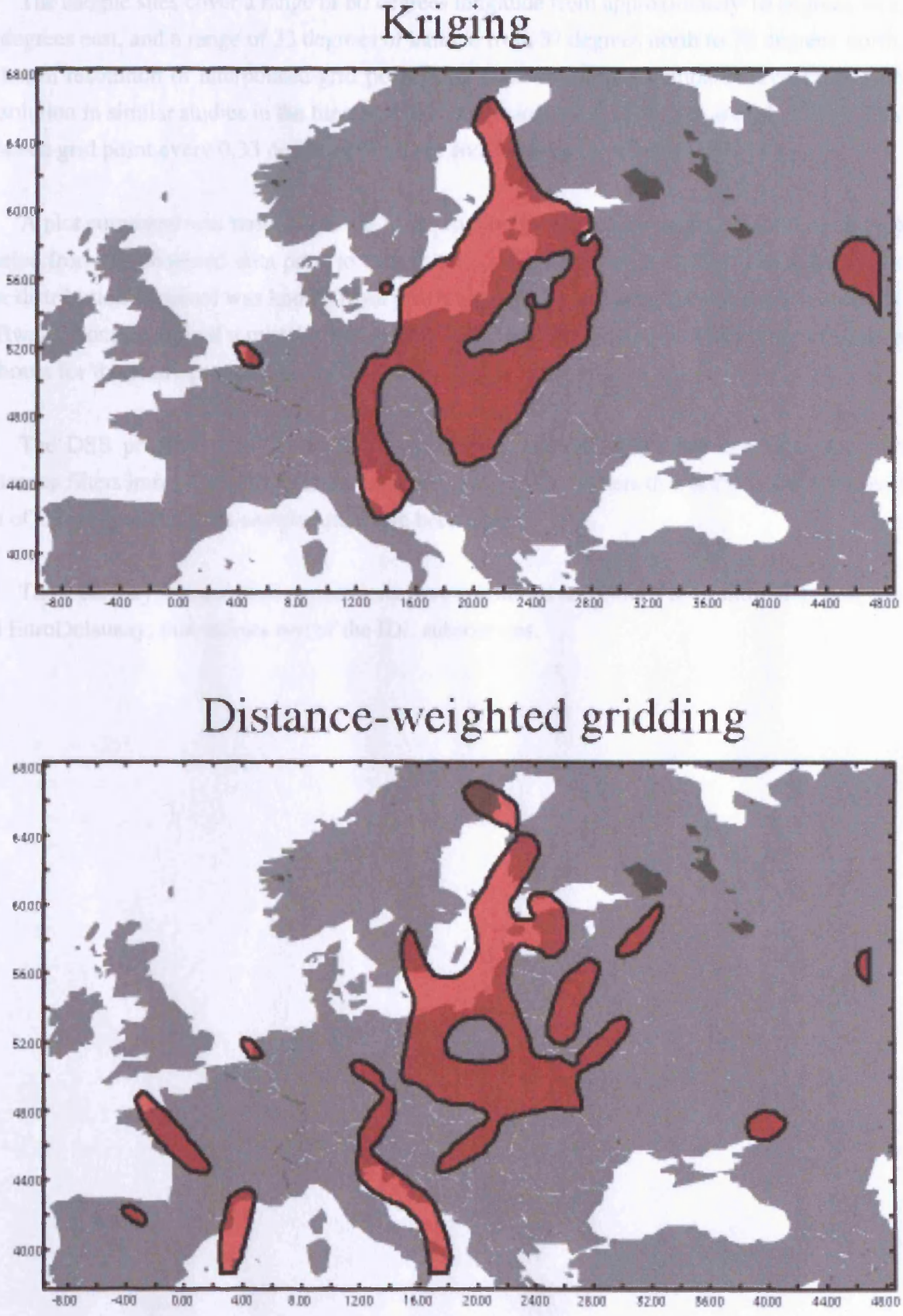


Figure 5.5: Comparison of barriers identified by Kriging and Distance-weighted gridding interpolations; the barriers are shown in transparent red and superimposed onto a map of Europe

The sample sites cover a range of 60 degrees longitude from approximately 10 degrees west to 50 degrees east, and a range of 33 degrees of latitude from 37 degrees north to 70 degrees north. The chosen resolution of interpolated grid points was 100x100, which compares favourably with the resolution in similar studies in the literature (Barbujani and Sokal 1990; Sokal et al. 1993). This represent a grid point every 0.33 degrees of latitude and 0.6 degrees of longitude.

A plot command was written into the EBS program so as to allow inspection of the summed derivative from the observed data prior to the initiation of the replication cycles. The approximate barrier distribution expected was known from previous work on the same dataset with the Surface III software. Consequently if a mistake was made it could be detected and rectified without waiting for 4 hours for the entire program to run its course.

The DSB program sets the level of significance for each filter and combines the two significance filters into a composite picture of significant genetic barriers that is then displayed over a map of Europe on which the sample sites have been plotted.

The Delaunay triangulation connections were calculated and plotted by a short IDL program called EuroDelaunay, that utilises one of the IDL subroutines.

Results of an analysis of European Y-chromosomal diversity

The sum of the derivatives of the surfaces from the distribution of 8 haplogroups in 43 European populations, prior to the application of any significance filters, was calculated by the EBS program and is displayed in figure 5.6. Figure 5.7 shows the same summed derivative when put through the threshold significance filter in the DBS program, thus only displaying the top ten percent of values. Figure 5.8 shows the result when solely the 95% significance filter from the permutation test is applied to the summed derivative of the observed dataset. The final picture of significant Y-chromosomal barriers in Europe is obtained by applying both significance filters to the data. This result is shown in figure 5.9.

It is clear that the threshold significance filter excludes a greater proportion of grid points than does the 95% permutation significance filter. Although there is great overlap between the grid points excluded by each filter they are by no means alternatives and are meant to operate in a complementary manner, as they are measuring significance in different ways.

The Northeast of Europe, specifically Poland and the Baltic, appears to be a region of high rate of change compared to the rest of Europe. The Polish sample was the smallest in the entire dataset and is notable for having by far the highest frequency of haplogroup 3 within the dataset. It is the dominance of this single haplogroup that accounts for the barriers between Poland and all other surrounding populations. Though the permutation procedure should correct for sampling effect, it is possible that another factor, perhaps population substructure has contributed to the anomalous nature of the Polish sample. Consequently the entire procedure was repeated for the same dataset with the Polish sample removed. The result of applying both significance filters to this new analysis is shown in figure 5.10.

It can be seen from a comparison of figure 5.9 and 5.10, that though the barriers look significantly difference in shape, they exist between the same populations and therefore removal of the Polish sample has little effect on the analysis. Given that this sample is the most influential of all it can be extrapolated that this analysis is relatively robust to the removal of individual populations.

The effect of varying the significance levels of the two filters was also tested. Figures 5.11 and 5.12 show the application of both significance filters to the barriers shown in figure 5.6, with one filter varied whilst the other is held constant. In figure 5.11 the permutation significance filter

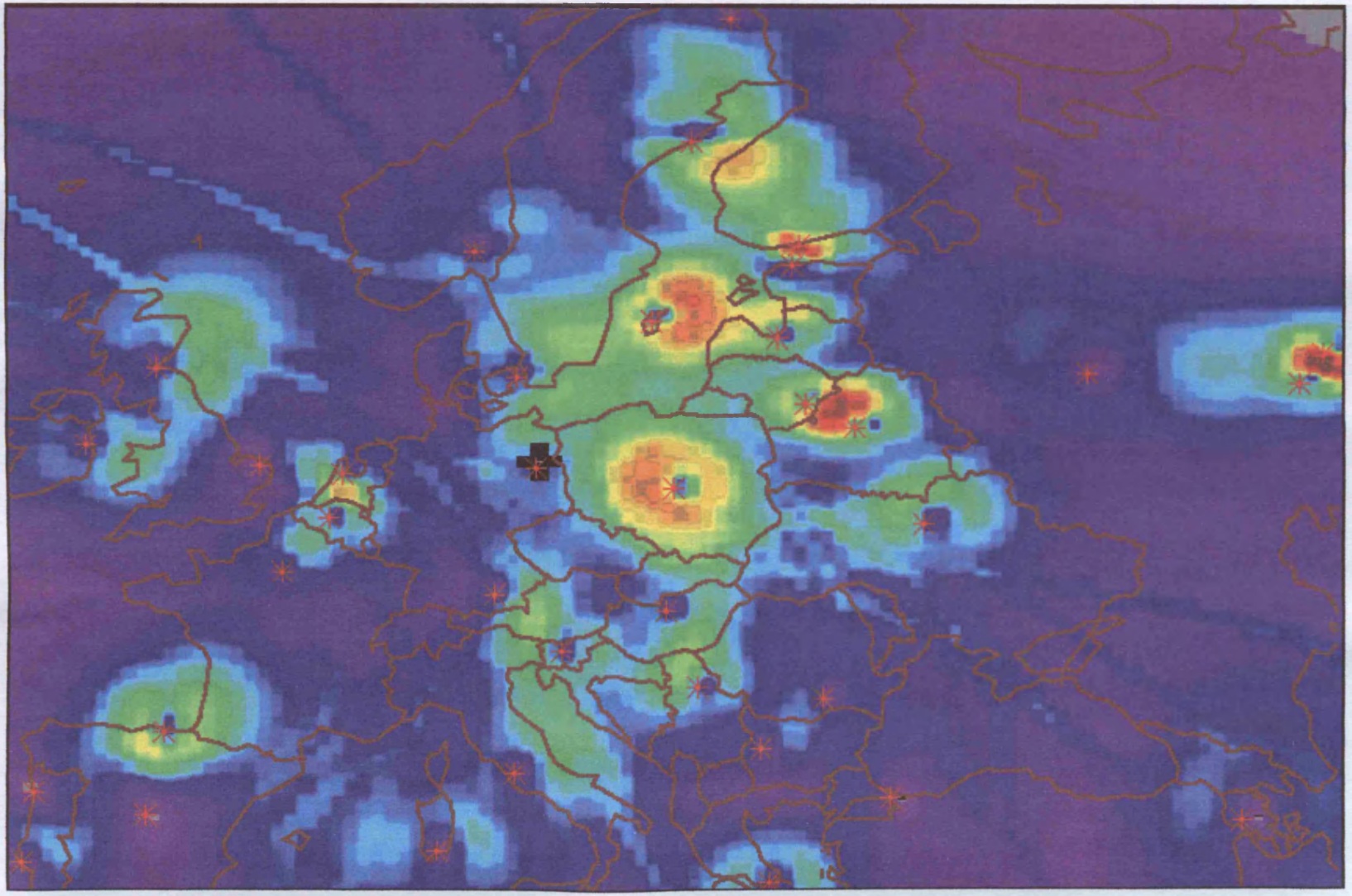


Figure 5.6: Lattice-wombling of the European Y chromosome diversity dataset superimposed on a map of Europe showing the position of the sample sites.

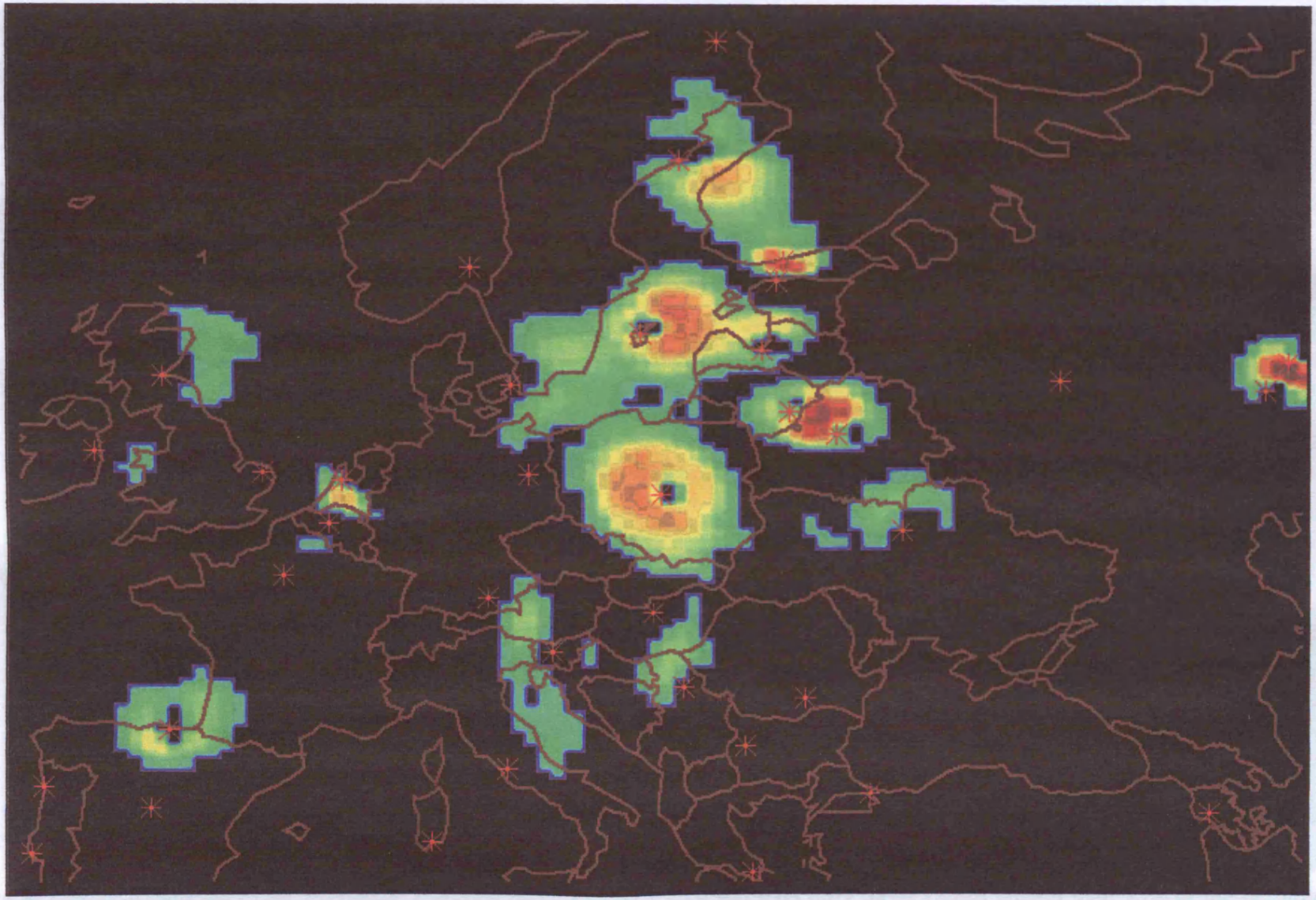


Figure 5.7: Lattice-wombling of the European Y chromosome diversity dataset, showing only the top ten percent of all values in the entire landscape

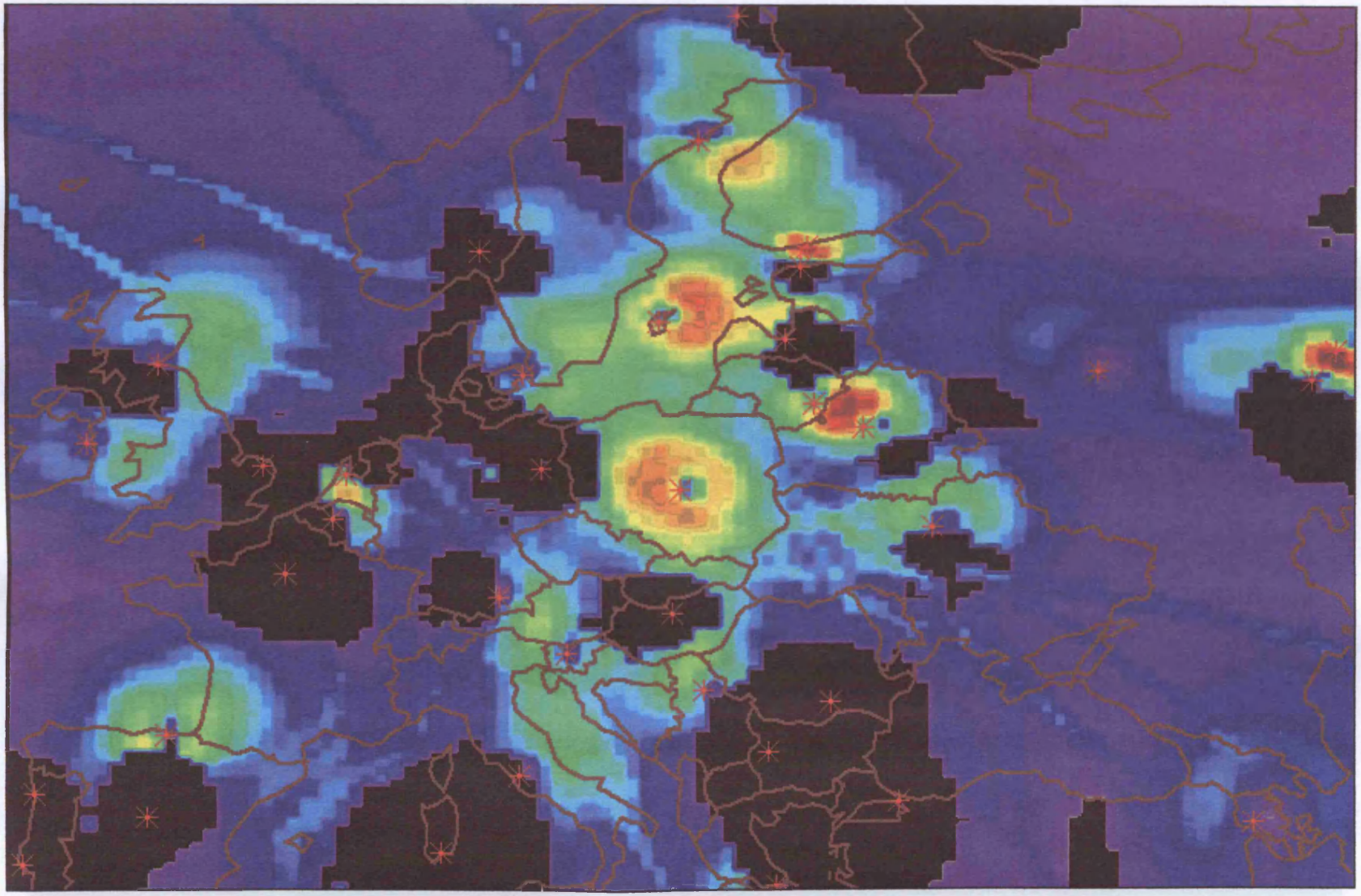


Figure 5.8: Lattice-wombling of the European Y chromosome diversity data set, with only those values greater than 95% of the permuted values shown.

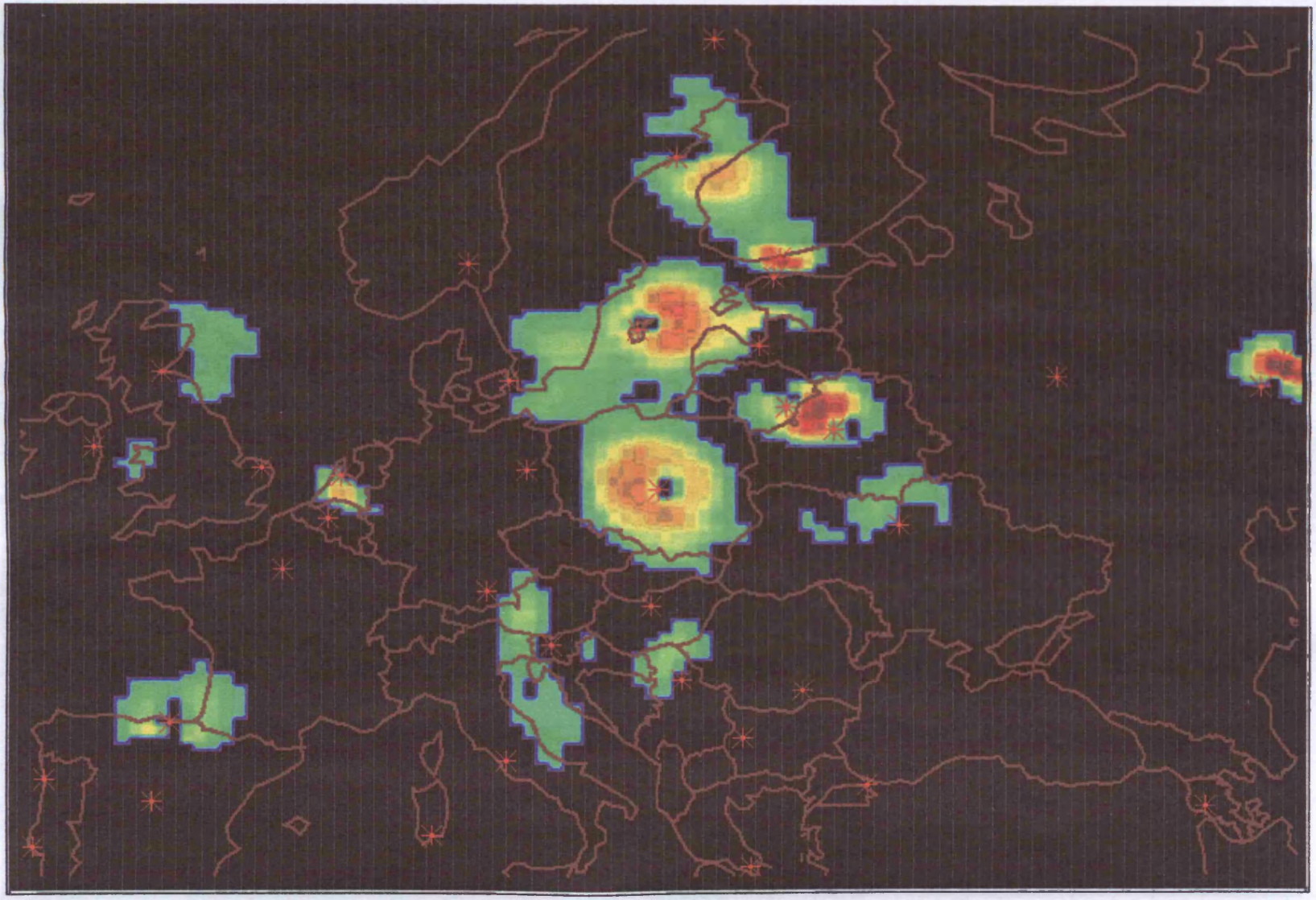


Figure 5.9: Lattice-wombling of the European Y chromosome diversity data set, showing only those values remaining after the application of both significance filters

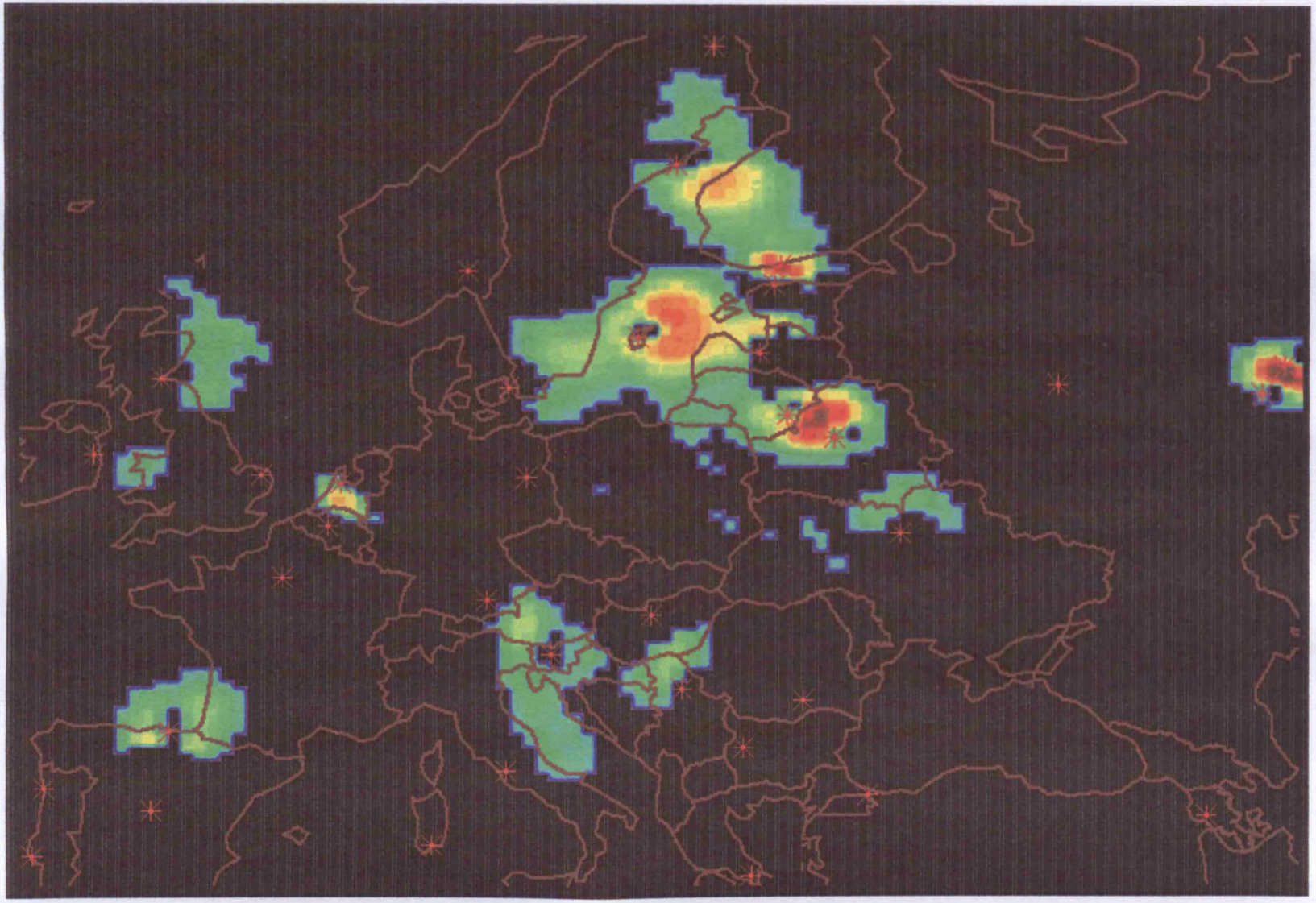


Figure 5.10: Lattice-wombling of the European Y chromosome diversity data set with the Polish sample removed, showing only those grid-points remaining after the application of the two significance filters

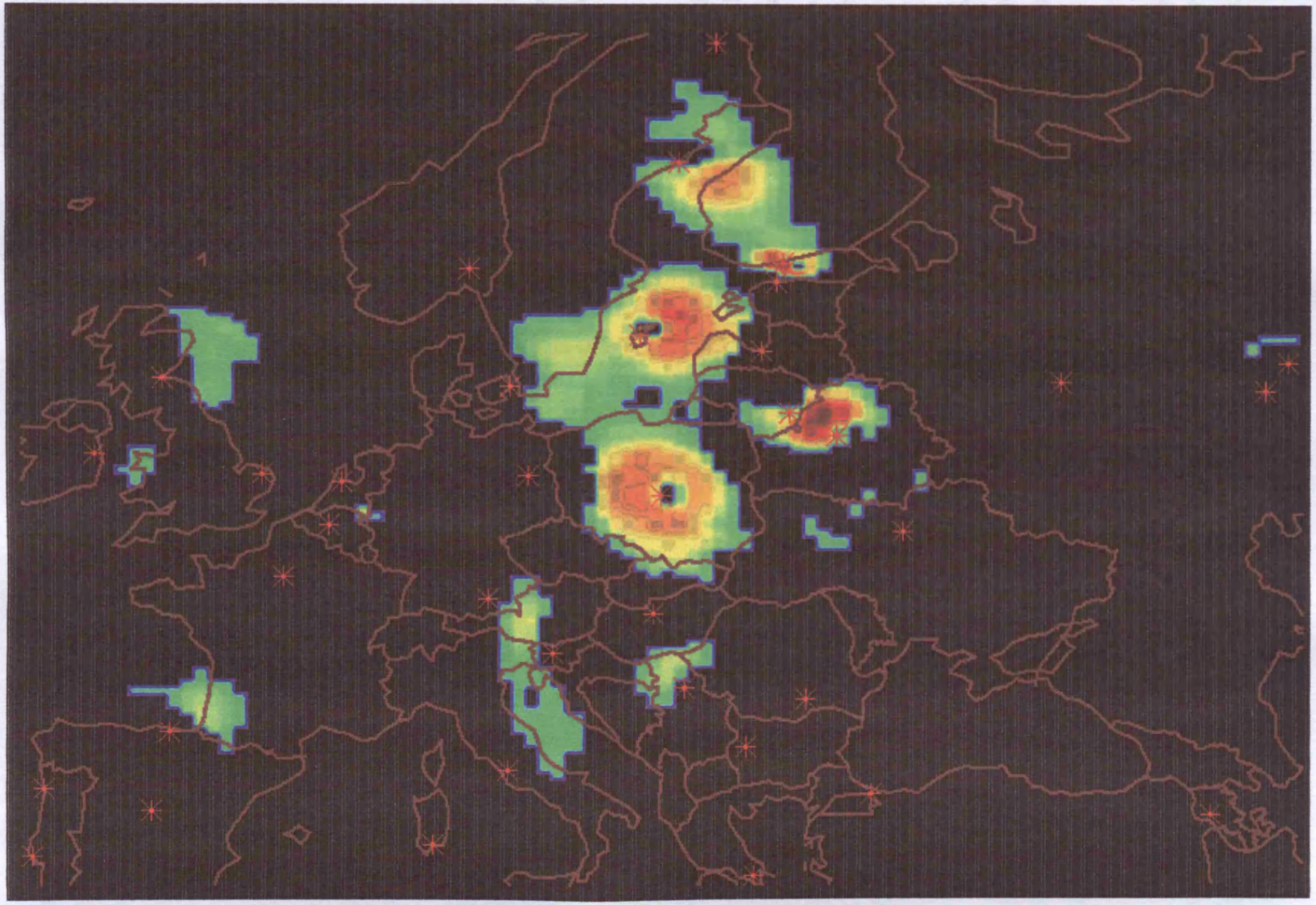


Figure 5.11: Lattice-wombling of the European Y chromosome data set, after application of a 99% permutation test significance filter and a top ten per cent threshold percentile significance filter.

has been raised from 95% to 99%, whereas figure 5.12 shows the effect of raising the threshold significance filter from the top ten percentile to the top five percentile.

1000 permutations are too few to be able to make reliable inferences at the 1% level of confidence, however figure 5.11 conveys the impression that many of the barriers are highly significant compared to the null hypothesis of a random distribution of haplogroups throughout Europe.

Figure 5.12 gives some idea of the robustness of the different barriers, in other words their level of significance compared to the surrounding landscape. It can been seen that the barriers in and around the British Isles and Italy disappear completely when the significance level is raised whereas the other barriers merely diminish in size.

Figure 5.13 shows the Y-chromosomal barriers mapped onto the map of European languages shown earlier in this chapter. At first sight it appears clear that genetic barriers do indeed coincide with language boundaries at the higher levels of classification: in other words at the level of the subfamily rather than the individual language within subfamilies. However some level of quantitation is required to support this inference. One of the ways of achieving this quantitation is described below.

Barbujani and Sokal (1990) were able to infer genetic barriers corresponding to linguistic boundaries at a high resolution, even inferring barriers between individual dialects within a language. This was as a result of the much finer sampling resolution of their dataset. With our dataset it would be disingenuous to attempt a similar level of resolution, simply because 37 languages are spoken amongst the 43 sample sites. Consequently barriers are likely, from purely a probabilistic perspective, to fall between sample sites speaking different languages.

However if linguistic differences do represent cultural barriers to gene flow we might expect that boundaries between higher levels of language classifications will present more of a barrier to gene flow that lower ones. This hypothesis can be tested quantitatively by the following analysis.

Unbiased connections can be drawn between the various sample sites by means of the Delaunay triangulation. The edges of each Delaunay triangle join two sample sites and can be classified into the level of language boundary that exists between the two sample sites. In this analysis three classes were chosen: within a subfamily (e.g. within the Romance subfamily of

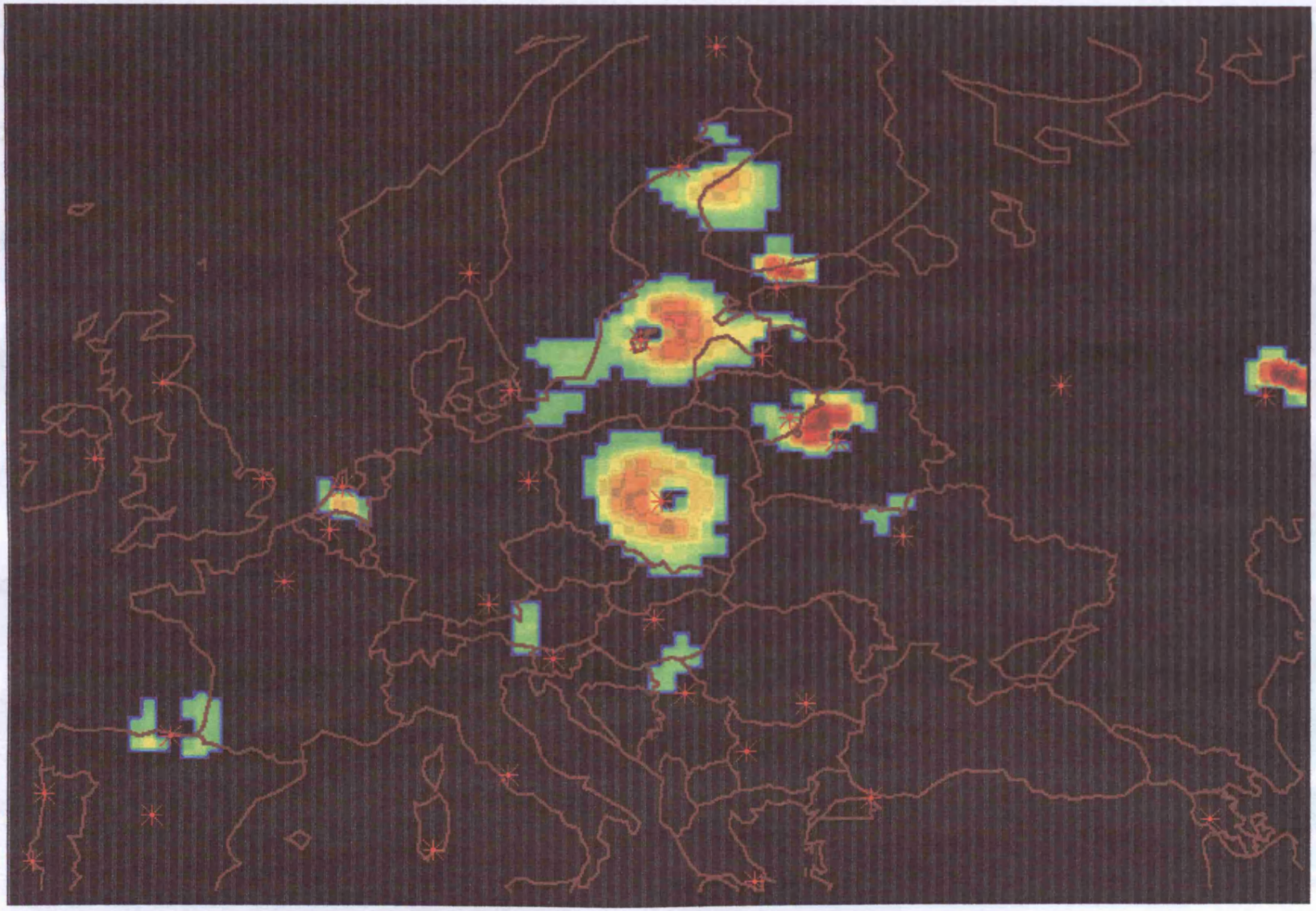


Figure 5.12: Lattice-wombling of the European Y chromosome data set, after application of a 95% permutation test significance filter and a top five per cent threshold percentile significance filter.

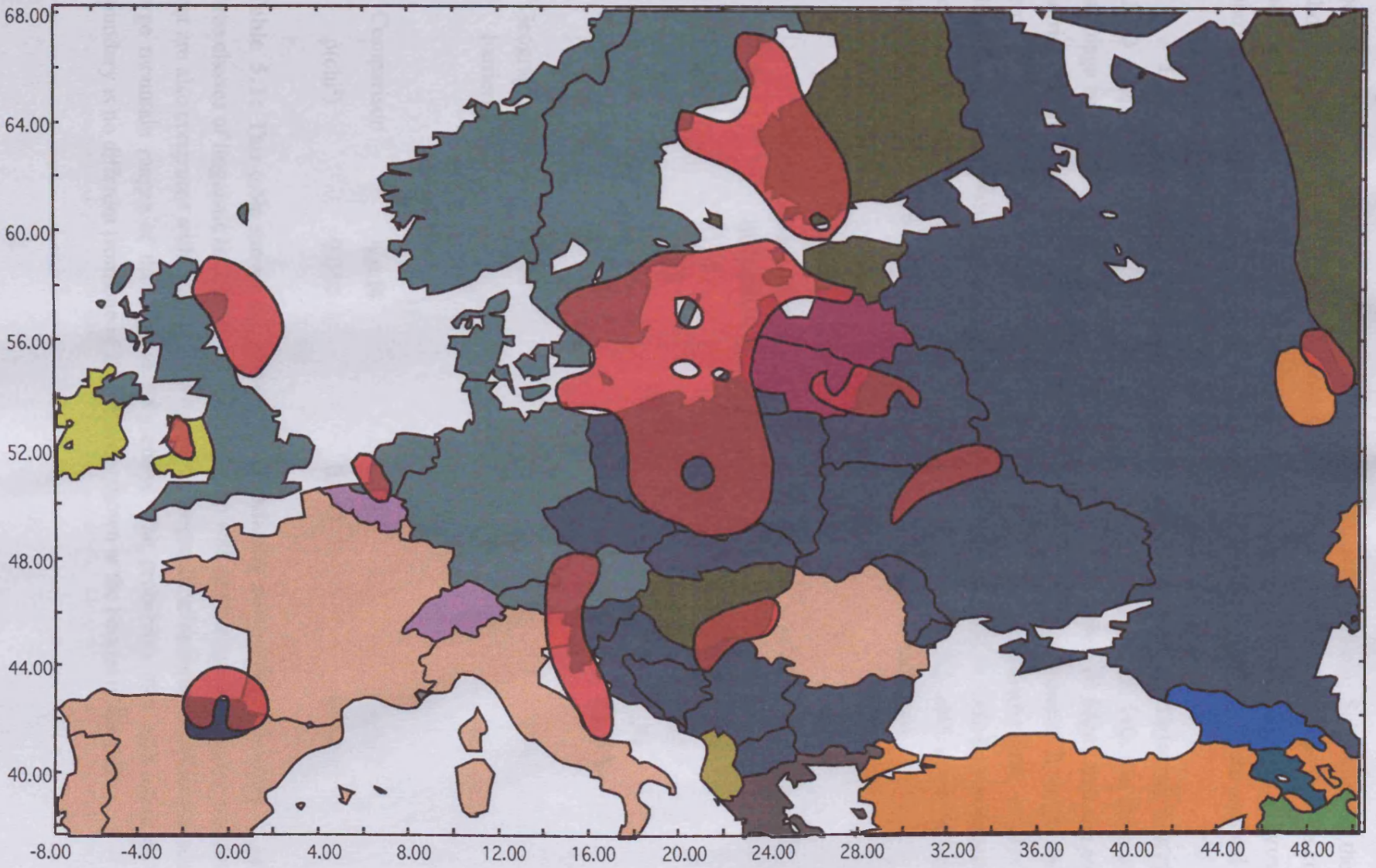


Figure 5.13: Barriers identified by lattice-wombling of the European Y chromosome diversity dataset after applying the two standard significance filters, superimposed on a map of the language diversity of Europe. The barriers are shown in transparent red.

Indo-European), between subfamilies (e.g. between Romance and Germanic subfamilies) and between families (e.g. between Indo-European and Uralic families). Subsequently the barriers shown in figure 5.13 can be mapped onto the Delaunay connections and those that are crossed by barriers noted. Finally the proportion of each class of Delaunay connection that is crossed by a barrier can be calculated and the significance between the classes tested by the chi-squared statistic.

Figure 5.14 shows the Y-chromosomal barriers mapped onto the Delaunay triangulation of all 43 sample sites. There are 99 different triangle edges, which fall into the three classes of language boundary in the ratio 31:37:31 respectively. The number of edges that are crossed by barriers is summarised in Table 5.1, together with their significance. It might occur that geographical barriers are also correlated with levels of language classification, in which case a discrepancy in the number of barriers found crossing each class of Delaunay connection might result purely from the indirect effect of geography. Thus table 5.1 also notes the number of linguistic boundaries that correspond with geographical barriers in each class.

	A	**B**	**C**
	Connection within subfamilies	Connection between subfamilies	Connection between families
Barriers	8	16	19
No barriers	23	21	12
Total	31	37	31
% barriers	**25.8%**	**43.2%**	**61.3%**
Geographical barriers	6/8	8/16	9/19
Comparison	**A=B**	**B=C**	**A=C**
$p(chi^2)$	0.050	0.043	6.3×10^{-6}

Table 5.1: This table shows the proportion of Delaunay connections belonging to each of the three classes of linguistic boundary that are congruent with genetic barriers and the number of these that are also congruent with geographical barriers. Geographical barriers are defined as seas and the large mountain ranges of the Alps and Pyrenees. The probability that each class of linguistic boundary is no different from one of the others is shown at the bottom of the table.

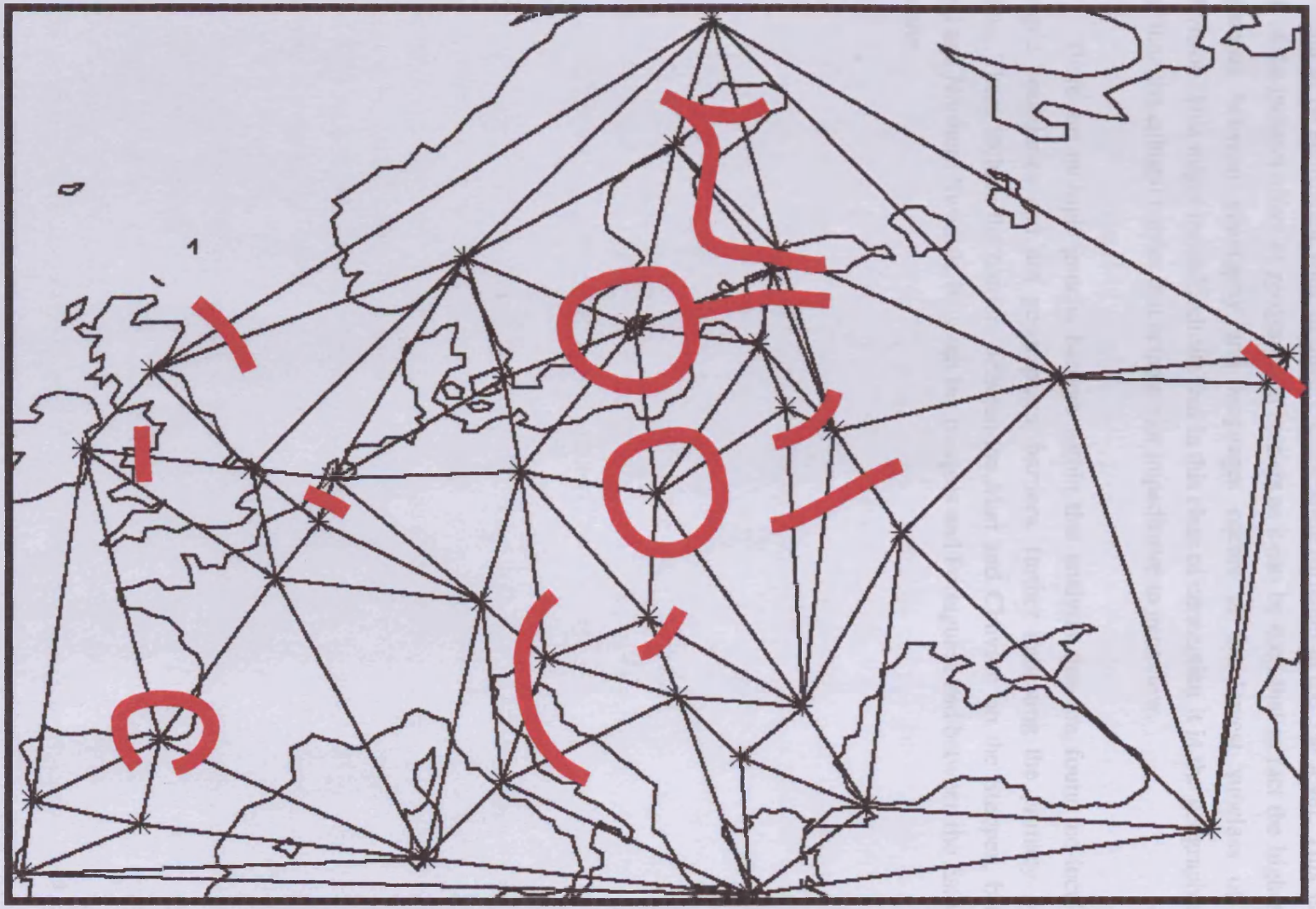


Figure 5.14: Map of Europe showing the Delaunay triangles connecting all 43 sample sites and a schematic representation of the significant barriers identified by the EBS program shown in red

It can be clearly seen that the hypothesis is borne out by the analysis, with successively higher levels of language classification representing larger barriers to gene flow. These differences are found to be significant with all three classes being different at the 5% level. This is not as a result of the indirect effect of geographical barriers as it can be seen that in fact the highest level of congruence between geography and languages occurs at the lowest subclass of language subdivision. This might indeed indicate that in this class of connection it is the geographical barrier rather than the cultural barrier that is the major impediment to gene flow.

There are multiple genetic barriers within this analysis that are found co-localised with linguistic boundaries and not geographical barriers, further indicating the primacy of cultural barriers. These include the barriers between the Mari and Chuvash on the Steppes, between the Saami and Northern Swedish, between the Basques and Portuguese and between the Estonians and Latvians.

Discussion

This analysis attempted to use a comprehensively sampled single locus study of European diversity to explore the relationship between genetic barriers and language barriers. It has been shown that the significant genetic barriers within Europe are spatially congruent with language barriers. Clearly this indicates that cultural differences, specifically linguistic differences, can and do represent significant barriers to male gene flow. It has also been shown that this analysis is likely to be robust to the removal of individual samples within the region. Designing a bootstrapping procedure analogous to that used in phylogeny construction is not straight-forward and is likely to be highly computationally intense.

The program EuroBarrierSig, in using two significance filters, a permutation test as well as a threshold percentile, provides a new way of assessing the significance of barriers identified by lattice-wombling. Although in this case the permutation significance filter did not exclude any major barriers identified by the threshold percentile approach to significance, a number of individual grid points were excluded, altering the shape of the barriers. It is reasonable to expect that the Y chromosome, through being the most geographically differentiated single human locus, would be the locus least affected by such a permutation test. The application of this program to other datasets from other loci would be useful to test this hypothesis, and would hopefully confirm the validity of using a permutation test in conjunction with a threshold percentile.

Any grid points excluded by the permutation significance filter represent a significant deviation from the null hypothesis of random haplogroup distribution throughout Europe. However not all remaining grid points represent genetic barriers; the clinal pattern expected under the model of isolation by distance will also generate significantly non-zero derivative values. Thus barriers must always be defined in the context of their landscape and so a threshold percentile significance filter must always be included.

The dependence of barrier delineation on the surrounding landscape must not be forgotten when interpreting genetic barriers. This is clearly illustrated in figure 5.15 where the samples surrounding the Baltic are considered separately from the rest of Europe. This analysis was performed using the Surface III software and thus only includes the threshold percentile significance filter. Comparing this figure with those shown in the Results section clearly indicates

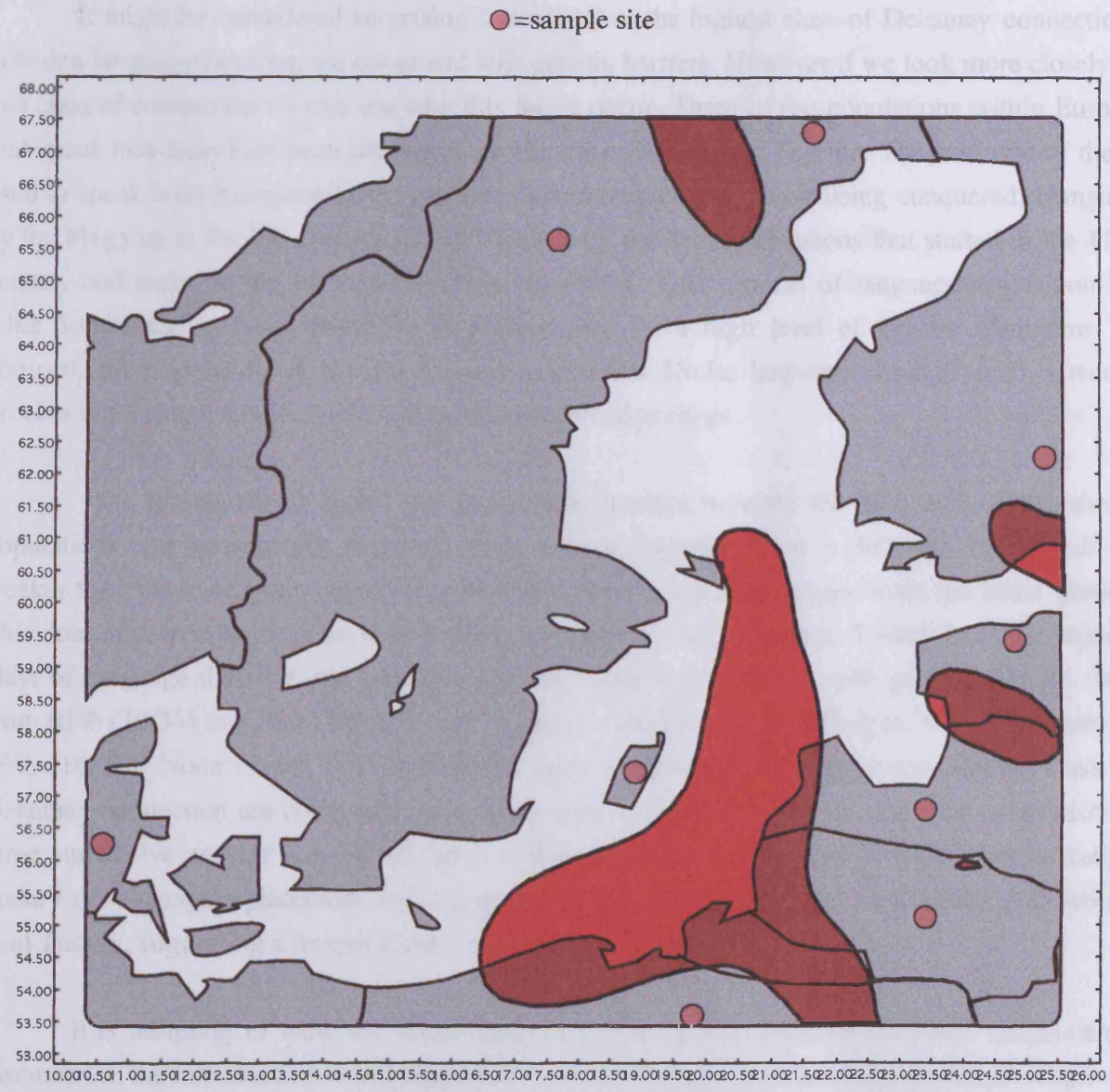


Figure 5.15: Lattice wombling of Y chromosome diversity in samples around the Baltic, Barriers are shown in transparent red and are superimposed on a map of the region.

that some of the barriers within this region that are significant in a European wide context are not significant when considered at the regional level.

It might be considered surprising that not all of the highest class of Delaunay connection, between language families, are congruent with genetic barriers. However if we look more closely at this class of connection we can see why this might occur. Three of the populations within Europe that speak non-Indo-European languages are Hungary, Turkey and Estonia. The first two of these used to speak Indo-European languages but adopted new languages on being conquered; Hungary by the Magyars in the 9th century AD and Turkey by the Turkic invasions that started in the 11th century and ended in the 15th century (Renfrew 1989). This process of language acquisition by 'elite dominance' is not expected to be accompanied by a high level of genetic admixture. In contrast, the population of Estonia has not gained it's Uralic language through such a recent process but through a more ancient population historical heritage.

It is reasonable to expect that any genetic barriers between the first two of the above populations and surrounding populations speaking a language from a different family will be weaker than those between populations that have never spoken languages from the same family. This does appear to be the case. If we exclude all Delaunay connections to Turkey from the highest class of language division, the proportion of this class that coincides with genetic barriers rises from 61% (19/31) to 73% (19/26), if we also reject connections from Hungary we get a figure of 76% (16/21). None of the five connections around Turkey belonging to the highest class of Delaunay connection are congruent with genetic barriers, whereas for Hungary the proportion is three out of five and for Estonia the figure is five out of five. Note that Hungary has an earlier history of language replacement and a greater proportion of barriers to surrounding populations than Turkey, suggesting a temporal component to barrier formation.

It is tempting to view the above analysis in the light of the two proposed mechanisms, discussed in the introduction to this chapter, by which language boundaries can be found congruent with genetic barriers. These can be summarised as (i) pre-existing genetic and linguistic differences through different population histories and (ii) different languages allowing genetic differences to accumulate by relative reproductive isolation. The above examples of language replacement by 'elite dominance' indicate that the second mechanism does indeed operate, but the greater proportion of this type of barrier in populations that have retained their ancestral language (e.g. Estonian) than those that have had it imposed (e.g. Hungary) suggests that divergent population histories underpin

the strongest genetic barriers. Although undoubtedly in such cases both mechanisms can operate synergistically.

This analysis contrasts markedly with a recent attempt to correlate maternal lineage diversity in Europe with languages (Sajantila et al. 1995). That study failed to find any correlation between mtDNA and linguistic boundaries. It is difficult to fully compare these studies because a barrier analysis was not attempted in the mtDNA study and the sampling was less than comprehensive with only 14 very irregularly distributed populations sampled. However the recent compilation of a higher resolution mtDNA study (Guido Barbujani, personal communication) should allow a more direct comparison to test the hypothesis mentioned earlier in the introduction to this chapter that gene flow across language barriers is higher for females than males.

In principle it would be possible to compare the barriers identified here with those identified from the autosomal study mentioned earlier in this chapter (Barbujani and Sokal 1990). This would be worth attempting if the presence of a barrier in one study but not in the same region in the other study could be deemed indicative of some underlying population historical cause. Two reasons suggest that this is not the case: firstly the different sampling in the two studies and secondly the relativistic nature of barrier definition (putting aside the fact that the barriers were defined differently in the two studies). For example, the presence of a barrier between the Chuvash and Mari tribes found in this study could not be replicated in the autosomal study by virtue its not having sampled these populations. Furthermore by not defining this barrier the autosomal study has a 'spare quotient of significant grid-points' which it can assign to the next highest barrier not previously defined as being significant. All other things being equal, this barrier would not appear to be significant in the Y-chromosomal study having already exhausted it's 'quotient of significant grid-points.' Ideally, any direct comparison of barriers produced from the analysis of different loci would only be attempted where the different loci were typed in the same samples.

In summary, this barrier analysis of European Y-chromosomal diversity indicates that the broad picture of clinal variation in Europe is fine-tuned by the regional heterogeneities within the landscape, and that these heterogeneities are both cultural and geographic in origin.

Chapter 6: Phylogeography of the Y chromosome in Southeast Asia and the Pacific

Introduction

The Polynesians were skilful ocean-going navigators at a time when the Greeks and Romans were little more than coastal sailors. Prior to 1500 AD they constituted the most geographically widespread people on Earth (Bellwood 1989) and the remarkable settlement of the remote Pacific islands has been extensively studied from the viewpoint of the archaeological record and cultural and linguistic affiliations (Jennings 1979). The Pacific islands have been traditionally classified into three geographical areas. Melanesia includes Papua New Guinea and islands to the East including the Solomon Islands and Vanuatu. To the North of Papua New Guinea, the dispersed archipelagos of Micronesia stretch from Palau in the West to Kiribati in the East. The islands to the East of Fiji make up Polynesia, a vast ocean triangle with sides roughly 6,500 km in length and apices at Hawaii, Aotearoa (New Zealand) and Rapanui (Easter Island). These cultural and geographical affiliations are shown on a map in figure 6.1. In addition the populations of a number of islands in Melanesia and Micronesia are thought to be more closely related to Polynesians than they are to their geographical neighbours. Consequently these islands are known as 'Polynesian outliers'.

The islands of the Pacific have received much attention from many different disciplines. Such geographically-separated islands are assumed to be isolated and as such to represent microcosms of biogeographic history. In particular those interested in evolutionary processes of all types often adhere to the 'islands as laboratories' axiom, representing the presumed multiple opportunities to observe independent evolution (Kirch and Green 1992). This approach can be traced back to Darwin's studies of finches from the Galapagos Islands.

The primary focus of this chapter is anthropology and not analytical methodology, and consequently it is necessary to introduce the background of this region in much greater anthropological depth than has previously been attempted in this thesis. Whilst I focus on issues surrounding the origins of the Polynesians, the geographical field of view extends, by necessity, into Island Southeast Asia. Firstly a broad summary of the major population movements that have dominated this region is presented, followed by a brief introduction to the major hypotheses for

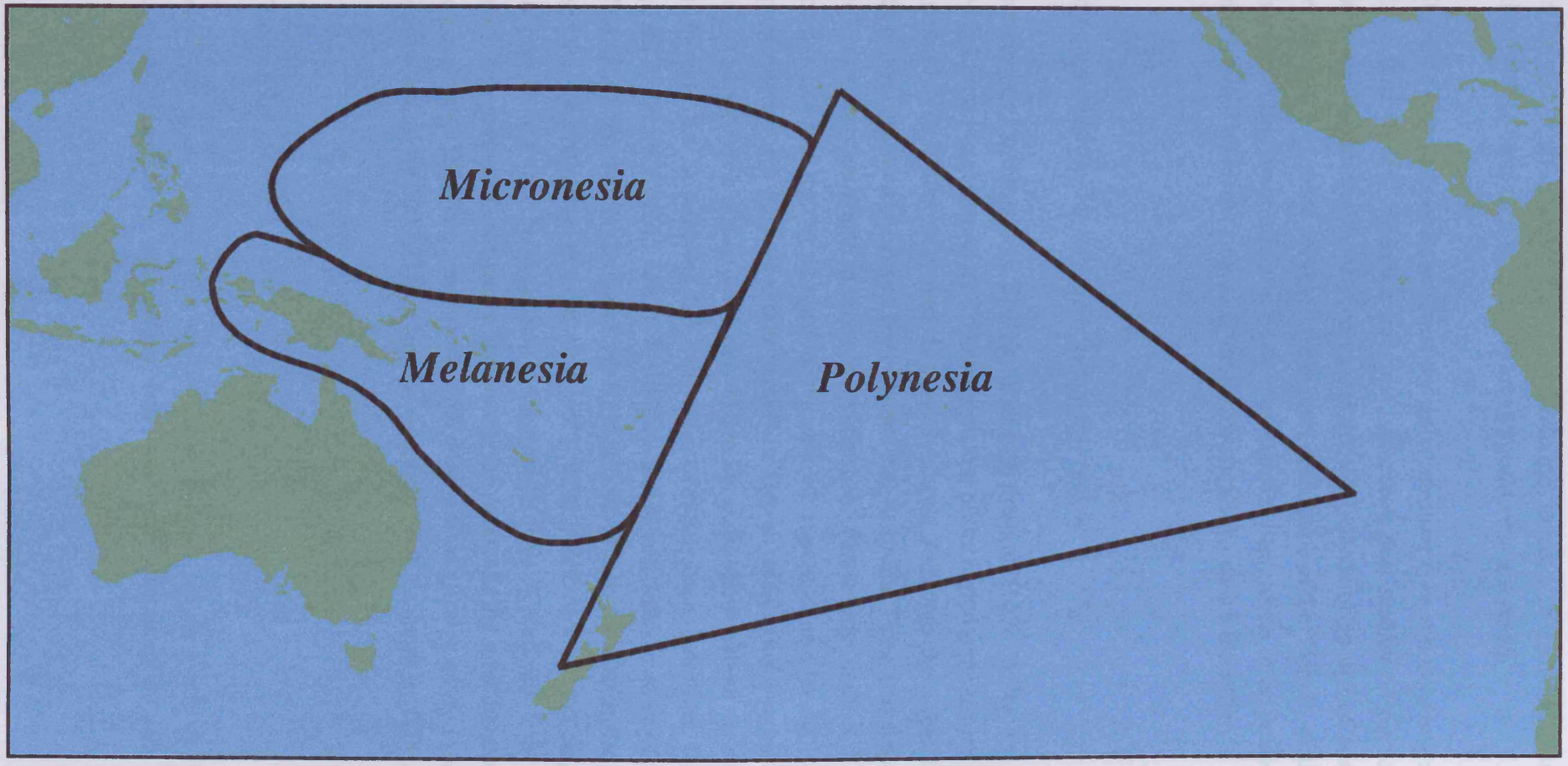


Figure 6.1 - Map of Southeast Asia and the Pacific indicating the three major cultural divisions in the Oceanic islands

Polynesian origins. Subsequently the finer resolution attained in individual studies from the different disciplines of archaeology, linguistics and genetics is discussed.

Whilst the complementary nature of the different disciplines can only be emphasised, there often exists a fundamental difference in scope. Archaeological studies, in particular, often limit themselves to studies of specific time periods in specific islands or groups of islands. In contrast most molecular genetic studies attempt a large-scale description of the diversity of the entire region. Consequently marrying these different disciplines together has proven to be problematic and most attention has focused on identifying compatible hypotheses from each discipline.

A consensus background to human migrations in this region

Archaeology, anthropology and linguistics (Bellwood 1987; Bellwood 1991; Bellwood 1989) have been used to construct the following picture of the region's prehistory. 30-50,000 years ago, Australoid hunter-gatherers moved through Southeast Asia into the Sahul landmass, which comprised the present islands of Papua New Guinea and Australia. These peoples are thought to be ancestral to the extant populations of native Australians and Papuan highlanders. The second major prehistoric migration into the region is associated with the expansion of a Mongoloid Austronesian-speaking population which began 5-6,000 yrs BP in Taiwan and coastal Southeast China. Over a relatively short period, their highly developed navigation skills allowed them to settle the islands of Melanesia, beyond the Solomons, that had not previously been colonised by the Australoid population. By 3,000 yrs BP, Fiji in East Melanesia and Samoa and Tonga in West Polynesia had been reached. This initial colonisation of island Melanesia is associated with the 'Lapita' culture, characterised by its distinctive pottery, and was quickly followed by the migration of other Melanesian populations that had acquired ocean-going skills. Colonisation of remote Oceania followed a lengthy period of adaptation in Western Polynesia during which a distinct Polynesian culture had developed. The Marquesas were reached by 300 AD, followed by Hawaii by 500 AD, Rapanui (Easter Island) by 800 AD and, finally, Aotearoa (New Zealand) by 1200 AD.

Models of Polynesian origins

Linking the proto-Polynesians to the Lapita culture, whose sites are found in island Melanesia, suggests relatively little genetic contribution of Australoid peoples to Polynesians.

Rather it was thought that the Polynesians had their origins in the relatively rapid maritime expansion from Southeast Asia (Bellwood 1979). An alternative model postulates that the proto-Polynesians evolved within Melanesia from the resident population that had been there for at least 30,000 years (Terrell 1986; Terrell 1997). A third model proposes a substantial colonisation of Polynesia from the Americas (Heyerdahl 1950). In addition there has been substantial contact with Europeans over the past 300 years, who may have contributed genetically to the extant population.

Linguistic analyses

Polynesian languages are closely related to each other, and belong to the Oceanic subgroup of the Austronesian language family (Pawley and Ross 1993). A summary of linguistic relations in these linguistic groups is given in figure 6.2. These subgroupings are by no means without contention, though they represent the closest thing to a consensus currently available (Kirch and Green 1992; Pawley and Ross 1993). It was the bringing of word lists from Indonesia, Madagascar and Polynesia to Europe in 1708 that first revealed the existence of a language family. Subsequently 1000-1200 languages belonging to this family have been identified, making it the largest well-established language family (Pawley and Ross 1993).

The use of the comparative method for the construction of proto-languages for groups of related languages, for example proto-Austronesian, proto-Oceanic and proto-Polynesian almost inevitably has led to an attempt to identify geographically an origin for the related dispersals (Pawley and Ross 1993). The reconstructed vocabularies of proto-languages inform on the cultures in which these languages were spoken and to the changes that accompanied the migration routes; an example is given below. Malagasy is also an Austronesian language, related most closely to languages spoken in Borneo, which has lead to a hypothesis of Indonesian colonisation of Madagascar (Grimes 1996). However, all words for the husbandry of large animals in the different Malagasy dialects come from Swahili. Transportation of domestic animals obviously did not survive the canoe trips that lead to Madagascan colonisation, and correspondingly all large animals farmed in Madagascar are of African origin. In Southeast Asia the words for root crops such as the yam and taro are only found in proto-Malayo-Polynesian and not proto-Austronesian indicating a shift in agriculture accompanying the latitudinal movement into the tropics from a more Northerly homeland (Pawley and Ross 1993). In contrast the words for grain crops present in proto-Austronesian become lost before the emergence of proto-Oceanic. Prior to historical European and Indian influences Austronesian languages were spoken, and not written ones. Therefore there does

not exist a written record of extinct languages, evidence that has aided the reconstruction of linguistic relationships in the Indo-European language family (Renfrew 1989).

Austronesian languages are spoken in a continuum throughout Island Southeast Asia, into Island Melanesia and Micronesia, and out into the remote Pacific Islands. Austronesian languages are not spoken exclusively in these regions - another group of languages is spoken, mainly in Melanesia. Dubbed 'Papuan', these highly diverse languages are classified together more through not being Austronesian than through shared characteristics (Pawley and Ross 1993). In Melanesia Austronesian languages are largely restricted to coastal regions and islands. There is no Austronesian penetration into central Papua New Guinea, the highlands of which exhibit massive numbers of diverse Papuan languages. There is considerable evidence that the Austronesian languages of Western Melanesia have borrowed substantial number of terms and grammatical features from their Papuan neighbours (Pawley and Ross 1993).

There is good agreement that the greatest diversity within Austronesian languages is apparent in Taiwan and that consequently this is the most likely homeland for these peoples (Bellwood 1991; Pawley and Ross 1993). It has been suggested that the wide distribution of Austronesian languages was fuelled by the expansions of a Neolithic culture out of Southeast China and Taiwan roughly 6000 years ago. It is largely the phylogenetic arrangement of Austronesian languages that has led to the hypothesis of a rapid expansion of a relatively homogenous people through Melanesian and into Polynesia. The finding of linguistic sublineages exhibiting wider geographic distributions than their ancestral lineages is closely analogous to the definition of evidence for a range expansion as argued by (Cann et al. 1987) and (Templeton 1998).

Archaeological analyses

The achievements of archaeology in reconstructing Oceanic prehistory have been overshadowed to a certain degree by the successes of historical linguistics. Widespread archaeological evidence of substantial human occupation prior to the Austronesian expansions can be found in the Bismarck archipelago, in island Melanesia, just off the coast of Papua New Guinea, but no further into the Pacific. The Lapita culture is characterised primarily by pottery with a distinctive dentate patterning but includes a characteristic tool kit, earth ovens and shells, and is found spread from Papua New Guinea to Tonga and Samoa (Bellwood 1987; Kirch and Green 1992). Sites found to contain Lapita artefacts are dated to a short time window from 3,600-3,000

years BP. This distribution of sites together with an apparent older diversity of pottery in the Western edge has led to the idea that this represent a rapid spread of a people, probably speaking an Austronesian tongue from a homeland in West Melanesia (Bellwood 1987). It is likely that these peoples had the use of the double-hulled canoes that were to prove invaluable for Oceanic exploration. In addition they are likely to have adopted the tropical agricultural crops of taro and yam. The widespread dispersal of obsidian, used extensively for tool-making, from its source in New Britain is further evidence that there were extensive trading networks throughout this region prior to expansion into Polynesia (Bellwood 1987).

Whilst islands falling into the Polynesian grouping were first colonised some 3000 years ago the first archaeological evidence of spread from Samoa and Tonga into central Polynesia is only found in the Marquesas some 1000 years later (Bellwood 1987). In this intervening period the obsidian- and pottery-based Lapita culture developed into the distinctive Polynesian culture that relied on neither of these innovations, thus stripping archaeologists of their most useful evidence for tracing prehistoric movements. It is only recently that the two-way voyaging between Polynesian islands has been detected archaeologically through the application of a new technique of isotope analysis allowing alternative sources of the basalt used for most tool-making to be discriminated (Weisler and Kirch 1996; Weisler and Woodhead 1995). This study has revealed that voyaging occurred not only between islands within the same archipelago but also between islands of different archipelagos.

The dates usually offered for the first evidence of human occupation of the Eastern Polynesian islands detailed above have recently come under scrutiny (Spriggs and Anderson 1993). After a number of studies proposing successively earlier dates than those given above, a controversial paper offered a more conservative assessment of all the available samples and argued that the previously determined dates for the first colonisation of Eastern Polynesia are more tenable (Spriggs and Anderson 1993).

At the time of first European contact the sweet potato (*Ipomaea batatas*) was one of the dominant crops in Eastern Polynesia. It has subsequently been ascertained that this crop originated in South America and that its wide distribution in Eastern Polynesia does not result from trading during historical times (Hather and Kirch 1991). Archaeological evidence, though slow to materialise, does support a prehistoric origin for this crop in Polynesia with storage pits in New Zealand and carbonised yet identifiable tubers from Hawaii and Mangaia in the Cook Islands. A three-stage model for this crop's introduction to the region has been proposed with an initial introduction into central Eastern Polynesia followed by dispersal to the marginal extremes of

Eastern Polynesia and a subsequent dispersals to Southeast Asia and Papua New Guinea by Spanish and Portuguese colonists in the 15th and 16th centuries (Hather and Kirch 1991). It is often glossed over or ignored altogether that this scenario has the absolute requirement that either Polynesians voyaged to and returned from South America or that South Americans themselves made the journey. This reticence undoubtedly relates to an unwillingness to resurrect the now heavily discredited theory of Thor Heyerdahl; that the first colonists of Polynesia were Amerindian (Heyerdahl 1950).

Genetic analyses

Analysis of classical, nuclear-encoded, markers in Polynesia weakly supported a Southeast Asian origin with perhaps some Melanesian admixture (Hill and Serjeantson 1989), a contribution which was further supported by the discovery, in Polynesia, of a specific thalassaemia allele of Melanesian origin (Hill et al. 1985). There was no strong support for an American origin although data from classical loci were never able to exclude this possibility because native Amerindians themselves have an origin in Asia. In general surprisingly little polymorphic information has been found in autosomal loci to discriminate between the very different population histories of Melanesians and Polynesians.

Diversity at six autosomal minisatellites has been used to show that whilst Melanesian populations show little evidence of reduced diversity, Polynesian populations show much lower heterozygosity resulting from the recent population bottlenecks and founder effects that characterised the colonisation process (Flint et al. 1989).

In contrast to these uncertainties, studies utilising mitochondrial DNA (mtDNA) have been particularly informative (Lum et al. 1994; Melton et al. 1995; Redd et al. 1995; Sykes et al. 1995). A common lineage cluster comprising three or four characteristic base substitutions in the control region and a 9bp deletion elsewhere in the mitochondrial genome has been found in 94% of all Polynesian mtDNAs. It has also been found at moderate frequency in coastal Papua New Guinea. Ancestral haplotypes have been traced back to Indonesia, the Philippines and Taiwan (Melton et al. 1995; Sykes et al. 1995). Other maternal lineages identified in Polynesia (Lum et al. 1994) were confirmed as Melanesian Australoid admixture of about 4% (Sykes et al. 1995). A common finding in all genetic studies has been a striking lack of diversity in Polynesia compared to source populations in Melanesia and Southeast Asia (Flint et al. 1989; Lum et al. 1994; Sykes et al. 1995).

Together with a cline of diversity within mitochondrial lineage groups from high in West Polynesia to low in East Polynesia (Sykes et al. 1995), this suggests there have been, not surprisingly, severe population bottlenecks during the colonisation of Polynesia. There is no mtDNA evidence for a substantial input from the Americas or Europe.

mtDNA has also been used to address the question of the geographical location of dispersal origins. One study used mtDNA sequences and autosomal *Alu* insertions to support the hypothesis that Taiwan represent the proto-Austronesian homeland (Melton et al. 1998). Another study suggested that mtDNA diversity is more congruent with an origin of the Polynesian expansion in Eastern Indonesia (Richards et al. 1998).

Simulation studies based on the mtDNA found amongst New Zealand Maori and sampling from an ancestral central Polynesian population have been used to estimate the size of the founding population given its known age (Murray-McIntosh et al. 1998). A founding population of between 50 and 100 women was deemed to be most likely. This fits nicely with the oral traditions of the Maori which describe 8-10 founding canoes, each containing 10-20 people.

Two anthropologically well-informed recent papers from (Lum and Cann 1998; Lum et al. 1998) have investigated correlations between genetic markers and language in Oceania. The first investigated Micronesian and Polynesian mtDNA and found significant correlations with language, using Mantel tests (Lum and Cann 1998). In addition the data indicated that Polynesian islands were relatively isolated compared to the higher gene-flow apparent amongst Micronesian islands (Lum and Cann 1998). The second, impressive, study compared diversity in mtDNA and autosomal microsatellites in 28 Oceanic and Islands Southeast Asian populations (Lum et al. 1998). Again mtDNA and linguistic differences were found to be well-correlated, however the autosomal diversity was better correlated with geographical distance, as would be expected under an isolation by distance model of relatively high gene-flow. This dichotomy was taken to suggest that while mtDNA diversity retained the imprint of the initial settlement patterns the distribution of autosomal diversity was more determined by the post-colonisation inter-island contacts. Consequently a higher male-specific migration rate was proposed.

Little work has been done on the Y-chromosomal diversity within Oceania. One early paper used the 49f polymorphic marker to delineate a number of different Y-chromosomal lineages in Polynesia (Spurdle et al. 1994). One of the three major lineages was found in one third of New Zealand Maori and was hypothesised to constitute European admixture; this lineage was also found at the same frequency in a small Rarotongan sample (n=3) and at lower frequency in Samoans.

Otherwise the only major conclusion from this paper was that "Polynesians have greater affinity to Caucasoids than to African populations."

A study of deviations from the single step-wise mutational model at the Y-chromosomal microsatellite *DYS390* revealed the presence of two monophyletic lineages defined by two independent deletions (Forster et al. 1998). One was found at high frequency in Australian Aborigines and the other in Papua New Guineans and Samoans. As such they constitute potentially useful population-specific Y-chromosomal markers for further study of these populations.

A single recent study has sought to compare Y-chromosomal microsatellite diversity with mtDNA and HLA diversity in a sparse sampling of eight populations (Hagelberg et al. 1999). Little of interest to the region as a whole was concluded excepting the fact that whilst mtDNA diversity was found to decrease throughout the Pacific Y-chromosomal diversity was more constant, perhaps suggesting the sex-specific difference in migration patterns described earlier. This paper focused mainly on the genetic affinities of the Tolai, Roro and Trobriand islanders which share morphological and cultural features with both Papuans and Austronesians. Sex-specific differences were impossible to conclude in this respect due to the inadequacy of the Y-chromosomal data in not allowing monophyletic lineages to be defined.

Humans are not the only species whose genetic diversity can inform studies of Oceanic origins. Other species were carried in the colonising canoes. This provides the possibility that the distribution of genetic diversity in these introduced species may well mirror the colonisation process. This approach is related to comparative phylogeography; the investigation of a common landscape through the investigation of genetic diversity in multiple species (Bermingham and Moritz 1998). However in this case the landscape is not a geographic but an anthropogenic scenario. Two species have been studied to this end, the Pacific rat (Matisoo-Smith et al. 1998) and a hitch-hiking lizard (Austin 1999). The distributions of sampling points were substantially different. The study of the lizard, *Lipinia noctua*, throughout Melanesia, Micronesia and Polynesia strongly supported an express train model of colonisation (Austin 1999). The study of the rat, *Rattus exulans*, mainly from Central and East Polynesia indicated multiple contacts between a central Polynesian sphere of influence, encompassing the Cooks and Society islands, and the more distant archipelagos of Hawaii and New Zealand (Matisoo-Smith et al. 1998). This study also revealed a relative isolation of the Marquesas from this central sphere of influence.

Some contentious issues in Polynesian prehistory

There are a number of contentious issues in Oceanic prehistory that it may be possible to address using genetic information, specifically that from Y-chromosomal markers. A number of these are described briefly below.

The genetic relationship of the Polynesia outliers to their Melanesian and Micronesian neighbours and to their cultural and linguistic kin in Polynesia is relatively unknown. At present it has been suggested that these outliers were colonised by Polynesian voyaging from Western Polynesia (Pawley and Ross 1993), though the timing of these colonisations are poorly characterised.

The times and routes of the colonisation events in Eastern Polynesia are also poorly understood. Whilst some remain convinced that the Marquesas were the launching pad of subsequent dispersals (Bellwood 1987), particularly to Hawaii and Easter Island, New Zealand seems more likely to have been colonised from the Cook Islands.

The hypothesised dispersal centres of the hierarchy of proto-languages might well represent homelands or significant staging posts in the expansion of Autronesian peoples. At present dispersal centres tend to be defined on the basis of the greatest linguistic diversity within that particular subgroup of languages (Pawley and Ross 1993). It would be of interest to see if these are congruent with dispersal centres identified through genetic diversity.

It has been suggested that the existence of primitive isolates in Oceania is a myth (Terrell 1986; Terrell 1997), although it has been noted that isolation is a relative term and not an absolute. It has also been suggested that there are sex-specific differences in gene flow (Lum et al. 1998). Quantification of the extent and variation of this isolation throughout Oceania would help to settle this question which is central to much of the disagreement surrounding other contentious issues. For example a related issue is whether Oceania could be better sub-classified into Near and Remote (Kirch and Green 1992). A recent proposal has claimed that this classification might better reflect the relative importance of recent settlement versus ancient occupation within Melanesia (Kirch and Green 1992).

The extent of admixture in different extant Oceanic populations has been little addressed by archaeologists and linguists more interested in prehistory. However identifying and excluding the 'noise' of admixture is vital to making reliable prehistorical inferences on the basis of extant genetic

data. Western Europe would seem to be a likely candidate as a source for admixture and has already been implicated in the low resolution Y-chromosomal study mentioned above (Spurdle et al. 1994). In addition the investigation of potential Amerindian admixture may well help to resolve the questions surrounding the origin of sweet potato. Such information may also prove to be useful to the forensic communities in this region of the world.

Polynesia has found itself the source of appreciable archaeological angst over methodology (Durham 1992). Evolutionary Culture Theory (ECT) seeks to explain archaeological diversity through appreciation of evolutionary processes. It emphasises distinguishing between homologous and analogous forms, and resolving divergence and convergence. It considers that much recent archaeology has focused on ethnogenesis and convergence to analogous forms (Durham 1992). Unsurprisingly ECT borrows heavily from concepts developed in the study of biological evolution. Isolation is a powerful mechanism underpinning divergence, and as such the isolated islands of Polynesia have become a testing ground for this novel methodology. To a biologist it would seem that much of the heated disagreement centres on confusion between the use of Darwinian evolution as a metaphor and as a mechanism. In addition there appears to be an uncommon resistance to making explicit assumptions which have clearly been used implicitly in archaeology for many years.

Outline of this study

Oceania has a simple, well documented and well dated history of population movements from the point of view of archaeology, linguistics and maternal lineages; it thus represents an ideal region for studies of paternal lineages.

Mating practices, the cultural phenomenon of patrilocality and the small effective population size of the Y chromosome result in a high degree of geographical differentiation that has been utilised to investigate prehistoric migration events e.g.(Underhill et al. 1996; Zerjal et al. 1997). It is the potential presence of different Y-chromosomal lineages in Southeast Asia, the Americas and Europe that might allow us to investigate the origins of Polynesian Y chromosomes. The Y chromosome is likely to be more sensitive than other loci to certain kinds of admixture, for example recent male-dominated admixture between populations normally separated by geographical distance.

This study was performed in two parts which will be described here in chronological order. The major aim of the preliminary study was to ascertain whether the present informative capacity of Y-chromosomal polymorphisms was sufficient to allow inferences on population history. In practice this meant whether a sufficient number of informative lineages could be delineated in the population samples. This first study complements previous mtDNA analysis by examining the paternal lineages of two populations, one from coastal Papua New Guinea (Port Moresby) and the other from Polynesia (Cook Islands) using Y-chromosomal markers which can be assayed by PCR. Thirty-three unrelated Cook Island and 58 unrelated Papua New Guinea samples were assayed for all polymorphisms. Thirty out of the thirty-three Cook Islander samples have the common Polynesian mtDNA haplotype characterised by control region transitions at 16189, 16217, 16247 and 16261. Seven base substitutions and one insertion/deletion were used to distinguish ten possible Y chromosome haplogroups, and diversity within observed haplogroups was assayed using seven microsatellite loci and the minisatellite, MSY1. This preliminary study has been published (Hurles et al. 1998), the resulting paper is in appendix C, and therefore is only briefly summarised in the text.

Having determined that Y-chromosomal polymorphisms were indeed of sufficient utility, the study was expanded to include samples from Southeast Asia and Oceania. Whilst not all of these samples had previously been typed for mtDNA, as they had been in the preliminary study, the landscape of mtDNA diversity in this region has been well characterised in other studies (Sykes et al. 1995; Lum and Cann 1998). Including the two populations typed previously, 420 chromosomes were analysed in 18 populations from throughout the region. In the latter study two new biallelic polymorphisms were assayed, while others that had been assayed before were excluded on the basis of being uninformative. In addition MSY1 codes were ascertained in all samples.

The analytical tools used to make population historical inferences from these data have been introduced and further adapted in the previous two chapters.

Materials and Methods

Samples

All samples in the two studies described in this chapter were kindly donated by Bryan Sykes and John Clegg. Cook Islander samples came from the island of Rarotonga and Papua New Guinean samples from Port Moresby. Four aboriginal Taiwanese tribes were sampled, the Atayal, Ami, Bunumi and Paiwan. A sample of autochthonous males from the Philippines was also assayed for diversity. Two Borneo samples were typed for all markers, from Kota Kinabalu in the Northern Malaysian province of Sabah and from Banjarmasin on the southern coast in the Indonesian province of Kalimantan. Two islands from Micronesia were sampled; from Majuro in the Marshall Islands and from the Polynesian outlier, Kapingamarangi. Two populations from the Island Melanesian group Vanuatu were sampled, from Port Olry and from the island of Maewo. Three different Tongan samples were used in this study. Additional Polynesian samples came from the islands of Western Samoa and Rapa, in French Polynesia. These samples are shown geographically in Figure 6.3.

Data

Biallelic markers and minisatellite MSY1 were assayed by me in Leicester, seven-locus microsatellite haplotypes were typed by Catherine Irven and Jayne Nicholson in Bryan Sykes' lab in Oxford. I performed all the analytical work presented here.

Markers

Two additional, unpublished, biallelic markers, RPS4Y (Bergen et al. 1999) and Lly22g were typed in the second study; they were kindly provided by Andrew Bergen and Chris Tyler-Smith respectively.

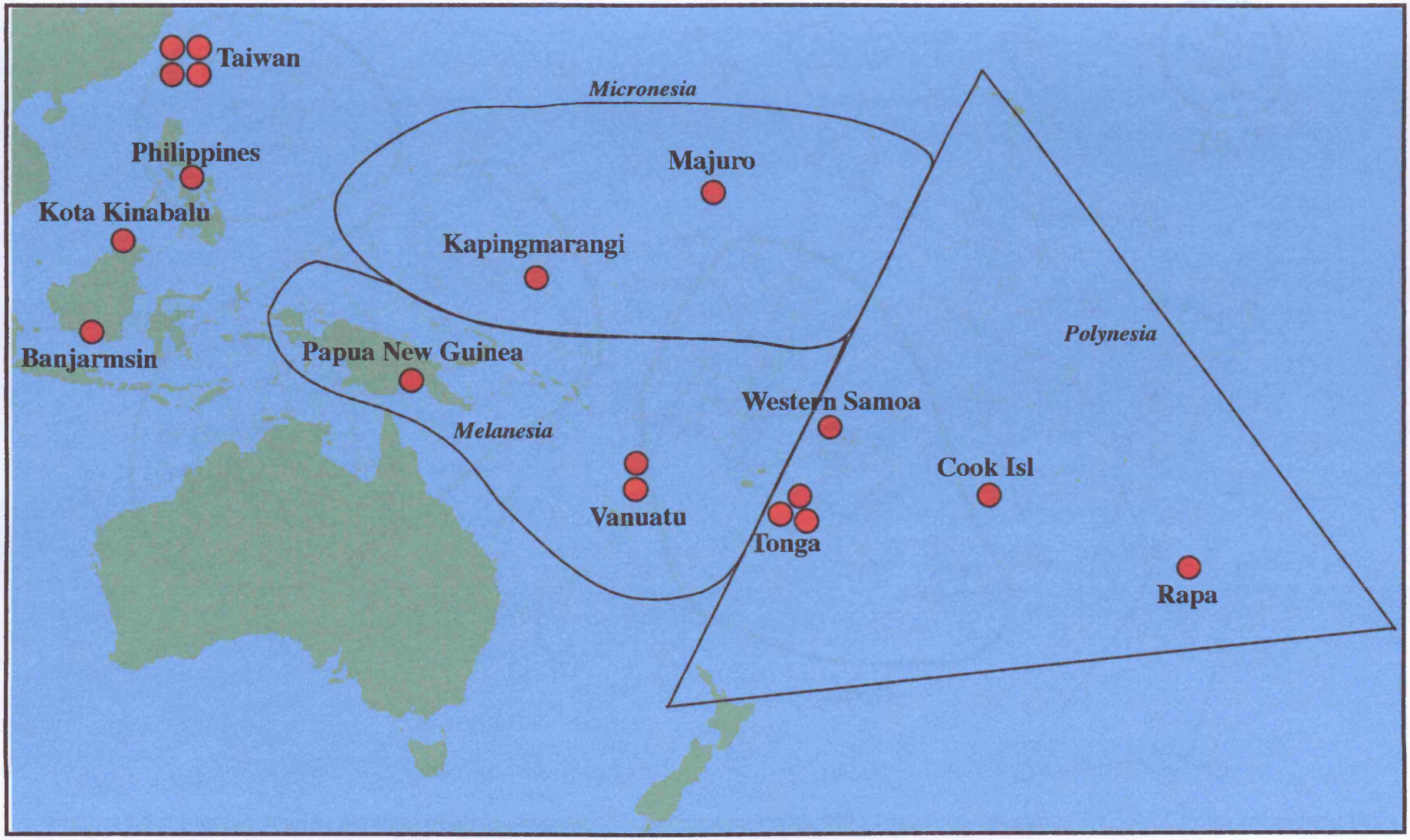


Figure 6.3: Map of the samples sites included in the extended study

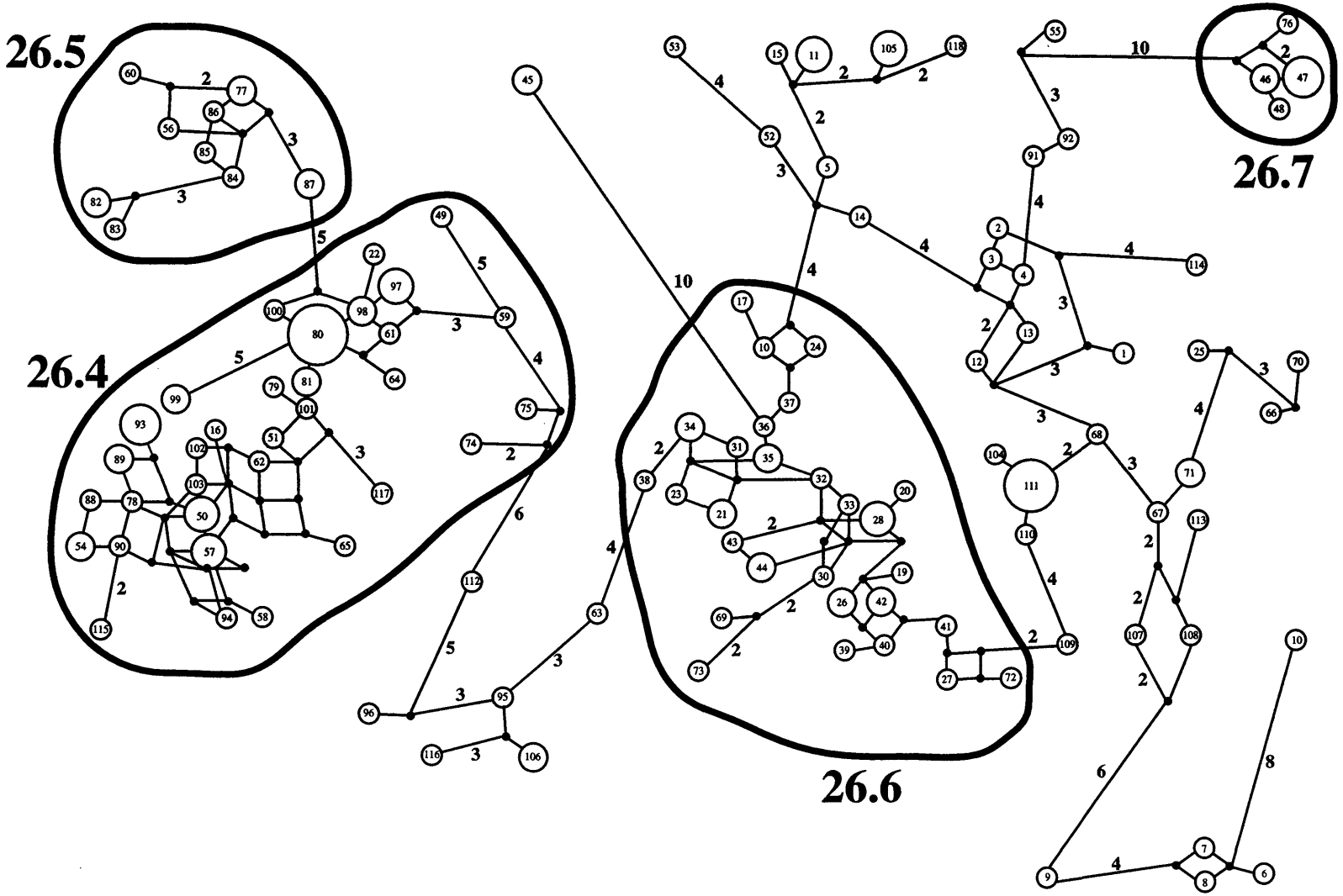


Figure 6.5a: Median-Joining network of all haplogroup 26 chromosomes with MSY1 modular structure (3,1,3,4). Circles are individual MSY1 codes each with its own number identifier. All branch lengths are of a single mutational step, except where indicated. The bold lines and numbering indicate the sublineages defined in this study. Circle area is proportional to frequency.

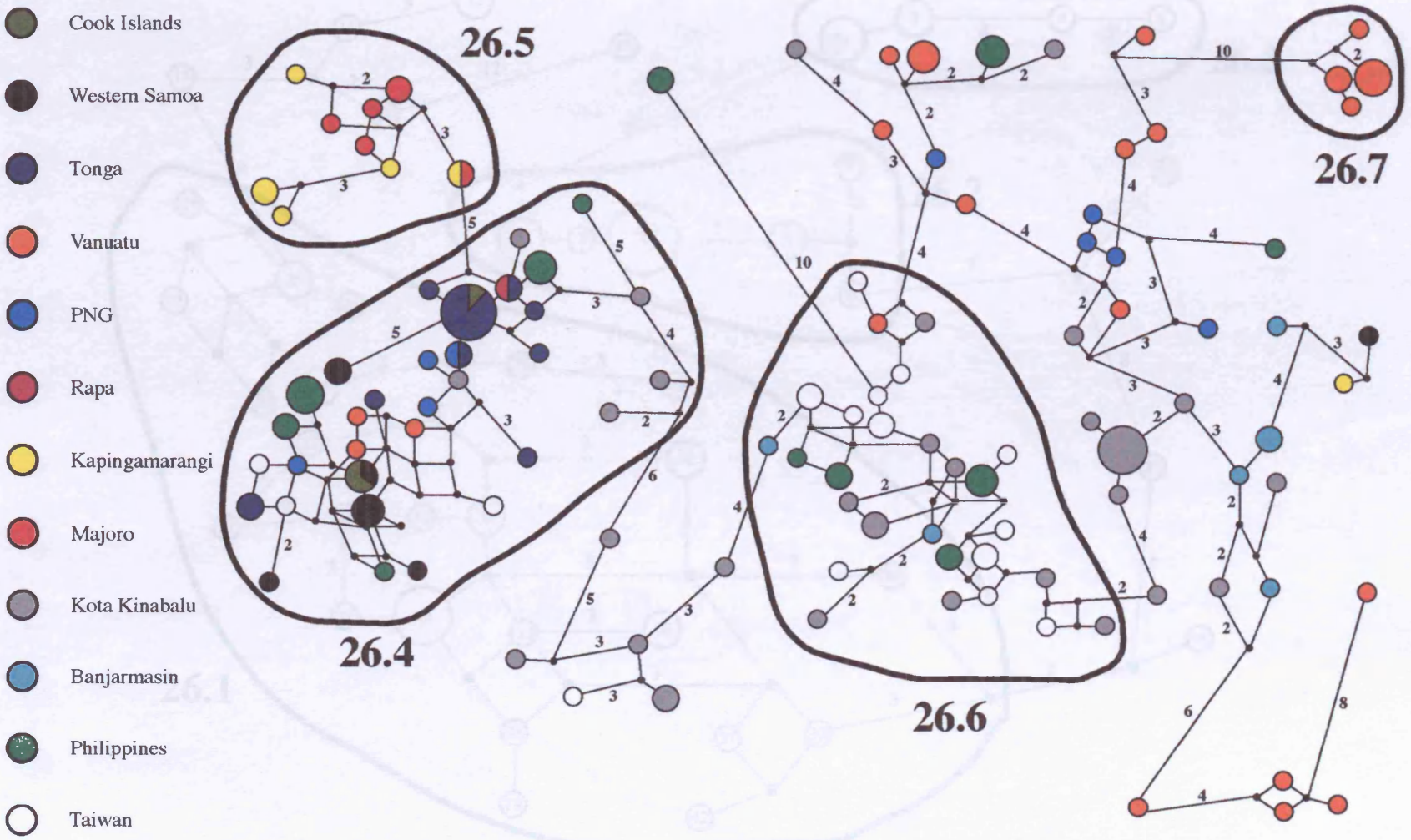


Figure 6.5b: Median-Joining network of all haplogroup 26 chromosomes with MSY1 modular structure (3,1,3,4). Circles are individual MSY1 codes coloured according to their population affiliation. All branch lengths are of a single mutational step, except where indicated. The bold lines and numbering indicate the sublineages defined in this study.

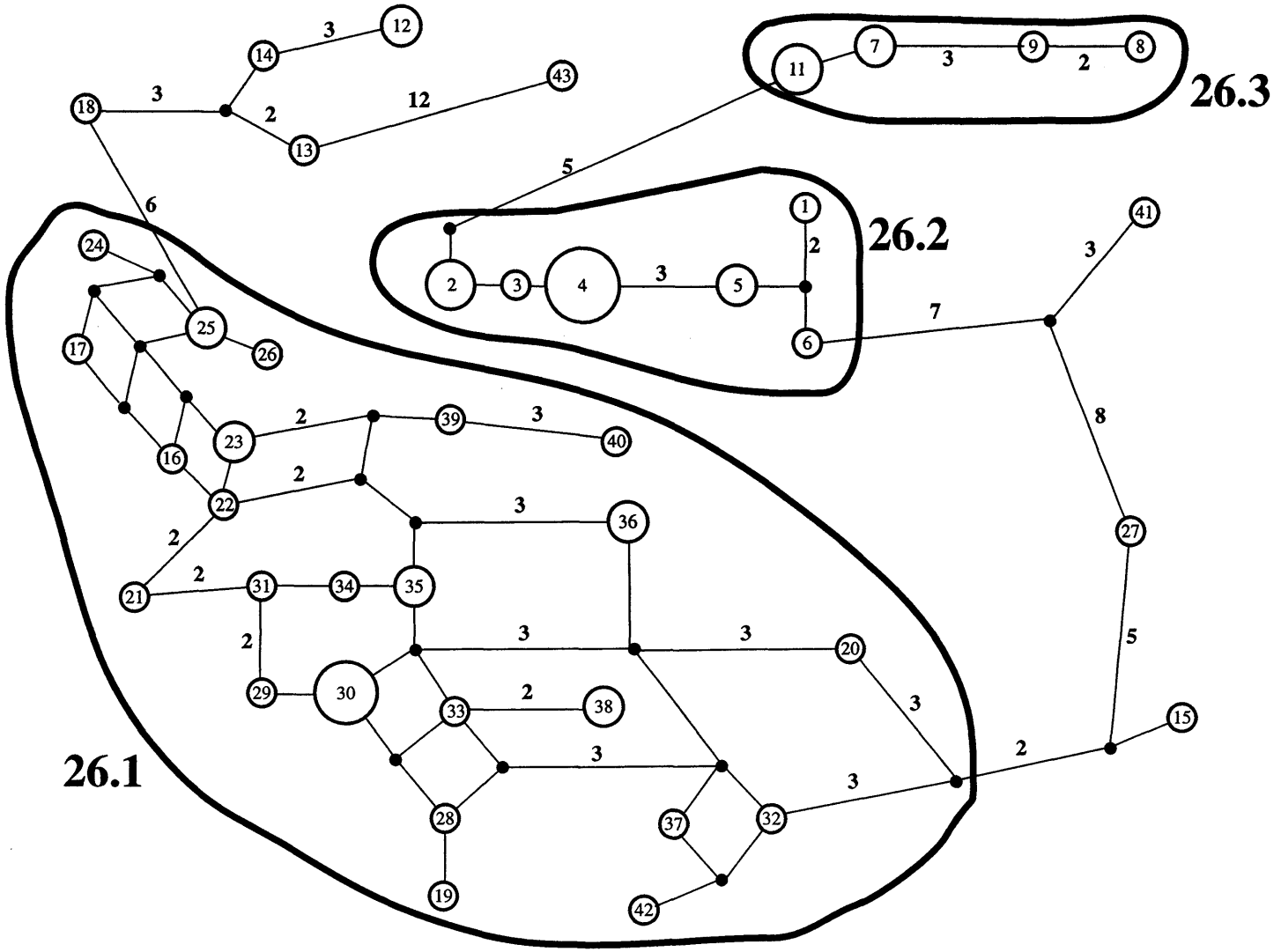


Figure 6.6a: Median-Joining network of haplogroup 26 chromosomes having the (1,3,4) MSY1 modular structure. Circles are individual MSY1 codes each with its own number identifier. All branch lengths are of a single mutational step, except where indicated. The bold lines and numbering indicate the sublineages defined in this study. Circle area is proportional to frequency.

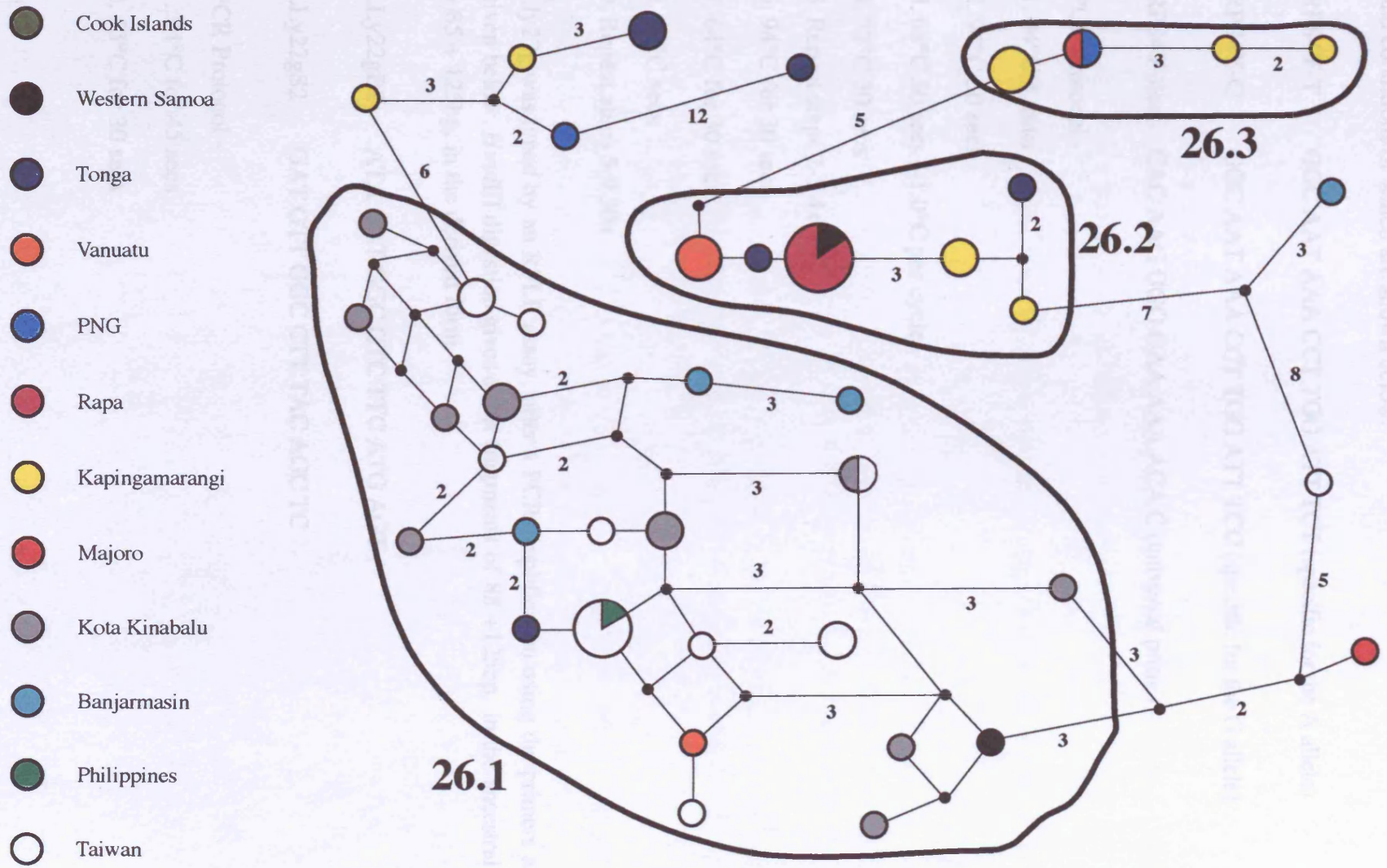


Figure 6.6b: Median-Joining network of all haplogroup 26 chromosomes with MSY1 modular structure (1,3,4). Circles are individual MSY1 codes coloured according to their population affiliation. All branch lengths are of a single mutational step, except where indicated. The bold lines and numbering indicate the sublineages defined in this study.

RPS4Y was typed as an allele-specific amplification in a touchdown protocol, the primers and conditions of which are shown below:

RPS4Y-T GGC AAT AAA CCT TGG ATT TCT (specific for the A allele)

RPS4Y-C GGC AAT AAA CCT TGG ATT TCC (specific for the G allele)

RPS4Y-non CAC AAG GGG GAA AAA ACA C (universal primer)

PCR Protocol -

1. 94°C 4 mins
2. 94°C 30 secs
3. 68°C 30 secs (-1.0°C per cycle)
4. 72°C 30 secs
5. Repeat steps 2-5 4x
6. 94°C for 30 secs
7. 64°C for 30 secs
8. 72°C secs
9. Repeat steps 5-9 30x

Lly22g was typed by an RFLP assay, after a PCR amplification using the primers and conditions given below. *Hind*III digestion gives either fragments of 85 +125bp, in the ancestral state, or 210 + 85 + 125bp, in the derived form.

LLy22gB2 ATA GAT GGC GTC TTC ATG AGT

LLy22gS2 GAT GTT GGC CTT TAC AGC TC

PCR Protocol -

1. 94°C for 45 secs
2. 55°C for 30 secs

3. 72°C for 30 secs

4. Repeat steps 1-3 33x

Software

The program Microsat used to calculate ASD and $\delta\mu^2$ was written by Eric Minch and obtained from the web site, 'http://lotka.stanford.edu'.

Median-Joining Networks were calculated using the program 'Network 2.0' from Arne Röhl (Bandelt et al. 1999).

The Microsoft Excel spreadsheet used to calculate dates for the lineages was based on a spreadsheet designed to calculate distances between microsatellite haplotypes for the drawing of minimum-spanning trees that was developed and kindly supplied by Fabricio Santos.

Permutation testing of pairwise F_{ST} comparisons, diversity calculations and Mantel tests were made using the beta-test version of Arlequin v2.0 software available from the web site 'http://anthropolgie.unige.ch/arlequin'.

Principal Components Analysis was performed using the statistical package Minitab 8.0. Results were exported into Microsoft Excel from where the graphs were drawn.

Results

Preliminary study

Five different haplogroups were defined by the biallelic markers within the Papua New Guinean (PNG) and Cook Islander (CIs) samples: haplogroups 1, 2, 3, 26 and 24. Two of these lineages - haplogroups 2 and 26 - were further subdivided into assumed monophyletic sublineages on the basis of unique MSY1 modular structures, making a total of 7 well-defined lineages. Both of these MSY1 modular structures, (3,1,3+,4-) within haplogroup 26 and (...4,0,4) within haplogroup 2, are significantly different from any others present in a database of 465 codes from around the world.

Four of these lineages were found at relatively high frequency in these samples, sufficient to allow population historical inferences. The Papua New Guinean sample was dominated by the presence of a single lineage found at a frequency of 64%; haplogroup 24, defined by the M4 base substitution. This lineage was completely absent from the Polynesian sample. The question arises as to whether this absence is due to population history, for example because of little admixture between the proto-Polynesian and Papua New Guineans, or due to the young age of this lineage that might post-date this population movement. This lineage has only been found at low frequency in neighbouring populations making it highly likely that it originated in PNG. Consequently this lineage was dated using the microsatellite and MSY1 diversity within it. Wide confidence limits on the age of this lineage were calculated that made it impossible to exclude an origin within the past 3,600 years. This is the age of the oldest archaeological site deemed to belong to the Lapita culture, and defines the likely timing of the migration of proto-Polynesians across Melanesia. However an origin for haplogroup 24 before this date was deemed to be more probable with an age of 4-6,000 years being most likely. Thus this absence from Polynesia may well reflect little contact between the proto-Polynesians and the Melanesians of Papua New Guinea; however 3,000 years ago the frequency of haplogroup 24 is likely to have been very much less than it is today and initial sampling followed by subsequent bottlenecks may well have erased all traces of this lineage from Polynesia. Certainly the absence of this lineage indicates that there was little or no contact between PNG populations and Polynesians during the past 3000 years, thus arguing against substantial trading contacts between these two populations.

Three lineages were found at high frequencies (>10%) in the Polynesian sample. They are haplogroup 1, haplogroup 2 chromosomes with the (...4,0,4) MSY1 modular structure and haplogroup 26 chromosomes with the (3,1,3+,4-) modular structure. The two sublineages defined by the MSY1 modular structures were also found in the PNG sample and so could be traced to a Southeast Asian heritage. The third lineage, haplogroup 1, was found at a frequency of 27% in the Cook Islander sample. All examples of this chromosome have the same MSY1 modular structure. The world-wide distribution of this haplogroup is well known to be extensive, though haplogroup 1 chromosomes with the MSY1 modular structure (1,3,4) are only known to be found at appreciable frequencies in Western Europe (Jobling et al. 1998). Asian examples of this haplogroup tend to have the alternative modular structure (3,1,3,4), whereas the modular structure of Amerindian haplogroup chromosomes is largely unknown, as only a single MSY1 code from a haplogroup 1 chromosome of this provenance has been determined. It was found to have the modular structure (1,3,4) (Jobling et al. 1998). Two other lineages defined by a combination of biallelic markers and MSY1 subtype are known to be present at appreciable frequencies in Western Europe. These are haplogroup 2 chromosomes with the MSY1 modular structure (3,1,3,4), and haplogroup 3 chromosomes with the (1,3+,4) MSY1 modular structure. These lineages are almost entirely absent from Eastern Asia (Jobling et al. 1998; Chris Tyler-Smith, personal communication, and unpublished observations) and were also found in the Cook Islander sample. In contrast, the lineage present at highest frequency in Amerindians - haplogroup 18 - was absent from this sample. The relative proportions of these non-Eastern Asian lineages were found to be almost exactly the same as in Western Europe. Probabilistic arguments were used to exclude an American origin for these chromosomes, and to hypothesise that they represent significant European admixture with the almost exclusively male post-contact groups. This admixture had not previously been seen with mtDNA-defined maternal lineages and represented 33% of all Y chromosomes in this sample. This fits with an earlier study that analysed the 49f marker system and postulated that a third of the Y chromosomes from two Polynesian populations (Cook Islands and New Zealand Maori) represent European admixture (Spurdle et al. 1994).

The two lineages found in the Cook Islander sample that were shared with the PNG sample comprised the majority (55%) of CIs Y chromosomes. One of these lineages, haplogoup 2 chromosomes with the (...4,0,4) MSY1 modular structure, was also characterised by one of the *DYS390* deletion alleles. Although sample sizes were small it appeared that diversity within both these lineages was substantially reduced in the CIs sample compared to the PNG sample, as would be expected from population bottlenecks.

Extended study

At the time of writing microsatellite data were not available from our collaborators on this project, and therefore all subsequent analyses presented here come from my assaying of biallelic and MSY1 diversity in the samples described above. The complete data on all samples the new samples included in the extended study are given in appendix E.

Due to their geographical proximity and small sample sizes the samples within Taiwan were pooled in most analyses, as were the samples within Vanuatu and within Tonga.

Lineage Analysis

The biallelic markers defined nine different haplogroups that were present in these populations. The majority of the 420 Y chromosomes assayed in this study belong to a single lineage, haplogroup 26. The next most represented lineage was defined by the new marker, RPS4Y, and subdivided the chromosomes that had previously been assigned to haplogroup 2. Haplogroups 10 and 11 in the tree of haplogroups presented in the general introduction are defined by RFLP assays scored by Southern hybridisation. A number of diverse chromosomes from these two lineages have been typed with RPS4Y (data not shown). All show the derived state of this polymorphism. Consequently it is assumed here that the RPS4Y marker is congruent with the RFLP marker that defines haplogroup 10. Thus, for simplicity's sake, rather than assign a new haplogroup number to define this lineage, chromosomes exhibiting the derived state of the RPS4Y marker in this study are designated haplogroup 10. In actuality the chromosomes thus defined may belong to haplogroup 11 or belong to a new haplogroup intermediate between haplogroups 2 and 10. There is not sufficient DNA available to resolve this issue by typing the relevant Southern hybridisation assays. However this is purely a problem of nomenclature and not of lineage definition. All the haplogroups assayed and their frequency in this data set are summarised in figure 6.4.

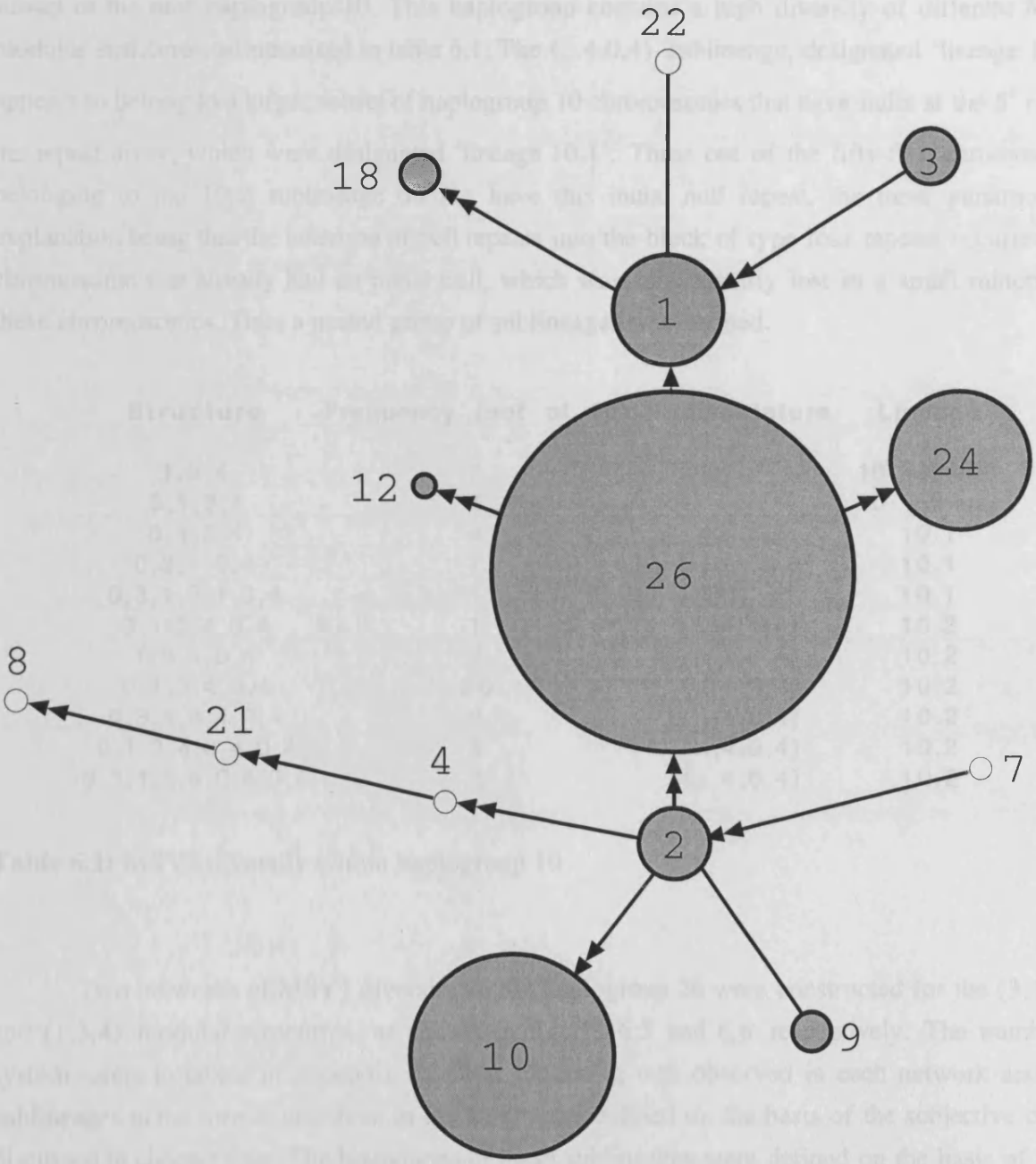


Figure 6.4: Tree of assayed haplogroups. An unfilled circle represents an absence of that lineage in the sample set. Circle area is proportional to frequency in the entire sample of 420 chromosomes. Numbers indicate haplogroup nomenclature. Arrows indicate known ancestral state.

In order to improve the resolution of lineages within the data set, the MSY1 diversity within the two major lineages was further investigated to see if further sublineages could be discerned.

The chromosomes belonging to the (...4,0,4) sublineage defined in the preliminary study are a subset of the new haplogroup 10. This haplogroup contains a high diversity of different MSY1 modular structures, summarised in table 6.1. The (...4,0,4) sublineage, designated 'lineage 10.2', appears to belong to a larger subset of haplogroup 10 chromosomes that have nulls at the 5′ end of the repeat array, which were designated 'lineage 10.1'. Three out of the fifty-five chromosomes belonging to the 10.2 sublineage do not have this initial null repeat, the most parsimonious explanation being that the insertion of null repeats into the block of type four repeats occurred in a chromosome that already had an initial null, which was subsequently lost in a small minority of these chromosomes. Thus a nested group of sublineages was defined.

Structure	Frequency (out of 80)	Nomenclature	Lineage
1,3,4	7		10 (others)
3,1,3,4	4		10 (others)
0,1,3,4	4	(0,...)	10.1
0,3,1,3,4	7	(0,...)	10.1
0,3,1,3,1,3,4	1	(0,...)	10.1
3,1,3,4,0,4	1	(...4,0,4)	10.2
1,3,4,0,4	2	(...4,0,4)	10.2
0,1,3,4,0,4	50	(...4,0,4)	10.2
0,3,1,3,4,0,4	2	(...4,0,4)	10.2
0,1,3,4,0,4,0,4	1	(...4,0,4)	10.2
0,3,1,3,4,0,4,0,4	1	(...4,0,4)	10.2

Table 6.1: MSY1 diversity within haplogroup 10

Two networks of MSY1 diversity within haplogroup 26 were constructed for the (3,1,3,4) and (1,3,4) modular structures, as shown in figures 6.5 and 6.6 respectively. The numbering system refers to tables in appendix E. Clear clustering was observed in each network and four sublineages in the former and three in the latter were defined on the basis of the subjective criteria discussed in chapter four. The boundaries of these sublineages were defined on the basis of larger mutational distances between a sublineage and the next closest MSY1 code than occur within the lineage itself. On rare occasions single cases of blatant saltatory mutations within sublineages were not excluded from the lineage, for example code number 99 in the 26.4 sublineage. Geographical clustering of the lineages defined on the basis of MSY1 codes could not be used as a criterion for defining sublineages as the subsequent analysis of the geographical distributions of these sublineages would introduce circularity into the analysis. However the fact that many of the geographical analyses presented later in this chapter, and based on the lineages defined here, find

significant geographical differentiation with respect to random permutations of the the same data strongly argues that the lineages defined here are not analytical contrivances. In addition the haplogroup 26 sublineage defined in the preliminary study, the monophyletic nature of which was confirmed with microsatellite analysis, was also reconstituted here. The sublineage (3,1,3+,4-) identified in the preliminary study falls into the sublineage 26.4 in this extended study. Furthermore the sublineages 26.1, 26.2, 26.3, 26.4, 26.5, 26.7, 10.1 and 10.2 are all defined by MSY1 modular structures and block size ranges that if found previously in a database of 465 world-wide codes have only been observed within their respective haplogroups and only in this region of the world.

This analysis allowed a total of 18 lineages to be defined in this data set, vastly improving the resolution of this study. These lineages were arranged into a tree structure, shown in figure 6.7. Although haplogroup 10 could be arranged into nested sublineages, the sublineages belonging to haplogroup 26 were arranged independently rather than making any unsubstantiated assumptions about the relationships among sublineages. It was not even assumed that sublineages having the (1,3,4) modular structure were more closely related to one another than to sublineages having the (3,1,3,4) modular structure. It is known from previous work on MSY1 diversity (Jobling et al. 1998) that switching between these two modular structure occurs within many lineages and so in a relatively ancient lineage such as haplogroup 26 switching is likely to have occurred multiple times.

The population distributions of these 18 monophyletic lineages are summarised in table 6.2.

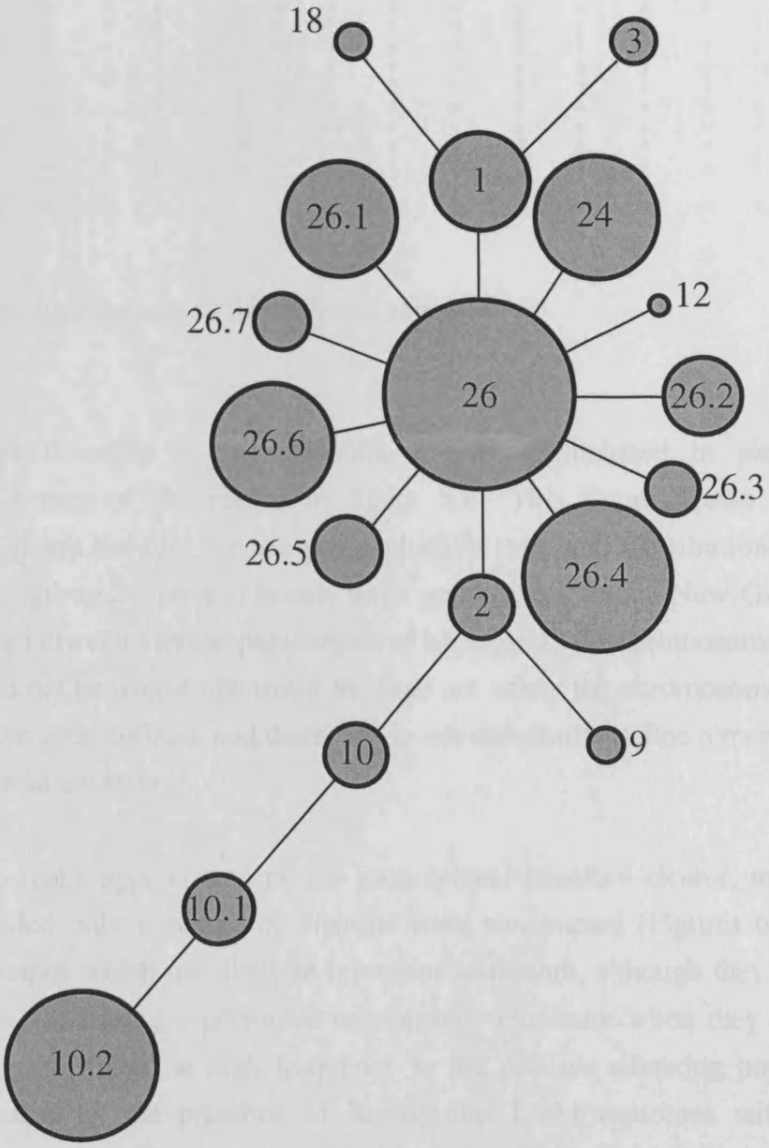


Figure 6.7: Tree of lineages found in this dataset. Circle area is proportional to frequency. Numbers indicate lineage nomenclature.

Lineage / Population	1	2	3	9	10.2	10.1	10	12	18	24	26.4	26.1	26.7	26.2	26.3	26.5	26.6	26	Total
Ami	0	0	0	0	0	0	0	0	0	0	3	3	0	0	0	0	5	3	14
Atayal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	7
Banjarmasin	0	1	1	1	0	0	5	0	0	0	0	3	0	0	0	0	2	10	23
Bunumi	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	3	0	8
CI	9	3	1	0	14	2	0	0	0	0	3	0	0	0	0	0	0	1	33
Filipino	1	1	0	0	0	0	0	0	0	0	11	1	0	0	0	0	8	8	30
Kapamarangi	0	0	0	0	6	0	0	0	0	0	0	0	0	3	5	6	0	1	21
Kota Kinabalu	1	2	3	1	0	6	5	1	0	0	5	12	0	0	0	0	11	25	72
Maewo	1	0	0	1	0	3	1	0	0	4	3	0	1	4	0	0	0	9	27
Majuro	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	6	0	3	11
Paiwan	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	1	3	10
PNG	0	1	0	0	2	0	0	0	0	28	4	0	0	0	1	0	0	8	44
Port Olry	2	0	0	0	5	0	0	0	0	1	0	1	7	0	0	0	0	15	31
Rapa	10	0	0	0	6	0	0	0	3	0	1	0	0	6	0	0	0	1	27
Tongan	0	1	0	0	3	0	0	0	0	2	3	0	0	0	0	0	0	0	9
Tongan Iysate	0	0	0	0	3	0	1	0	0	3	10	1	0	0	0	0	0	1	19
Vav'ua	0	0	0	0	1	0	0	0	0	0	4	0	0	1	0	0	0	1	7
W. Samoa	0	0	0	0	15	1	0	0	0	0	8	1	0	1	0	0	0	1	27
Total	24	10	5	3	55	12	12	1	3	38	55	33	8	15	7	12	37	90	420

Table 6.2: Lineage distributions in the different samples

The lineage diversity of each population was summarised in pie chart form and superimposed on a map of the region in figure 6.8. This figure shows clear geographical substructure, with many lineages having geographically restricted distributions. Note the limited distribution of haplogroup 24, present in only three populations: Papua New Guinea, Vanuatu and Tonga. The sharing between various populations of haplogroup 26 chromosomes not belonging to a sublineage should not be over-emphasised as these are solely the chromosomes that remain after the sublineages have been defined, and therefore do not necessarily define a monophyletic grouping with a recent common ancestor.

In order to make appreciation of the geographical structure clearer, maps of simpler pie charts which included only a subset of lineages were constructed (Figures 6.9-11). Figure 6.9 identifies those lineages which are likely to represent admixture, although they may only do so in certain populations. Lineages are presumed to represent admixture when they are found together with other lineages also found at high frequency in the putative admixing population. European admixture is indicated by the presence of haplogroup 1 chromosomes with MSY1 modular structure (1,3,4) together with haplogroup 2 chromosomes of the (3,1,3,4) modular structure, whereas Amerindian admixture is indicated by haplogroup 18 chromosomes together with chromosomes belonging to haplogroup 1 (Karafet et al. 1999). Whereas haplogroup 2 chromosomes in the Cook Islander sample resemble closely European chromosomes belonging to the same lineage, not all haplogroup 2 chromosomes in Southeast Asia have the same, diagnostic, MSY1 modular structure and so may not represent admixture, merely a low frequency of other

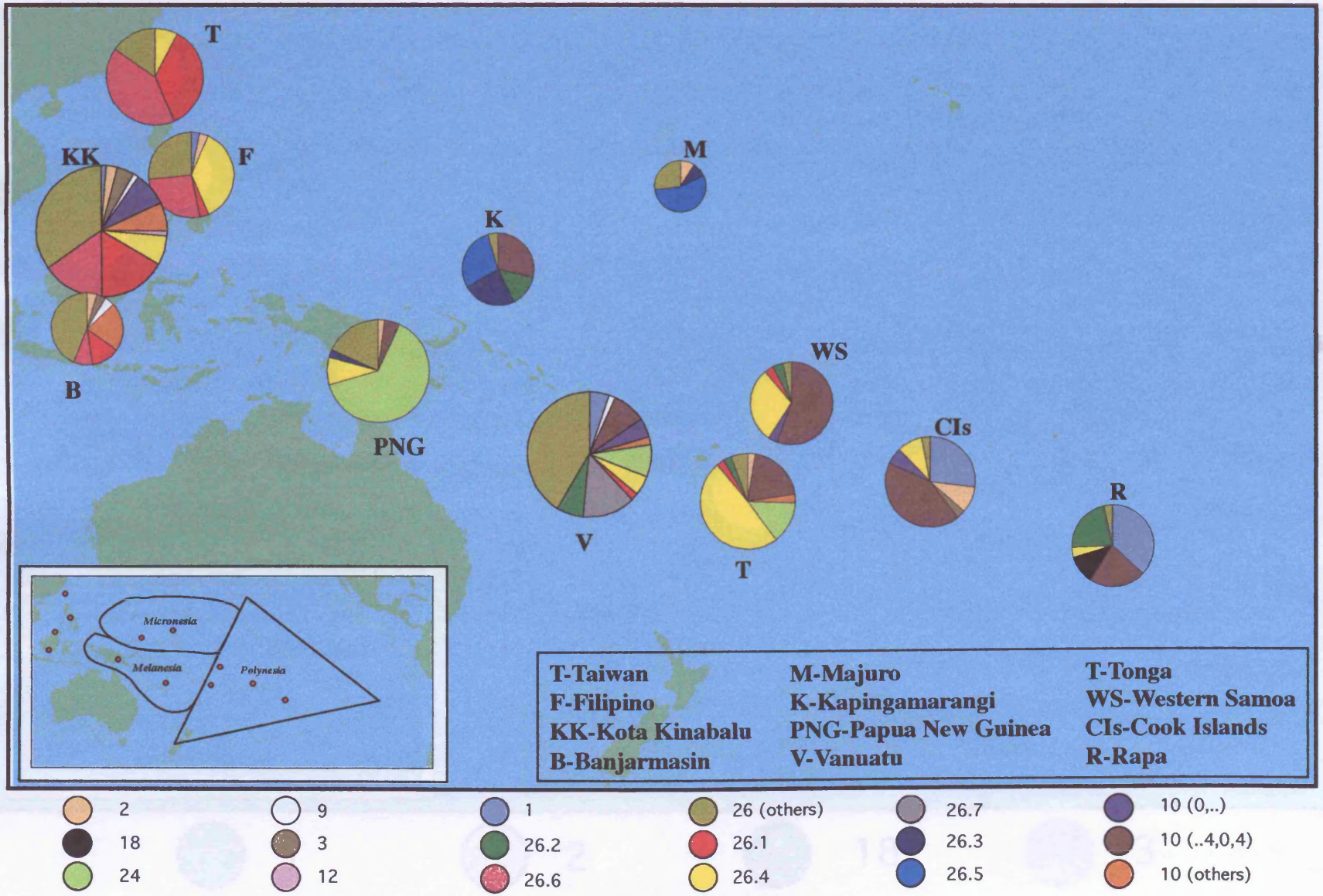


Figure 6.8: Map showing the distribution of lineages in each of the twelve populations sampled here. Circle area is proportional to sample size.

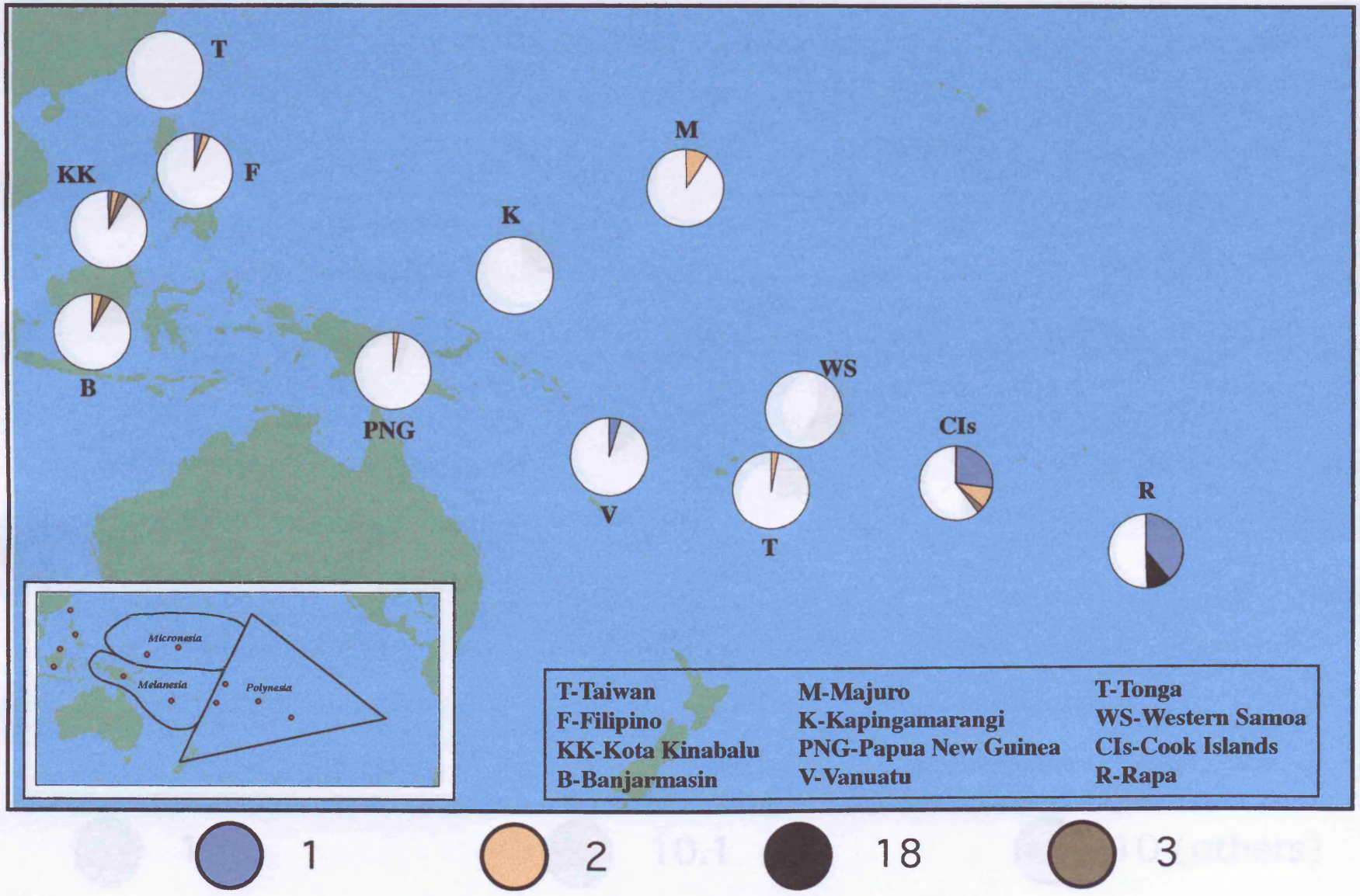


Figure 6.9: Map indicating the distribution of lineages that might constitute admixture.

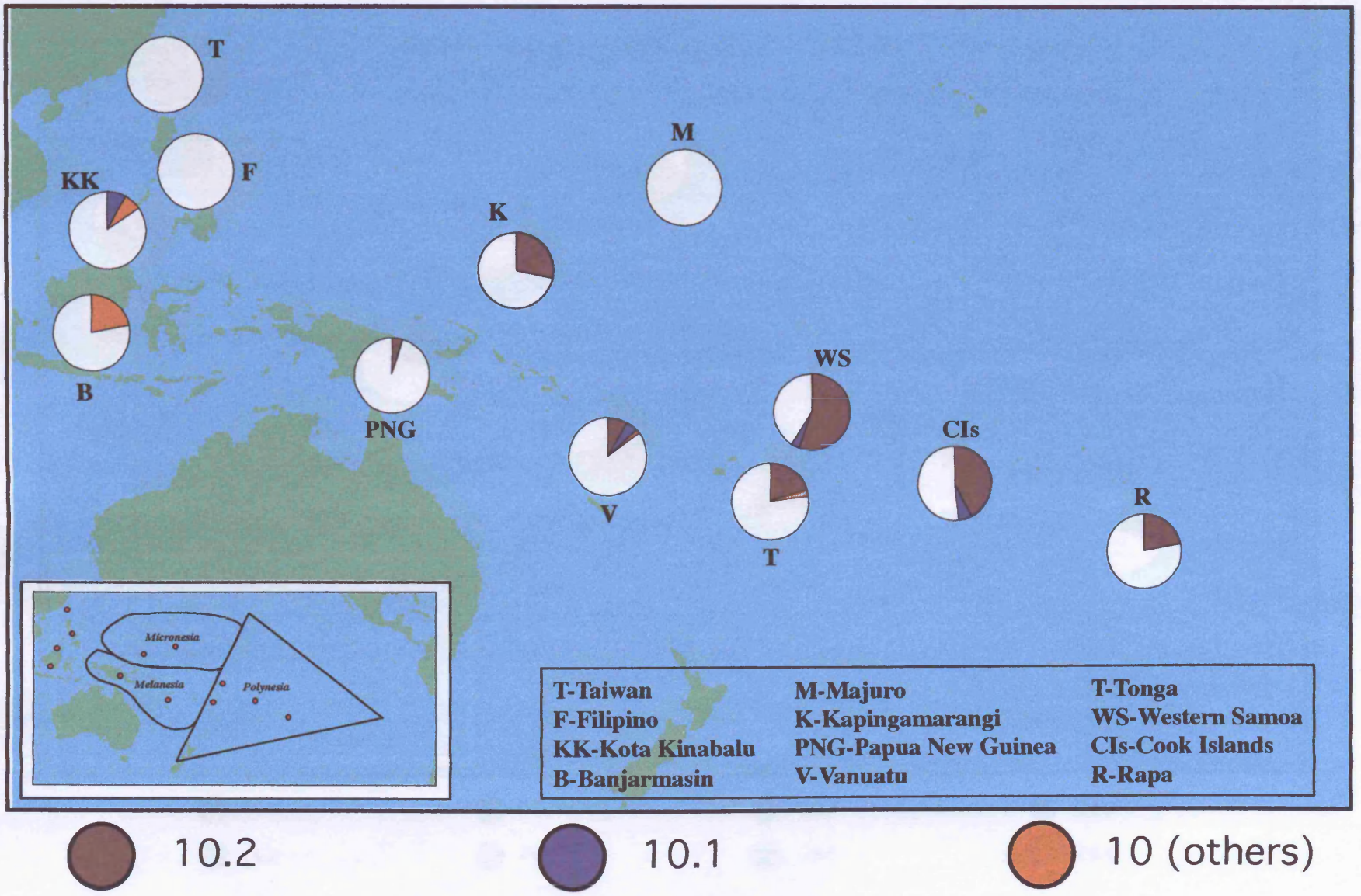


Figure 6.10: Map showing the distribution of lineages belonging to haplogroup 10

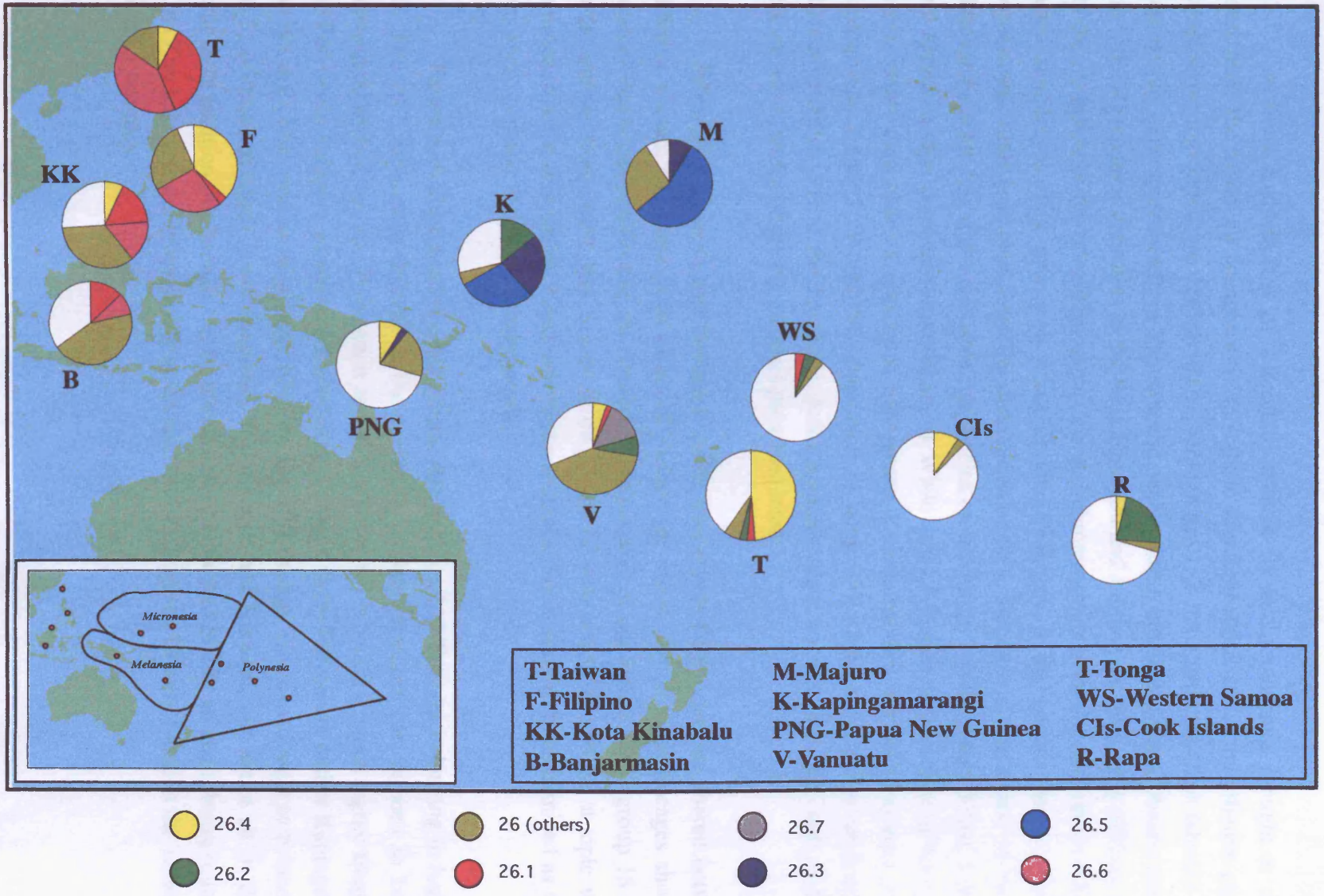


Figure 6.11: Map showing the distribution of lineages belonging to haplogroup 26

diverse chromosomes. Haplogroup 2 is the most ancestral Y-chromosomal lineage defined in this study.

In future analyses it is of interest to exclude those chromosomes thought to represent admixture. The following lineages were thought to represent admixture in the different population samples: haplogroups 1 and 18 from Rapa, haplogroups 1, 2 and 3 from the Cook Islander sample, haplogroup 1 chromosomes from Vanuatu, and haplogroup 1 and haplogroup 2 chromosomes with the (3,1,3,4) modular structures in the Kota Kinabalu and Filipino samples. It is difficult to argue for the exclusion of the other haplogroup 2 and 3 chromosomes from the more diverse Southeast Asian samples as they may well occur at low frequency in indigenous people from these populations. Little is known about the distribution of these lineages in this region of the world, although both are found at reasonable frequencies on the Indian sub-continent (Chris Tyler-Smith and Arpita Pandya, personal communication). While the (1,3,4) modular structure of haplogroup 1 is reasonably diagnostic of non-Asian haplogroup 1 chromosomes, no such evidence exists for haplogroups 2 and 3. Whilst this denotation of admixture might seem rather arbitrary, using alternative definitions in the Southeast Asian samples (data not shown) does not substantially change the inferences from the analyses presented later.

Whereas lineages of true Southeast Asian ancestry in Polynesian are shared between the different islands, and some populations of Melanesian origin, the admixed lineages show much greater heterogeneity, reflecting the island-specific histories of admixture. Haplogroup 18 is found in the sample from Rapa; this lineage is diagnostic of Amerindian admixture, a topic which be returned to in the discussion. Tonga and Western Samoa do not seem to be as admixed as the other two, more isolated, Polynesian populations.

Figure 6.10 illustrates the geographical distribution of all lineages belonging to haplogroup 10. The 10.2 sublineage defined by the (...4,0,4) modular structure can be seen to be shared between all Melanesian and Polynesian samples, although it is found at much higher frequency in the Polynesian samples. In addition this lineage is present in the Polynesian outlier Kapingamarangi but not in its Micronesian neighbour, Majuro. The 10.1 lineage, which is thought to be ancestral to 10.2, is found at much lower frequency in Polynesia and Vanuatu, but not at all in the PNG sample. In addition this lineage can be traced further back into Southeast Asia, being found in the Kota Kinabalu sample. Interestingly no examples of haplogroup 10 were found in the Taiwanese or Filipino samples.

Figure 6.11 illustrates the geographical distribution of the various haplogroup 26 sublineages, but not the remainder of chromosomes belonging to this haplogroup that are not classified into sublineages. Lineage 26.1 is found mainly in Southeast Asia with little penetration into Oceania although individual examples are found in Vanuatu, Tonga and Western Samoa. Lineage 26.2 is found in all Polynesian populations excepting the Cook Islands sample but including the outlier, Kapingmarangi. It is also found in Vanuatu. Lineage 26.3 is shared between the two geographically Micronesian populations and the PNG sample. Lineage 26.4 is found in most populations, but at its highest frequencies (>10%) in Taiwan, Filipino, Tongan and Cook Islands populations. Lineage 26.5 is found only in the two geographically Micronesian populations, where it has attained high frequency in both. Lineage 26.6 is found solely in Southeast Asian populations, shared at reasonable frequencies by the Taiwanese, Filipino and Kota Kinabalu samples. Lineage 26.7 is found solely within Vanuatu, within both of the samples from this group of islands.

All of the non-admixed monophyletic sublineages were dated using the weighting for repeat block size used previously and the three different methods for lineage dating introduced in chapter 4. The results of this analysis are summarised in table 6.3.

Lineage	Prop. mutants (95% CI)	Variance (95%CI)	ASD (95% CI)
26.1	**1679** (916-10076)	**4231** (2587-46377)	**4584** (2500-27503)
26.2	**611** (333-3667)	**1156** (755-9552)	**1278** (697-7665)
26.3	**655** (357-3929)	**1216** (793-10107)	**1250** (682-7500)
26.4	**2076** (1132-12455)	**5707** (2661-67335)	**6318** (3446-37910)
26.5	**1319** (720-7917)	**3296** (1588-31629)	**3055** (1666-18330)
26.6	**1599** (872-9595)	**3587** (1721-35247)	**3795** (2070-22770)
26.7	**313** (170-1875)	**343** (174-2585)	**313** (171-1880)
24	**2684** (1464-16103)	**9771** (4350-195833)	**10102** (5510-60610)
10.2a	**1003** (547-6019)	**1929** (650-16063)	**2190** (1195-13140)
10.2b	**1427** (779-8565)	**4057** (1156-38130)	**3911** (2133-23468)
10.1	**2635** (1437-15808)	**10272** (2785-170934	**12020** (6556-72118)

Table 6.3: Table showing the ages (in years) of the different Southeast Asian lineages found in this study.

As noted in chapter 4 whereas the ages produced by the ASD and Variance methods are in good agreement, the mean number of mutations methods consistently underestimates the other two

methods by a variable amount. It appears that the older the lineage, the greater the factor of underestimation, and thus there is no simple linear correction factor between the methods.

There is a striking correlation between the geographical distribution of lineages and their age, with one notable exception. Haplogroup 24 is the second oldest of all lineages (the inclusion of examples from samples other than PNG has increased the age estimate from that obtained in the preliminary study), yet it is only found in three populations. Compare that with the younger lineage 26.4, the next oldest, which is found in all but three of the twelve samples analysed here. Note also that this is the only major haplogroup 26 sublineage found in Polynesia at appreciable frequencies (>23%). The next oldest lineage, 26.1, is only found at low frequency (<6%) in Polynesia, whereas the only other haplogroup 26 lineage found in Polynesia, 26.2, would appear to have had a recent Oceanic origin. Two lineages found with a restricted geographical distribution within the Oceanic samples, 26.3 (Micronesia + PNG) and 26.7 (Vanuatu), also seem to be very young and are likely to have had origins in the regions in which they are found. The Micronesian-specific lineage 26.5 is surprisingly old given its limited distribution. The 26.6 lineage is the fourth oldest, yet is not found at all in Oceania.

The 10.2 lineage which constitutes the highest frequency lineage within Polynesia has been dated in two separate methods; the first (10.2a) only takes into account chromosomes with the modular structure (0,1,3,4,0,4) and gives a relatively young age (younger than the age of the first colonisation of Polynesia), whereas the second (10.2b) method includes those chromosomes with different modular structures: (0,3,1,3,4,0,4), (1,3,4,0,4) and (3,1,3,4,0,4) by considering the longest modular structure of which all others could be considered a structural (but not evolutionary) subset and counting absence of a repeat block as zero repeats. This gives an age more in keeping with the known history of colonisation. The 10.1 lineage appears to be relatively old despite it's limited distribution, but it is not known whether this age has been biased by the over-representation of lineage 10.2 chromosomes within it. Figure 6.12 shows a network of lineages 10.2 chromosomes with the modular structure (0,1,3,4,0,4).

The dates and topologies of population splits could have been, but were not calculated from the MSY1 diversity within each sublineage. The small numbers of chromosomes belonging to each lineage within each sample and the possible effect of saltatory mutations at MSY1 would make any age estimates prone to large stochastic variation. Microsatellite data would be more applicable to this end due to a lower rate of saltatory mutation (see chapter 4). However it is clear from the networks of MSY1 diversity within lineages in figures 6.5, 6.6 and 6.12 that there is substantial population structure within these lineages - many identical codes are found within individual

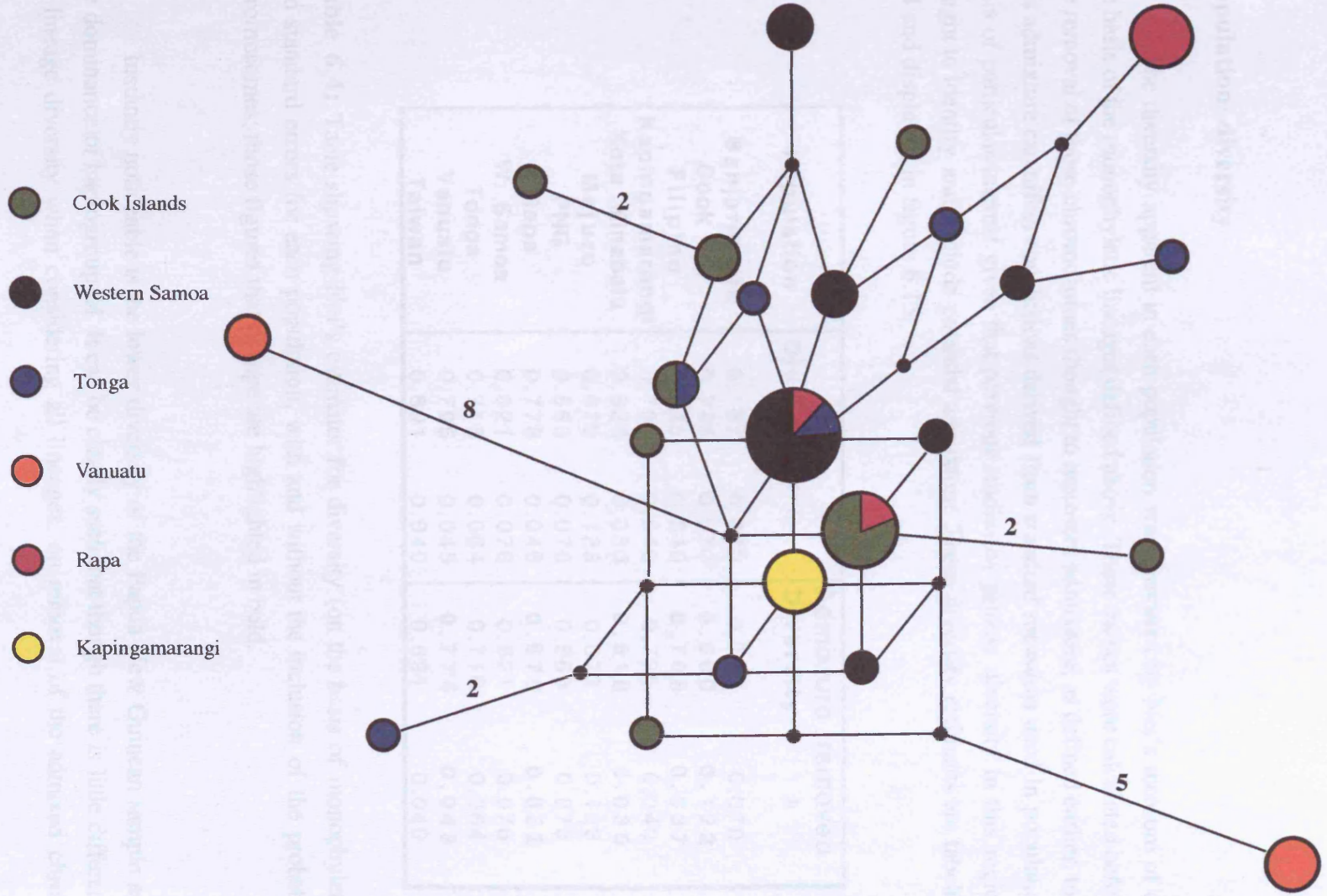


Figure 6.12: Median-Joining network of chromosomes belonging to lineage 10.2 with the MSY1 modular structure (0,1,3,4,0,4). Circles are individual MSY1 codes coloured according to theirpopulation affiliation. All branch lengths are of a single mutational step, except where indicated.

populations and few are shared between populations. For example it can clearly be seen in figure 6.12 that the Vanuatu examples of lineage 10.2 chromosomes are outliers compared to those in Polynesia.

Population diversity

The diversity apparent in each population was estimated by Nei's measure of diversity on the basis of the monophyletic lineages defined above. These values were calculated before and after the removal of those chromosomes thought to represent admixture, as defined earlier, to reveal how this admixture can affect conclusions derived from standard measures used in population genetics. This of particular interest given that previous studies of genetic diversity in this region have not sought to identify and exclude potential admixture. These diversity estimates are tabulated in table 6.4 and displayed in figure 6.13.

Population	Diversity	±	Admixture removed Diversity	±
Banjarmasin	0.767	0.070	0.767	0.070
Cook Isl.	0.746	0.070	**0.500**	**0.122**
Filipino	0.745	0.040	**0.706**	**0.037**
Kapingamarangi	0.795	0.040	0.795	0.040
Kota Kinabalu	0.820	0.030	**0.810**	**0.030**
Majuro	0.673	0.123	0.673	0.123
PNG	0.563	0.076	0.563	0.076
Rapa	0.778	0.046	**0.670**	**0.082**
W. Samoa	0.621	0.076	0.621	0.076
Tonga	0.718	0.064	0.718	0.064
Vanuatu	0.795	0.045	**0.774**	**0.048**
Taiwan	0.691	0.040	0.691	0.040

Table 6.4: Table showing Nei's estimator for diversity (on the basis of monophyletic lineages) and standard errors for each population, with and without the inclusion of the probable admixed chromosomes; those figures that change are highlighted in bold.

Instantly noticeable is the lower diversity of the Papua New Guinean sample as a result of the dominance of haplogroup 24. It can be clearly seen that though there is little difference in terms of lineage diversity when considering all lineages, on removal of the admixed chromosomes a

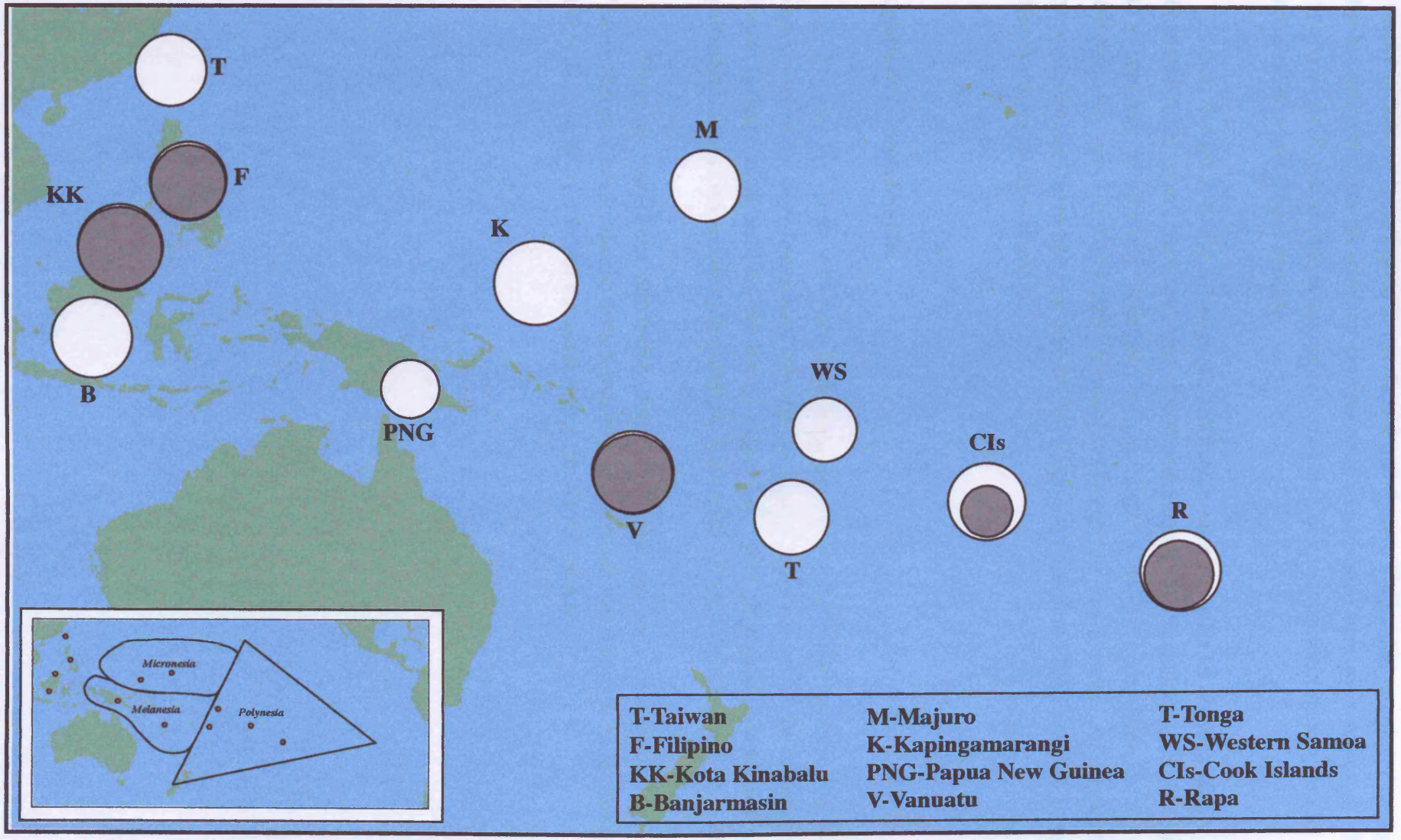


Figure 6.13: Map showing the diversity in each population before (white) and after (grey) admixed chromosomes were removed, circle area is proportional to Nei's esimator of diversity given in table 6.4.

diversity deficiency is noticeable in the Central and Eastern Polynesian samples. Consequently in all subsequent analyses the admixed chromosomes as identified here were removed. The higher diversity of the Rapan sample compared to that of the Cook Islands is surprising given its relatively later settlement and greater isolation.

The lineage diversity in the pooled Taiwanese samples belies an obvious lack of diversity within each tribal group: there are many shared codes within these small samples. The figures for Nei's estimator of diversity calculated on the basis of MSY1 code diversity are: Atayal - 0.0905±0.103; Ami - 0.978±0.035 ; Bunumi - 0.786±0.151 and Paiwan - 0.978±0.054. These diversity estimates do not take into account that many of the different MSY1 codes in these samples differ by only a single repeat unit.

Population comparisons

Pairwise F_{ST} values were calculated between all pairs of populations in two ways; the first solely considers the allele frequencies whereas the second takes account of the mutational distance between lineages as defined by the tree in figure 6.6. 1000 permutations were used to test the significance of these interpopulation distances. The results of both of these are given in table 6.5.

There are only three population comparisons that are not significant in both analyses; Western Samoa v Cook Islands, Kota Kinabalu v Banjarmasin and Majuro v Kapingamarangi. The latter is likely to be an effect of the small size of the Majuro sample.

a

	Banjarmasin	Cook Isl.	Filipino	Kapingamarangi	Kota Kinabalu	Majuro	PNG	Rapa	W. Samoa	Tonga	Vanuatu
Banjarmasin											
Cook Isl.	0.34845										
Filipino	0.11641	0.32161									
Kapingamarangi	0.20266	0.18681	0.22139								
Kota Kinabalu	*0.00133*	0.2837	0.06016	0.17718							
Majuro	0.17407	0.41907	0.22703	0.08801	0.15817						
PNG	0.289	0.43148	0.29339	0.31765	0.24685	0.36243					
Rapa	0.25409	0.15251	0.25399	0.09533	0.21689	0.3152	0.3709				
W. Samoa	0.29339	*-0.00041*	0.22307	0.15416	0.23416	0.35151	0.37376	0.10825			
Tonga	0.2316	0.21297	0.09156	0.19462	0.17705	0.28731	0.24481	0.19276	0.09502		
Vanuatu	0.04078	0.26985	0.11938	0.16914	0.04605	0.16889	0.21343	0.18403	0.22722	0.18286	
Taiwan	0.14554	0.37905	0.1138	0.25647	0.07459	0.28625	0.35164	0.30604	0.31294	0.25382	0.20396

b

	Banjarmasin	Cook Isl.	Filipino	Kapingamarangi	Kota Kinabalu	Majuro	PNG	Rapa	W. Samoa	Tonga	Vanuatu
Banjarmasin											
Cook Isl.	0.51703										
Filipino	0.16848	0.64439									
Kapingamarangi	0.1114	0.31384	0.25339								
Kota Kinabalu	*-0.00704*	0.51891	0.08754	0.1402							
Majuro	0.20972	0.63431	0.30309	*0.10124*	0.17277						
PNG	0.2892	0.65829	0.32899	0.30878	0.24499	0.38648					
Rapa	0.21489	0.20265	0.3753	*0.04832*	0.24583	0.33707	0.42446				
W. Samoa	0.30485	*0.03345*	0.42651	0.15219	0.3371	0.41699	0.48118	*0.05605*			
Tonga	0.12925	0.38056	0.1196	0.11857	0.12499	0.23407	0.22755	0.14294	0.17151		
Vanuatu	*0.034*	0.47532	0.13913	0.0872	0.0401	0.16611	0.18808	0.16911	0.28446	0.08911	
Taiwan	0.17598	0.6691	0.14065	0.30072	0.09154	0.34101	0.38476	0.42568	0.48691	0.27009	0.20273

Table 6.5: Table giving the pairwise F_{ST} between different populations (a) not taking mutational distance into account and (b) taking mutational distance into account. Italics indicates a lack of significance at the 5% level.

Principal Components Analysis can be used on the lineage frequency data to construct a graphical picture of the diversity apparent within the sample. Three-dimensional representations of PCA are often difficult to interpret accurately, consequently only two-dimensional representations were used. The first two dimensions only account for less than 60% of the overall diversity therefore the first three dimensions are represented in two 2D representations. PCA can be performed in one of two ways; by converting the input data to a covariance matrix or considering it as a correlation matrix. Although the practical implications are unclear, the two approaches differ in how they scale the data from the different lineages. Figure 6.14 shows the result of this analysis.

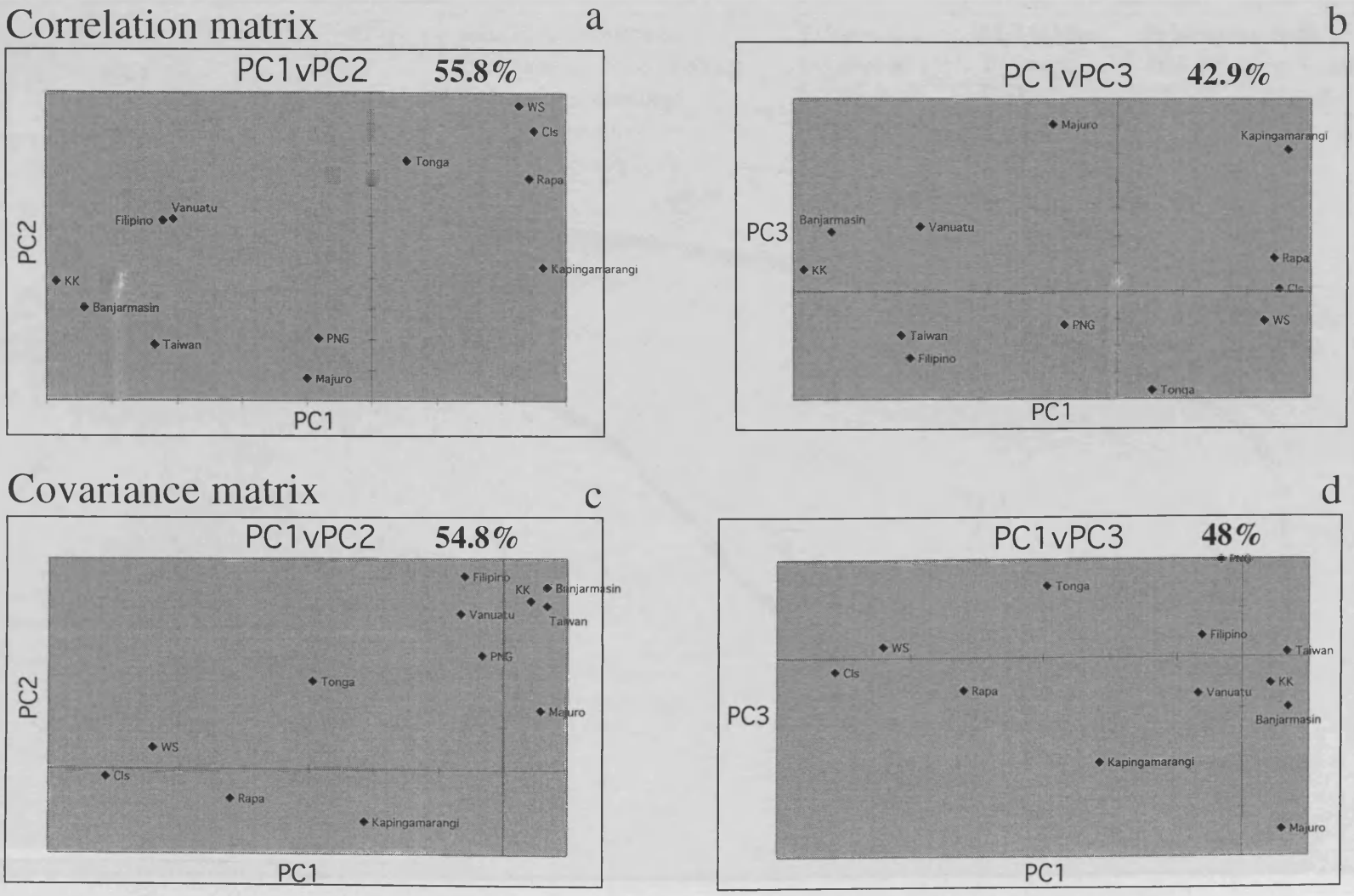


Figure 6.14: Graphs showing the results of Principal Components Analysis (a) PC1 v PC2 correlation (b) PC1 v PC3 correlation (c) PC1 v PC2 covariance (d) PC1 v PC3 covariance. Figures next to each plot indicate the amount of variance summarised in those two PCs. KK- Kota Kinabalu; WS - Wstern Samoa; CIs -Cook Islands; PNG - Papua New Guinea.

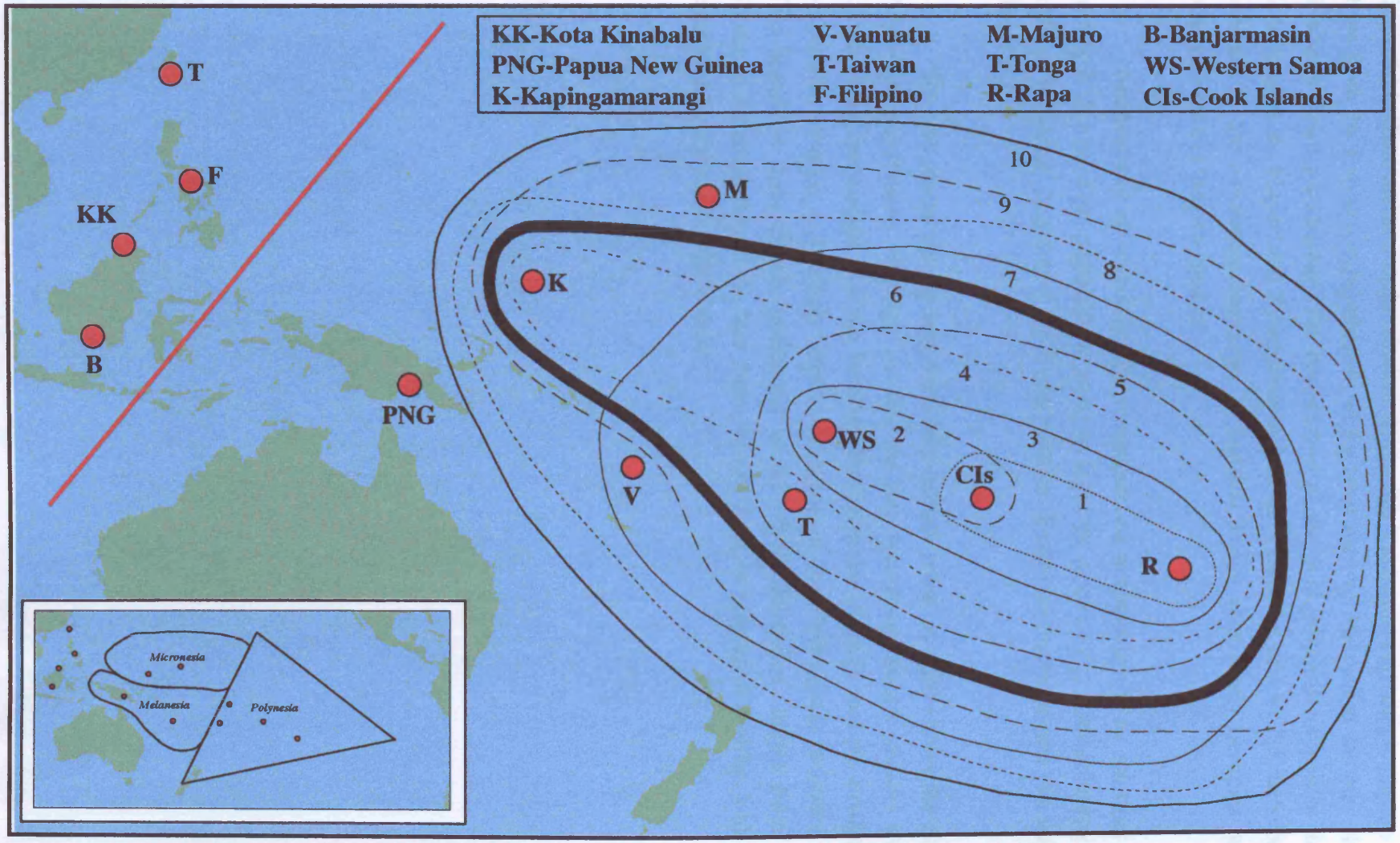


Figure 6.17: Map indicating the different groupings considered in the AMOVA classification of Oceania. The bold line indicates the grouping best supported by this analysis. The red line separates the Southeast Asian samples not included in this analysis from the Oceanic islands that were.

The striking conclusions from this analysis are common to both the approaches to PCA. In figure 6.14a there are two major clusters; in one the Kapingmarangi sample clusters together with those from the Polynesian populations, in the other the two Melanesian samples cluster with the Southeast Asian populations. The introduction of the third principal component in figure 6.14b distinguishes the Majuro and Kapingamarangi samples from the others. The same conclusions are apparent in the covariance approach. In addition Tonga consistently occupies an intermediate position between the two clusters.

Another way of visualising the interpopulation relationships is to construct a Neighbour-Joining (NJ) tree of the populations, using as input the pairwise F_{ST} distance matrix shown above and the 'Neighbor' program of the Phylip package. A tree constructed in just this way is shown in figure 6.15.

The tree makes clear geographical and cultural sense. It again shows the clustering of the Polynesian populations, although this time the Kapingamarangi sample clusters with the Majuro sample in an intermediate position between the Polynesian samples and the Southeast Asian and Melanesian samples. A strikingly long branch is seen to the PNG sample indicating its distinction. This branch is most closely positioned to the other Melanesian sample from Vanuatu. The Southeast Asian portion of the tree traces its way back in geographical order, with the Taiwanese sample being the most distal branch.

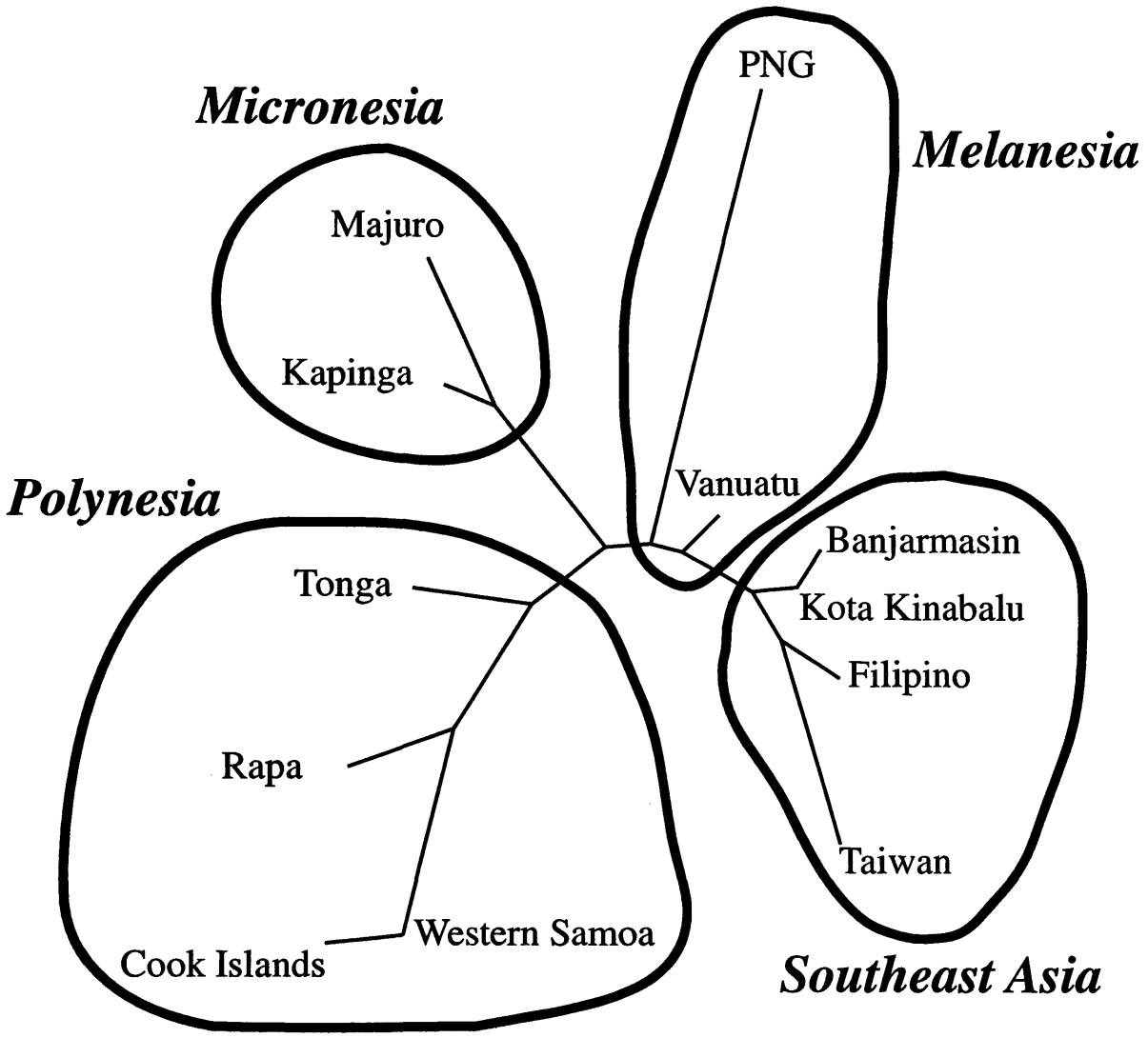


Figure 6.15: Neighbour-Joining tree of the 12 populations, showing the regional affiliations, constructed using Phylip.

Landscape analyses

As well as making comparisons between individual populations we can also use the dataset as a whole to make inferences on the properties of the entire genetic landscape studied here. One such analytical technique seeks to identify the nature of the relationships between geographical, genetic and linguistic differences. Towards this end Mantel tests of partial correlations were applied to different subsets of the data. These subsets can be taken to represent different 'spheres of influence'. The paper by Lum et al. (1998) used four different 'spheres of influence'; however,

with fewer samples I have decided to only consider two: they are (a) all Austronesian-speaking populations (all those here) and (b) all populations speaking languages belonging to the Oceanic subgroup of Austronesian.

Mantel tests require distance matrices for all three forms of relationships between populations. F_{ST} measures were used for genetic distance. I wrote a short program in IDL to calculate the geographical distances between all populations taking into account the Earth's curvature by using Great Circle Distances. The resulting geographical distance matrix is given in Table 6.6. The program source code is detailed in appendix D.

	Borneo	Cook Isl	Filipino	Kapinga	Majuro	PNG	Rapa	W.Samoa	Tonga	Vanuatu
Cook Isl	5929									
Filipino	1051	5921								
Kapinga	2766	3431	2501							
Majuro	3948	2688	3515	1213						
PNG	2310	3623	2435	911	2011					
Rapa	6934	1077	6993	4508	3724	4653				
W.Samoa	5103	944	5003	2503	1766	2797	2017			
Tonga	4954	975	4979	2520	2004	2649	2019	459		
Vanuatu	3803	2126	3860	1501	1546	1497	3162	1343	1151	
Taiwan	1658	6115	625	2701	3541	2839	7191	5174	5221	4159

Table 6.6: Matrix of geographical distances (in km) between populations

Linguistic distances are less straightforward, but were calculated on the basis of a tree of the Austronesian languages spoken by the populations studied here. The matrix of linguistic differences is given in Table 6.7. The language tree presented in figure 6.16 is based on that produced in the Lum et al. (1998) paper which was constructed by the authors in consultation with linguists in the field. Not all of the languages considered here were considered in that tree and so additional information was gleaned from the trees presented in figure 6.2. Most notably this included the Taiwanese samples which were pooled and considered to be the most distinct from all other Austronesian languages in keeping with the consensus of the Formosan languages splitting first in the Austronesian phylogeny. Where some doubt remained about the relationships between groups of languages the branching orders are not resolved, as can be seen to be the case for the subgroups of the Oceanic group of languages.

Information was not available on the languages spoken in the two Borneo samples. However, both are almost certain to belong to Borneo branch of Western Malayo-Polynesian group of languages. Filipino languages also belong to this group of languages.

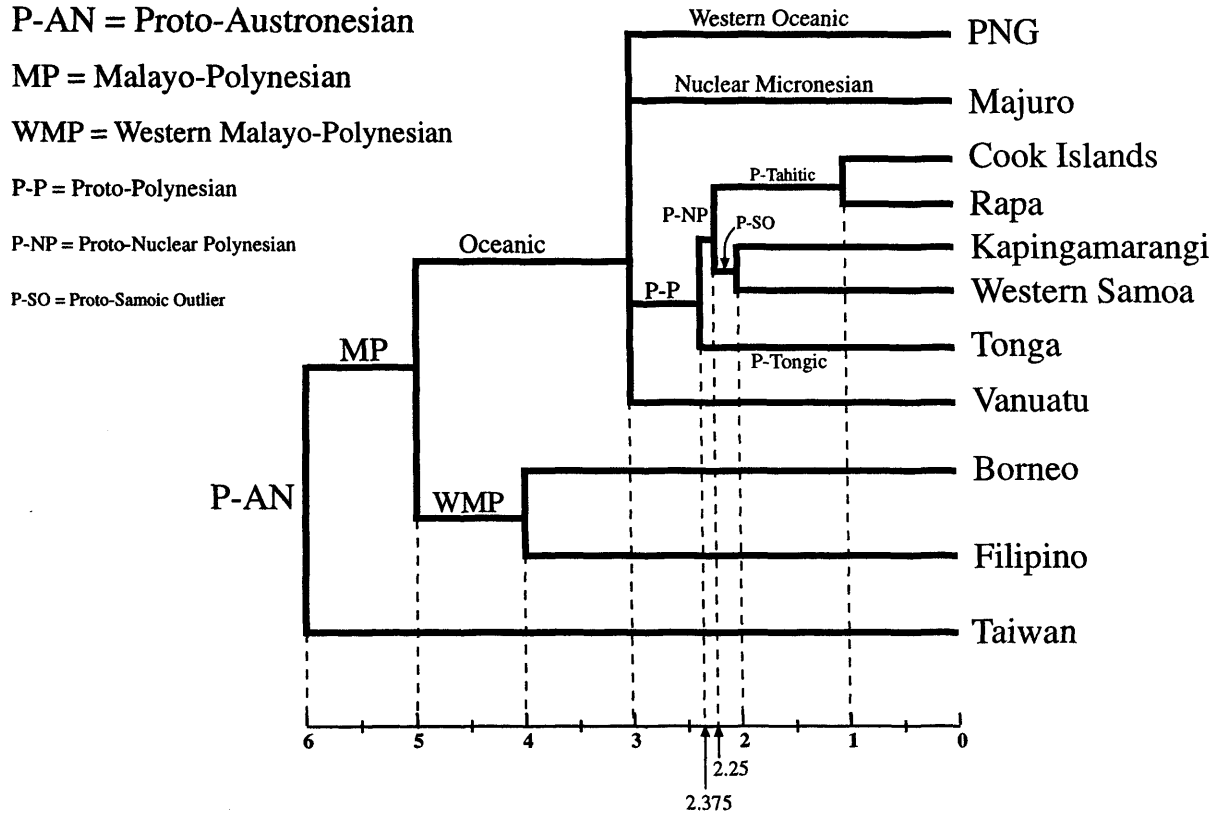


Figure 6.16: Linguistic tree relating 11 Austronesian linguistic groups, after Lum et al., 1998 (see text).

	Borneo	Cook Isl	Filipino	Kapinga	Majuro	PNG	Rapa	W.Samoa	Tonga	Vanuatu
Borneo										
Cook Isl	10									
Filipino	8	10								
Kapinga	10	4.5	10							
Majuro	10	6	10	6						
PNG	10	6	10	6	6					
Rapa	10	2	10	4.5	6	6				
W.Samoa	10	4.5	10	4	6	6	4.5			
Tonga	10	4.75	10	4.75	6	6	4.75	4.75		
Vanuatu	10	6	10	6	6	6	6	6	6	
Taiwan	12	12	12	12	12	12	12	12	12	12

Table 6.7: Matrix of linguistic differences between populations

The results of the Mantel tests on the two 'spheres of influence' are given in table 6.8. The significance of correlations and partial correlations is determined by permutation testing.

Pairwise and partial correlations	All Austronesians		All Oceanic Austronesians	
	r	p value	r	p value
Ychr v geography	**0.31**	**0.026**	0.38	0.06
Ychr v language	0.13	0.28	**0.53**	**0.011**
Ychr v geography (language)	**0.29**	**0.044**	0.37	0.078
Ychr v language (geography)	-0.05	0.544	**0.52**	**0.011**

Table 6.8: Results of Mantel tests on the two groupings, correlations significant at the 5% level are shown in bold. Where only two matrices are named the correlation coefficient (r) refers to a pairwise correlation between them. Where a third matrix is mentioned in brackets r refers to a partial coefficient of the first two matrices given the third. Significance is determined by permutation testing.

When all populations are considered genetic and geographical distances are significant at the 5% level even when linguistic distances are taken into account. However linguistic distances are not significantly correlated with genetic distances.

In contrast when only considering the populations speaking Oceanic languages there is a clear significant correlation between genetics and linguistic distances at the 5% level even when geographical distances are accounted for. When geographical and genetic distances are considered on their own they only just fail to be significant at the 5% level; however, when linguistic distances are taken into account they are more clearly non-significant.

In an attempt to consider the best way to subdivide the Oceanic samples on the basis of genetic diversity into two groupings, AMOVA calculations were performed on different groupings of samples. AMOVA apportions the diversity within a landscape at three different hierarchical levels; within populations, between populations within a group and between these groups. The grouping that maximises the latter variance can be considered to subdivide the samples most completely whereas the grouping that minimises variance due to the middle level of the hierarchy can be thought of as splitting the samples into the two most homogenous groups. Though not independent these two criteria subdivide the landscape into two regions by focusing on different considerations.

Ten different groupings were calculated. Although these do not represent all possible groupings of the eight Oceanic samples, they are the only ones which make any geographical or cultural sense. The results of the AMOVA are presented in Table 6.9 and the different groupings are shown geographically in figure 6.17.

Grouping	% Variance within pop	% Variance Among Pop	% Variance Among group	p value
1	73.78	21.23	4.99	0.164
2	72.01	18.48	**9.51**	**0.035**
3	72.16	17.34	**10.5**	**0.020**
4	72.91	16.87	**10.23**	**0.030**
5	73.35	17.04	9.6	0.063
6	73.17	16.32	**10.5**	**0.040**
7	75.29	21.51	3.2	0.220
8	73.27	19.37	7.37	0.074
9	73.6	17.48	**8.92**	**0.047**
10	71.37	18.08	10.55	0.120

Table 6.9: Results of AMOVA analyses of different groupings. Values for among group variance significant at the 5% level are highlighted in bold.

Two groupings produce the same, highest, value for significant between group variance; one of these also produces the lowest between population variance and so can be considered to be the best supported on the basis of genetic diversity: this is grouping 6 in the table and the figure below, and groups together the samples from Rapa, Cook Isl., Western Samoa, Tonga and Kapingamarangi.

Discussion

This study illustrates the power of the methodological approach for defining lineages introduced in chapter 4. Comparing figures 6.4 and 6.7 indicates the vastly increased resolution obtained. Many of the conclusions obtained here are contingent upon this greater resolution. The present resolution of biallelic markers alone would not be sufficient to make most of these anthropological inferences.

The lineage definitions make geographical sense in that they have comprehensibly coherent distributions throughout the region investigated here. In addition the ages of lineages show close congruence with the extent of their geographical distributions. Neither of these concordances would be expected if these lineage definitions were in fact analytical contrivances. Further evidence for the resolving power and overall veracity of this analysis is the fact that the Neighbour-Joining tree reconstructs the four main groupings of Polynesia, Melanesia, Micronesia and Island Southeast Asia.

Despite their wide confidence limits the ages for lineages obtained fit nicely with age estimates of population movements within Oceania and thus supports the 'express train' model of colonisation. Only the oldest Southeast Asian lineage is found at high frequency in Polynesia, while all others are poorly represented and have a proposed origin within the past 4600 years. This oldest Southeast Asian lineage, 26.4, has an origin roughly 6000 years ago, the time at which the Austronesian expansion was thought to have begun. The oldest lineage in the dataset, haplogroup 24, was thought in the preliminary study to have had an origin in Papua New Guinea itself (Hurles et al. 1998). The earlier date for the origin of this lineage obtained here makes it more evident that the lack of spread of this lineage throughout Oceania must be due to minimal contact of the colonising populations with the ancestral populations of this island. This conclusion is supported by the position of PNG within the Neighbour-Joining tree at the end of a long branch from the other populations (figure 6.15).

The delineation of geographical and temporal origins of lineages found in Oceania might reveal the centres of dispersals of which linguists are fond (Pawley and Ross 1993). Previous genetic literature has proposed two dispersal centres, one for Proto-Austronesians in Taiwan (Melton et al. 1998) and the other for Proto-Polynesians in East Indonesia (Richards et al. 1998). These need not be mutually exclusive but operate at different levels of a linguistic and therefore temporal hierarchy (Pawley and Ross 1993). The two major Polynesian lineages that are also found

outside of Island Oceania, 26.4 and 10.2, support both of these hypotheses. Interestingly their ages fit with this temporal hierarchy. Lineage 10.2 is found furthest West in Papua New Guinea but its putative ancestral lineage 10.1 is found only in one site in Island Southeast Asia in the sample from Kota Kinabalu on the north coast of Borneo. The best estimate of the age of this lineage is roughly 4000 years ago. Lineage 26.4 is older, of the order of 6000 years old, and can be traced back through Southeast Asia as far as Taiwanese populations. Proving that the homeland of this lineage is indeed in Taiwan is more difficult; MSY1 diversity reveals that two of these four Taiwanese tribes show reduced diversity consistant with the idea that small tribal social organisations are prone to very low effective population sizes for paternal lineages that result in the elimination of much ancestral diversity (Jobling et al. 1998). This may also explain the absence of haplogroup 10 from these Taiwanese populations. Nei's estimator of diversity does not take into account the proximity of codes and so does not reveal that all of these tribes show reduced diversity.

Haplogroup 10 can be reasonably expected to have an origin that predates the Austronesian expansion; it is shared between populations on both sides of the Bering Strait and thus is likely to be at least 12000 years old, although coalescent dating gives it an even earlier origin (Karafet et al. 1999). Although the Taiwanese populations sampled here did not have this lineage another survey of this lineage in populations from this island did find it at a reasonable frequency (Karafet et al. 1999).

The power of adopting a genealogical approach is amply verified by the ability to resolve admixture within these samples. The European admixture in the Cook Islander sample reported in the preliminary study has not been duplicated to the same degree in other populations. However more intriguing is the discovery of the haplogroup 18 chromosomes in the Rapa sample. Haplogroup 18 is incontrovertibly of Amerindian origin (Karafet et al. 1997). This raises the spectre of Thor Heyerdahl's hypothesis of Amerindian origins for Polynesians (Heyerdahl 1950). However a more parsimonious hypothesis is at hand. By the middle of the nineteenth century the population of Rapa numbered 360 (Maude 1981). In 1863 the Rapan islanders captured a ship participating in the pernicious Peruvian labour trade that lasted for a mere seven months yet resulted in the deaths of some 6000 Polynesians (Maude 1981). The mainly Chilean crew of the captured ship remained on Rapa. The next year a Peruvian ship repatriated 15 enslaved Polynesians previously taken from different islands within Polynesia, from Peru to Rapa, bringing with it smallpox and dysentery. Two thirds of the Rapan population died during the resulting epidemic and it is documented that at one stage only 20 males survived (Maude 1981). This temporal coincidence of a tight population bottleneck and the presence of Amerindian sailors seems a far more plausible explanation for the origin of these Amerindian chromosomes than Heyerdahl's. It might be that the

haplogroup 1 chromosomes in the Rapan sample have two origins - from Amerindians and Europeans. Europeans are likely to have constituted the remainder of the non-Chilean crew. Despite the ships sailing under Peruvian flags is was not unusual for there to be no Peruvians on board (Maude 1981). A network of the haplogroup 1 chromosomes found in this study reveals that the Rapan examples fall into two groups: one is subsumed within the other haplogroup 1 chromosomes of European origin, whereas the other is present as outliers and may represent Amerindian chromosomes. The analysis of microsatellite data on these samples will help to resolve this presently tenuous hypothesis, as would the typing of a potential Amerindian source population. Furthermore the repatriation of diverse Polynesians and their documented taking of Rapan wives may well explain the surprisingly high diversity of the Polynesian lineages within Rapa compared to the Cook Islands.

Clearly, islands have radically different amounts and origins of admixture, dependent on their individual histories and demographies. Extrapolation of conclusions on admixture from one island to another should be avoided. Consequently the removal of these admixed chromosomes from subsequent analyses is vital for a faithful investigation of the prehistory of the region. This is well illustrated by the comparisons of intra-population diversities in table 6.4. Studies which use microsatellite data alone and do not define lineages can not hope to correct for this admixture and may well bias their conclusions significantly.

The genetic affiliations of Kapingamarangi are particularly well resolved by this analysis. Simple inspection of lineages indicates the sharing of Micronesian-specific lineages with the sample from the Marshall islands (Majuro) in addition to the substantial presence of lineages that are only found at high frequency within Polynesia. Kapingamarangi must represent an admixed population with both Polynesian and Micronesian origins, which would help to explain its surprisingly high diversity compared with most other Oceanic islands. In various of the graphical representations of inter-population comparisons it can be seen to cluster with either the Majuro sample or the other Polynesian samples.

After the removal of admixture it can be seen that there is a substantial reduction in diversity of monophyletic lineages in the more geographically isolated Oceanic islands. The low diversity of the Papua New Guinean sample is due to the dominance of haplogroup 24. MSY1 code diversity within this sample does not indicate the same lack of diversity (data not shown) compared to other populations. By contrast Tonga and Vanuatu do not appear to exhibit this reduction in diversity. These two populations also are the only other ones to contain haplogroup 24. The absence of this lineage from other Polynesian populations probably indicates that its presence in Tonga is due to

post-settlement movements that affected Polynesia minimally. This picture of events is reinforced by the position of Tonga in both the Neighbour-Joining tree and PCA as being intermediate between Polynesian samples and the rest. The Vanuatu sample includes both Melanesian and Polynesian lineages. The latter may have originated from the initial expansion or from subsequent movements back from Polynesia; the fact that the examples of lineage 10.2 from Vanuatu are outliers from the Polynesian examples in the network in figure 6.12 would seem to indicate that the former is more likely to be the case, and provides further evidence that Polynesia was peripheral to the proposed post-colonisation movements in Melanesia.

Lum et al. (1998) proposed extensive male-specific post-colonisation contacts within Oceania and thus a higher rate of male gene flow, from Mantel tests that indicated genetic distances from mtDNA but not autosomal STRs having a significant correlation with linguistic differences within Oceania. They argued that significant partial correlations of genetic distances and linguistic distances are indicative of genetic affinities that are largely determined by the pattern of initial settlement where as a more significant partial correlation of genetic and geographic distances indicates the impact of post-colonisation movements. The Mantel tests here verify the method of inference yet disagree with the inference itself. We can legitimately expect a higher rate of gene flow in Island Southeast Asia than in Oceania, and therefore the finding that Y-chromosomal genetic distances among all samples studied here are better correlated with geographical distances than linguistic distances is unsurprising. However the finding that genetic distances within Oceania are correlated significantly with languages but not geography emphasises that the pattern of settlement remains the major determinant of genetic affinity amongst the Oceanic populations studied here.

Consequently this analysis gives no support to the concept of higher Oceanic male-specific gene-flow, proposed by Lum et al. What are the possible causes of this discrepancy? Four possible explanations come to mind. Firstly the mutation rate of the markers used might explain the discord. Faster mutating markers such as STRs recover their diversity more rapidly after bottlenecks than do biallelic markers such as those used to construct the mtDNA and Y-chromosomal genetic distances. This recovering of diversity might mask the signal of original settlement. However the time since the events being investigated is so small that even STRs will have little time to recover. Secondly the discrepancy might be a function of the different samples used in each study. The paper by Lum et al. had many more Micronesian samples than Polynesian. Micronesia is known to exhibit far greater inter-island voyaging than Polynesia. However by using different 'spheres of influence' Lum et al. took account of this, and indeed no marker exhibited significant correlations with language when Micronesian populations alone were considered. Significant correlations between

mtDNA and linguistic distances were only found when considering (a) all Austronesian-speaking populations and (b) Melanesian and Polynesian populations. Whilst these findings negate this explanation for the above discrepancy it raises the interesting proposition of higher male gene-flow in Island Southeast Asia. The final two explanations are more likely to account for this discord. The lower effective population sizes of the Y chromosome and mtDNA means that bottlenecks are felt more severely by these loci and settlement patterns are more pronounced. Finally the exclusion of admixture in Polynesian and Melanesian samples may well be the underlying cause of this discrepancy. The 'noise' of admixture that detracts from the 'signal' of original settlement cannot be excluded by microsatellite data treated in a non-genealogical fashion. This 'noise' might well be confused with the 'noise' of post-colonisation contacts.

Any attempt to draw conclusions about sex-specific differences in prehistory should consider data on maternal and paternal lineages in the same samples. Consequently, though the present analysis fails to support previously published hypotheses regarding male-specific gene-flow in Oceania, it does not propose that this higher rate of male gene-flow has not occurred, but that it has not presently been demonstrated. However the final sentence of Lum et al. states "Thus, we see female settlement as an express train and male gene-flow as an entangled bank." This sentiment would seem to be untenable in light of the present analysis.

The comparisons of different groupings of Oceanic populations by AMOVA were intended to reveal the best way to subdivide these populations into two groups, on the basis of the genetic diversity revealed here. It has been suggested that Oceania would be better subclassified into Near and Remote rather than on the cultural definitions of Polynesian, Melanesian and Micronesian (Terrell 1997). However this analysis faithfully reconstitutes the Polynesian cultural grouping of the four Polynesian populations and the Polynesian outlier in Micronesia, Kapingamarangi. Thus this classification supports a genetic underpinning of these cultural definitions.

In summary; though the use of the metaphor of the 'entangled bank' is pretentious, it may well be useful in describing the post-colonisation movements throughout island Melanesia. However the reduction in diversity, clustering in PCA and Neighbour-Joining trees, and correlation with languages of Polynesian Y chromosomes would appear to exclude them from this metaphorical ecosystem. Rather than there being a sharp boundary to these post-colonisation movements a gradient of gene-flow magnitude throughout Oceania is more likely. Drawing a line across this gradient might always prove to be contentious, but here an objective line defined genetically maintains the cultural definitions long-since appreciated in anthropology.

What is there left to be done with paternal lineage analysis in this region of the world?

The paucity of present populations sampled, the small size of the samples and the absence of microsatellite data means that dating population splits as opposed to lineage ages is fraught with uncertainty to such an extent that inferences of anthropological significance cannot be drawn. A more comprehensive sampling and collection of larger samples would help towards this end. Sampling in Island Melanesia would help further investigate the interplay of initial colonisation and post-colonisation inter-island contacts on extant diversity. Sampling from East Indonesia, specifically the Moluccas and Sulawesi, would aid in defining the colonisation routes through this critical region and better define the homeland of lineage 10.2. Sampling from East Polynesia would provide an opportunity to compare genetic, linguistic and archaeological data together to best resolve the contentious migration routes and dates within this region. Sampling of Micronesian populations would allow a complete picture of the genetic diversity within the region to be constructed. Finally sampling various populations from South America will allow a better appreciation of potential Amerindian admixture.

The practical employment of this knowledge to an understanding of disease allele distributions in Oceania might well add a direct practical application to this, historically, mainly intellectual curiosity.

Chapter 7: General discussion

This thesis has been concerned with developing methodologies for Y-chromosomal diversity analysis and making anthropologically-relevant inferences from Y-chromosomal diversity. Chapter 3 emphasised the need for ever greater numbers of biallelic markers and introduced the recently developed mutation detection technique of DHPLC. Whilst the screen for new SNPs using this complex technology was far from optimal, its potential for high throughput SNP screening was demonstrated and a single new polymorphism discovered. Chapter 4 presented the concept of lineage analysis that is critical to the adoption of a genealogical approach to understanding extant Y-chromosomal diversity. This chapter revealed the power of network analysis compared to competing multivariate methods. Tools for the incorporation of multi-allelic diversity into a genealogical framework were introduced and further developed in the analysis of two relatively young Y-chromosomal lineages. It was shown that even the analysis of a single lineage can yield anthropologically-informative conclusions. Chapter 5 presented a departure from lineage-based methods. The application of a permutation test to 'Wombling' analysis in this chapter represents a new way to ascertain significance of genetic barriers. Whereas linguistic boundaries closely correspond to Y-chromosomal genetic barriers there is a relative lack of correspondence with mtDNA (Poloni et al., 1997; L. Simoni personal communication). This reveals the differential effect that cultural practices in prehistory (such as patrilocality) can have on the sex-specific substructuring of extant diversity. Chapter 6 presented the analysis of the major dataset collected during this thesis. The isles of Oceania and their inhabitants have been intensely studied from multiple viewpoints. Despite or even because of this intense interest many contentious issues remain in Oceanic prehistory. As well as demonstrating the congruence of Y-chromosomal diversity with the consensus on Polynesian origins I have attempted to address some of these issues.

What remains to be done in these regional studies? The wider sampling of more Oceanic islands will allow greater resolution of Y-chromosomal analysis. The relative genetic contributions of the original migrational settlement and subsequent inter-island voyaging are of specific interest, underpinning as they do many of the contentious issues alluded to above. In addition greater sampling may provide sufficient information to make a Wombling analysis practical. There now exists a vast corpus of genetic data on Oceanic diversity (Hill et al. 1985; Hill and Serjeantson 1989; Flint et al. 1989; Sykes et al. 1995; Lum and Cann 1998; Lum et al. 1998; Lum et al. 1994;

Hurles et al. 1998), and the collation of this diversity into a coherent database should be a goal of subsequent studies of this region.

In summary it has been revealed that in principle one can learn more about a single lineage on the Y chromosome than at any other human locus. However if these advantages are to be fully realised for regional studies, larger samples of individual populations will have to be analysed so as to cope with their subsequent subdivision.

As far as I know there is no single locus with comparable resolution within the field of ecological genetics. The extension of the methodologies discussed within this thesis to heterogametic chromosomes in other species might prove a powerful weapon within the ecologist's analytical arsenal.

Whilst there is a general consensus within much of the field as to the benefits of adopting a genealogical approach to non-recombining regions of the genome there exists a considerable diversity of analytical techniques used to address any given issue. Some groups (old dogs) exhibit an unwillingness to develop a breadth of techniques (learn new tricks) to the detriment of the field as a whole. Papers are published that could adequately be summarised by the title "Diversity of locus X in population Y using Z analysis". If genetics is to contribute to anthropological studies and be better appreciated by other disciplines within this broad church it must focus on question-driven research, and seek to test hypotheses rather than demonstrate compatibility.

What can be done towards this end?

Combining diversity information from multiple loci is an obvious goal for the near future. Individual alleles and even individual loci tell their own story of human prehistory. Combining loci together provides a way of escaping from the minefield of stochastic effects. At present the analytical framework to underpin this change of focus is not in place, and much work is to be done here if the benefits of the human genome project are to be fully realised. One way forward may be the use of tree comparison statistics (Penny et al. 1993) or similar analyses that seek to find congruence between loci (Jin et al. 1999). However demonstrating congruence and making inferences that are more than the sum of their parts, are presently worlds apart. Will a consensus view of human prehistory from the entirety of the genome tell any interesting stories or will it just be an relatively uninteresting average of them all?

The danger of stochastic effects is greatest when comparing Y-chromosomal and mtDNA diversity, because of their small effective population sizes. The number of papers seeking to show sex-specific differences in prehistory is growing rapidly (Bravi et al. 1997; Seielstad et al. 1998; Hammer et al. 1998; Hurles et al. 1998; Hurles et al. 1999; PerezLezaun et al. 1999). In truth we have little concept of the degree to which stochastic variation between these loci might cause us to draw spurious conclusions. There are no obvious control experiments. However a common picture is beginning to emerge of substantial reproducible differences between mtDNA and Y chromosome. I believe the greater geographic differentiation of the Y chromosome compared to mtDNA is just such a robust finding. In addition there are a number of examples of small, ancient, isolated populations that whilst retaining a large amount of mtDNA diversity find themselves almost monomorphic for Y-chromosomal diversity, witness the Taiwanese tribes in chapter 6. A similar picture seems likely for the Khoisan (H. Soodyall, unpublished observations). Furthermore it seems likely that the long-awaited publication of Peter Underhill's 160+ biallelic Y-chromosomal markers will reveal a Y-chromosomal tree with a number of long branches defined by multiple markers but with intermediate haplotypes being unrepresented - a topology that is notably very different from those of global mtDNA phylogenies. The tendency will be to interpret these data in terms of sex-specific movements extrapolated to a global generalisation. The algebraic reduction of these findings to single parameters such as migration rate and/or effective population size belies the multitude of factors that can influence these summary statistics. To take an example; the findings of Seielstad et al. (1998) could result not from differential migration rates between the sexes but from a difference in effective population size that results from a sex-specific difference in the social transmission of reproductive behaviour (Austerlitz and Heyer 1998). In other words, the sons of men that have many sons have many sons themselves, whereas daughters of women may not share the same inheritance pattern. We can learn much from historical demographers in this regard.

Related to the issues surrounding the combining of data from the phylogenies of multiple loci is the concept of combining data from different disciplines within a common analytical framework. At present, inferences from the different disciplines are combined post-analysis purely by inspection of compatibility. Underlying similarities in the raw data of different fields may allow another approach. There have been many historical attempts by linguists and archaeologists to model the evolution of languages and cultures on biological mechanisms (Kirch and Green 1992). Often these have been heartily rebuffed by experts within the fields themselves as being based on unrealistic assumptions. Lamarckian rather than Darwinian processes may be the dominant mechanism of cultural evolution (Kirch and Green 1992). Shoe-horning cultural and linguistic data into a framework of natural selection has been tried and has failed. However at present there is a movement in both historical linguistics and archaeology towards numerical analyses in general and

phylogenetic ones in particular (Dyen et al. 1992; Durham 1992). This is accompanied by a greater appreciation of both the distinction between homologous and analogous forms, and the sometimes reticulate nature of all evolutionary processes. Consequently the application of network analyses, which can cope with a limited amount of reticulation, to these different data sets may allow a more direct interaction between these disciplines (Forster et al. 1998).

A more practical application of this growing knowledge about global population histories and structures may be towards understanding the distribution of disease alleles (Barbujani 1999). An appreciation of population structure will prove vital if the ongoing construction of databases of genome-wide SNP variation is to result in meaningful genome association scans for common diseases, as is hoped by the many large pharmaceutical companies, government agencies and medical charities involved in these projects.

References

Affara NA, Lau YFC (1994) Report of the First International Workshop on Y chromosome mapping 1994. Cytogenet. Cell Genet. 67:360-386

Ammerman AJ, Cavalli-Sforza LL (1984) Neolithic transition and the genetics of populations in Europe. Princeton University Press, Princeton, New Jersey

Arguello R, Avakian H, Madrigal JA (1996) A high resolution typing method for the simultaneous identification of HLA class I alleles with 40 universal SSO probes. Human Immunology 49:163

Armour JAL, Anttinen T, May CA, Vega EE, Sajantila A, Kidd JR, Kidd KK, et al (1996) Minisatellite diversity supports a recent African origin for modern humans. Nature Genet. 13:154-160

Armour JAL, Jeffreys AJ (1992) Recent advances in minisatellite biology. FEBS Lett. 307:113-115

Arnason U, Gullberg A, Janke A (1998) Molecular timing of primate divergences as estimated by two nonprimate calibration points. Journal Of Molecular Evolution 47:718-727

Austerlitz F, Heyer E (1998) Social transmission of reproductive behaviour increases frequency of inherited disorders in a young expanding population. PNAS 95:15140-15144

Austin CC (1999) Lizards took express train to Polynesia. Nature 397:113-114

Avise JC (1989) Gene Trees and Organismal Histories - a Phylogenetic Approach to Population Biology. Evolution 43:1192-1208

Avise JC (1994) Molecular markers, natural history and evolution. Chapman & Hall, pp 126-138

Avise JC (1998) The history and purview of phylogeography: a personal reflection. Molecular Ecology 7:371-379

Avise JC, Arnold J, Ball RM, Bermingham E, Lamb T, Neigel JE, Reeb CA, et al (1987) Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. Annual Review of Ecology and Systematics 18:489-522

Baird DM, Jeffreys AJ, Royle NJ (1995) Mechanisms underlying telomere repeat turnover, revealed by hypervariable variant repeat distribution patterns in the human Xp/Yp telomere. EMBO J. 14:5433-5443

Bandelt H-J, Forster P, Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. Mol. Biol. Evol. 16:37-48

Bandelt H-J, Forster P, Sykes BC, Richards MB (1995) Mitochondrial portraits of human populations using median networks. Genetics 141:743-753

Barbujani G (1997) DNA variation and language affinities. American Journal Of Human Genetics 61:1011-1014

Barbujani G (1999) Geographical patterns: how to identify them, and why. Human Biology in press

Barbujani G, Jacquez GM, Ligi L (1990) Diversity Of Some Gene-Frequencies In European and Asian Populations .5. Steep Multilocus Clines. American Journal Of Human Genetics 47:867-875

Barbujani G, Magagni A, Minch E, CavalliSforza LL (1997) An apportionment of human DNA diversity. Proceedings Of the National Academy Of Sciences Of the United States Of America 94:4516-4519

Barbujani G, Oden NL, Sokal RR (1989) Detecting Regions Of Abrupt Change In Maps Of Biological Variables. Systematic Zoology 38:376-389

Barbujani G, Pilastro A, Dedomenico S, Renfrew C (1994) Genetic-Variation In North-Africa and Eurasia - Neolithic Demic Diffusion Vs Paleolithic Colonization. American Journal Of Physical Anthropology 95:137-154

Barbujani G, Sokal RR (1990) Zones of sharp genetic change in Europe are also linguistic boundaries. Proc. Natl. Acad. Sci. USA 87:1816-1819

Barbujani G, Sokal RR (1991) Genetic Population-Structure Of Italy .2. Physical and Cultural Barriers to Gene Flow. American Journal Of Human Genetics 48:398-411

Barbujani G, Sokal RR, Oden NL (1995) Indo-European Origins - a Computer-Simulation Test Of 5 Hypotheses. American Journal Of Physical Anthropology 96:109-132

Barbujani G, Vian P, Fabbris L (1992) Cultural Barriers Associated With Large Gene-Frequency Differences Among Italian Populations. Human Biology 64:479-495

Bellwood P (1987) The Polynesians. Thames and Hudson, London

Bellwood P (1991) The Austronesian dispersal and the origin of languages. Sci. Am. July:70-75

Bellwood PS (1979) Man's conquest of the Pacific: the prehistory of South-East Asia and Oceania. Oxford University Press, Oxford

Bellwood PS (1989) The colonization of the Pacific: some current hypotheses. In: Hill AVS, Serjeantson SW (eds) The Colonization of the Pacific: a genetic trail. Clarendon Press, Oxford, pp 1-60

Bergen AW, Wang CY, Tsai J, Jefferson K, Dey C, Smith KD, Park SC, et al (1999) An Asian-Native American paternal lineage identified by RPS4Y resequencing and by microsatellite haplotyping. Annals Of Human Genetics 63:63-80

Bermingham E, Moritz C (1998) Comparative phylogeography: concepts and applications. Molecular Ecology 7:367-369

Berta P, Hawkins JR, Sinclair AH, Taylor A, Griffiths BL, Goodfellow PN, Fellous M (1990) Genetic evidence equating *SRY* and the testis-determining factor. Nature 348:448-450

Bertorelle G, Barbujani G (1995) Analysis Of Dna Diversity By Spatial Autocorrelation. Genetics 140:811-819

Bertranpetit J, Calafell F (1996) Genetic and geographic variability in cystic fibrosis: evolutionary considerations. In: Chadwick D, Cardew G (eds) Variation in the human genome. John Wiley & sons, Chichester, pp 97-118

Bianchi NO, Bailliet G, Bravi CM, Carnese RF, Rothhammer F, Martínez-Marignac VL, Pena SDJ (1997) Origin of Amerindian Y-chromosomes as inferred by the analysis of six polymorphic markers. Am. J. Phys. Anthropol. 102:79-89

Biggs B (1971) The languages of Polynesia. In: Sebeok TA (ed) Current trends in linguistics. Mouton, The Hague

Blust RA (1990) Linguistic change and reconstruction methodology in the Austronesian language family. In: Baldi P (ed) Linguistic change and reconstruction methodology. Mouton de Gruyter, Berlin/New York

Boom R, Sol CJA, Salimans MMM, Jansen CL, Wertheim-van Dillen PME, Van Der Noordaa J (1990) Rapid and simple method for purification of nucleic acids. Journal of Clinical Microbiology 28:495-503

Bouzekri N, Taylor PG, Hammer MF, Jobling MA (1998) Novel mutation processes in the evolution of a haploid minisatellite, MSY1: array homogenization without homogenization. Hum. Mol. Genet. 7:655-659

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL (1994) High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368:455-457

Bravi CM, Sans M, Bailliet G, Martinez-Marignac VL, Portas M, Barreto I, Bonilla C, et al (1997) Characterization of mitochondrial DNA and Y-chromosome haplotypes in a Uruguayan population of African ancestry. Hum. Biol. 69:641-652

Brown MD, Hosseini SH, Torroni A, Bandelt HJ, Allen JC, Schurr TG, Scozzari R, et al (1998) mtDNA haplogroup X: An ancient link between Europe western Asia and North America? American Journal Of Human Genetics 63:1852-1861

Brown WRA (1988) A physical map of the pseudoautosomal region. EMBO Journal 7:2377-2385

Calafell F, Bertranpetit J (1994) Principal component analysis of gene frequencies and the origin of Basques. Am. J. Phys. Anthropol. 93:201-215

Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. Nature 325:31-36

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Lane CR, et al (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nature Genetics 22:231-238

Cavalli-Sforza L, Piazza A, Menozzi P, Mountain J (1989) Genetic and Linguistic Evolution. Science 244:1128-1129

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press, Princeton, N.J.

Cavalli-Sforza LL, Minch E (1997) Paleolithic and neolithic lineages in the European mitochondrial gene pool. American Journal Of Human Genetics 61:247-251

Chandley AC, Goetz P, Hargreave TB, Joseph AM, Speed RM (1984) On the nature and extent of XY pairing at meiotic prophase in man. Cytogenet. Cell Genet. 38:241-247

Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, Winkler J, et al (1996) Accessing genetic information with high-density DNA arrays. Science 274:610-614

Chikhi L, DestroBisol G, Pascali V, Baravelli V, Dobosz M, Barbujani G (1998) Clinal variation in the nuclear DNA of Europeans. Human Biology 70:643-657

Ciminelli BM, Pompei F, Malaspina P, Hammer M, Persichetti F, Pignatti PF, Palena A, et al (1995) Recurrent simple tandem repeat mutations during human Y-chromosome radiation in Caucasian subpopulations. J. Mol. Evol. 41:966-973

Collins R (1986) The Basques. Blackwell, Oxford

Comas D, Mateu E, Calafell F, Pérez-Lezaun A, Bosch E, Martínez-Arias R, Bertranpetit J (1998) HLA class I and class II DNA typing and the origin of Basques. Tissue Antigens 51:30-40

Comas D, Reynolds R, Sajantila A (1999) Analysis of mtDNA HVRII in several human populations using an immobilised SSO probe hybridisation assay. European Journal Of Human Genetics 7:459-468

Cooper G, Amos W, Hoffman D, Rubinsztein DC (1996) Network analysis of human Y microsatellite haplotypes. Hum. Mol. Genet. 5:1759-1766

Côrte-Real HBSM, Macaulay VA, Richards MB, Hariti G, Issad MS, Cambon-Thomsen A, Papiha S, et al (1996) Genetic diversity in the Iberian Peninsula determined from mitochondrial sequence analysis. Ann. Hum. Genet. 60:331-350

Cotton RGH (1997) Slowly but surely towards better scanning for mutations. Trends Genet. 13:43-46

D'Esposito M, Ciccodicola A, Gianfrancesco F, Esposito T, Flagiello L, Mazzarella R, Schlessinger D, et al (1996) A synaptobrevin-like gene in the Xq28 pseudoautosomal region undergoes X inactivation. Nature Genet. 13:227-229

de Knijff P, Kayser M, Caglià A, Corach D, Fretwell N, Gehrig C, Graziosi G, et al (1997) Chromosome Y microsatellites: population genetic and evolutionary aspects. Int. J. Legal Med. 110:134-140

Dorit RL, Akashi H, Gilbert W (1995) Absence of polymorphism at the ZFY locus on the human Y chromosome. Science 268:1183-1185

Dumolin-Lapegue S, Demesure B, Fineschi S, LeCorre V, Petit RJ (1997) Phylogeographic structure of white oaks throughout the European continent. Genetics 146:1475-1487

Durham W (1992) Applications of Evolutionary Culture Theory. Annual Review of Anthropology 21:331-355

Dyen I, Kruskal JB, Black P (1992) An Indo-European classification: A lexicostatistical experiment. American Philosophical Society, Philadelphia

Efron B (1982) The Jacknife, the Bootstrap and other resampling plans. Regional conference series in applied mathematics, Philidelphia

Ellis NA, Tippett P, Petty A, Reid M, Weller PA, Ye TZ, German J, et al (1994) PBDX is the Xg blood-group gene. Nature Genet. 8:285-290

Ellison JW, Ramos C, Yen PH, Shapiro LJ (1993) 2 protein isoforms are encoded by the human pseudoautosomal gene XE7. Clin. Res. 41:A 3-A 3

EyreWalker A, Smith NH, Smith JM (1999) How clonal are human mitochondria? Proceedings Of the Royal Society Of London Series B-Biological Sciences 266:477-483

Fisher EMC, Beer-Romero P, Brown LG, Ridley A, McNeil JA, Lawrence JB, Willard HF, et al (1990) Homologous ribosomal protein genes on the human X and Y chromosomes: escape from inactivation and possible implications for Turner syndrome. Cell 63:1205-1218

Fisher RA (1930) The genetical theory of natural selection. Clarendon Press, Oxford

Flint J, Boyce AJ, Martinson JJ, Clegg JB (1989) Population bottlenecks in Polynesia revealed by minisatellites. Hum. Genet. 83:257-263

Forster P, Harding R, Torroni A, Bandelt H-J (1996) Origin and evolution of Native American mtDNA variation: a reappraisal. Am. J. Hum. Genet. 59:935-945

Forster P, Kayser M, Meyer E, Roewer L, Pfeiffer H, Benkmann H, Brinkmann B (1998) Phylogenetic resolution of complex mutational features at Y-STR DYS390 in Aboriginal Australians and Papuans. Molecular Biology and Evolution 15:1108-1114

Forster P, Toth A, Bandelt H-J (1998) Evolutionary network analysis of word lists: visualising the relationships between Alpine Romance languages. Journal of Quantitative Linguistics 5:174-187

Fortin M-J, Drapeau P (1995) Delineation of ecological boundaries: comparison of approaches and significance tests. Oikos 72:323-332

Foster EA, Jobling MA, Taylor PG, Donnelly P, de Knijff P, Mieremet R, Zerjal T, et al (1998) Jefferson fathered slave's last child. Nature 396:27-28

Foster EA, Jobling MA, Taylor PG, Donnelly P, de Knijff P, Mieremet R, Zerjal T, et al (1999) The Thomas Jefferson paternity case. Nature 397:32

Fraser A (1993) The Gypsies. Blackwell Publishers, Oxford

Goldstein DB, Linares AR, Cavalli-Sforza LL, Feldman MW (1995) An evaluation of genetic distances for use with microsatellite loci. Genetics 139:463-471

Goldstein DB, Zhivotovsky LA, Nayar K, Linares AR, Cavalli-Sforza LL, Feldman MW (1996) Statistical properties of the variation at linked microsatellite loci: implications for the history of human Y chromosomes. Mol. Biol. Evol. 13:1213-1218

Goodfellow PN, Lovell-Badge R (1993) SRY and sex determination in mammals. Annual Review Of Genetics 27:71-92

Gough NM, Gearing DP, Nicola NA, Baker E, Pritchard M, Callen DF, Sutherland GR (1990) Localization of the human GM-CSF receptor gene to the X-Y pseudoautosomal region. Nature 345:734-736

Grimes BF (1996) Ethnologue. Summer Institute of Linguistics Inc., Dallas, Texas

Haff LA, Smirnov IP (1997) Single-nucleotide polymorphism identification assays using a thermostable DNA polymerase and delayed extraction MALDI-TOF mass spectrometry. Genome Research 7:378-388

Hagelberg E, Goldman N, Lio P, Whelan S, Schiefenhovel W, Clegg JB, Bowden DK (1999) Evidence for mitochondrial DNA recombination in a human population of island Melanesia. Proceedings Of the Royal Society Of London Series B-Biological Sciences 266:485-492

Hagelberg E, Kayser M, Nagy M, Roewer L, Zimdahl H, Krawczak M, Lio P, et al (1999) Molecular genetic evidence for the human settlement of the Pacific: analysis of mitochondrial DNA, Y chromosome and HLA markers. Philosophical Transactions of the Royal Society 354:141-152

Haldane JBS (1932) The causes of evolution. Longmans and Green, London

Haldane JBS (1940) The blood-group frequencies of European peoples and racial origins. Human Biology 12:457-480

Hammer MF (1994) A recent insertion of an Alu element on the Y chromosome is a useful marker for human population studies. Mol. Biol. Evol. 11:749-761

Hammer MF (1995) A recent common ancestry for human Y chromosomes. Nature 378:376-378

Hammer MF, Karafet T, Rasanayagam A, Wood ET, Altheide TK, Jenkins T, Griffiths RC, et al (1998) Out of Africa and back again: Nested cladistic analysis of human Y chromosome variation. Molecular Biology and Evolution 15:427-441

Hammer MF, Spurdle AB, Karafet T, Bonner MR, Wood ET, Novelletto A, Malaspina P, et al (1997) The geographic distribution of human Y chromosome variation. Genetics 145:787-805

Harding RM, Fullerton SM, Griffiths RC, Bond J, Cox MJ, Schneider JA, Moulin DS, et al (1997) Archaic African *and* Asian lineages in the genetic ancestry of modern humans. Am. J. Hum. Genet. 60:772-789

Harpending HC, Batzer MA, Gurven M, Jorde LB, Rogers AR, Sherry ST (1998) Genetic traces of ancient demography. Proceedings Of the National Academy Of Sciences Of the United States Of America 95:1961-1967

Hather J, Kirch PV (1991) Prehistoric sweet potato (Ipomoea batatas) from Mangaia Island, Central Polynesia. Antiquity 65:887-893

Hawkins JR, Taylor A, Berta P, Levilliers J, Vanderauwera B, Goodfellow PN (1992) Mutational analysis of SRY - nonsense and missense mutations in XY sex reversal. Human Genetics 88:471-474

Hazout S, Lucotte G (1986) Vers une généalogie du chromosome Y. Annales Génétiques 29:246-252

Hearne CM, Ghosh S, Todd JA (1992) Microsatellites for linkage analysis of genetic traits. TIG 8:288-294

Heyer E, Puymirat J, Dieltjes P, Bakker E, de Knijff P (1997) Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. Hum. Mol. Genet. 6:799-803

Heyerdahl T (1950) Kontiki: across the Pacific by raft. Rand McNally, Chicago

Hill AVS, Gentile B, Bonnardot JM, Roux J, Weatherall DJ, Clegg JB (1985) Polynesian origins and affinities: globin gene variants in eastern Polynesia. Am. J. Hum. Genet. 40:453-463

Hill AVS, Serjeantson SW (1989) The colonization of the Pacific: a genetic trail. Clarendon, Oxford

Hoogendoorn B, Owen MJ, Oefner PJ, Williams N, Austin J, O'Donovan MC (1999) Genotyping single nucleotide polymorphisms by primer extension and high performance liquid chromatography. Human Genetics 104:89-93

Huber CG, Oefner PJ, Bonn GK (1995) Rapid and accurate sizing of DNA fragments by ion-pair chromatography on alkylated nonporous Poly(styrene-divinylbenzene) particles. Analytical Chemistry 67:578-585

Hudson RR, Boos DD, Kaplan NL (1992) A statistical test for geographical subdivision. Molecular Biology and Evolution 9:138-151

Hurles ME, Irven C, Nicholson J, Taylor PG, Santos FR, Loughlin J, Jobling MA, et al (1998) European y-chromosomal lineages in Polynesians: A contrast to the population structure revealed by mtDNA. American Journal Of Human Genetics 63:1793-1806

Hurles ME, Veitia R, Arroyo E, Armenteros M, Bertranpetit J, Pérez-Lezaun A, Bosch E, et al (1999) Recent male-mediated gene flow over a linguistic barrier in Iberia suggested by analysis of a Y-chromosomal DNA polymorphism. Am. J. Hum. Genet. 65:1437-1448

Hurst LD (1999) The evolution of genomic anatomy. Trends In Ecology & Evolution 14:108-112

Ivanov PL, Wadhams MJ, Roby RK, Holland MM, Weedn VW, Parsons TJ (1996) Mitochondrial DNA sequence heteroplasmy in the Grand Duke of Russia Georgij Romanov establishes the authenticity of the remains of Tsar Nicholas II. Nature Genet. 12:417-420

Jager RJ, Harley VR, Pfeiffer RA, Goodfellow PN, Scherer G (1992) A familial mutation in the testis-determining gene SRY shared by both sexes. Human Genetics 90:350-355

Jakubiczka S, Arnemann J, Cooke HJ, Krawczak M, Schmidtke J (1990) A search for restriction fragment length polymorphism of the human Y chromosome. Hum. Genet. 84:86-88

Jeffreys AJ, Allen MJ, Armour JAL, Collick A, Dubrova Y, Fretwell N, Guram T, et al (1995) Mutation processes at human minisatellites. Electrophoresis 16:1577-1585

Jeffreys AJ, MacLeod A, Tamaki K, Neil DL, Monckton DG (1991) Minisatellite repeat coding as a digital approach to DNA typing. Nature 354:204-209

Jeffreys AJ, Murray J, Neumann R (1998) High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot. Molecular Cell 2:267-273

Jeffreys AJ, Neumann R, Wilson V (1990) Repeat unit sequence variation in minisatellites: a novel source of DNA polymorphism for studying variation and mutation by single molecule analysis. Cell 60:473-485

Jegalian K, Page DC (1998) A proposed path by which genes common to mammalian X and Y chromosomes become X inactivated. Nature 394:776-780

Jennings JD (1979) The prehistory of Polynesia. Harvard University Press, Cambridge

Jin L, Underhill PA, Doctor V, Davis RW, Shen PD, CavalliSforza LL, Oefner PJ (1999) Distribution of haplotypes from a chromosome 21 region distinguishes multiple prehistoric human migrations. Proceedings Of the National Academy Of Sciences Of the United States Of America 96:3796-3800

Jobling MA (1994) A survey of long-range DNA polymorphisms on the human Y chromosome. Hum. Mol. Genet. 3:107-114

Jobling MA, Bouzekri N, Taylor PG (1998) Hypervariable digital DNA codes for human paternal lineages: MVR-PCR at the Y-specific minisatellite, MSY1 (*DYF155S1*). Hum. Mol. Genet. 7:643-653

Jobling MA, Fretwell N, Dover GA, Jeffreys AJ (1994) Digital coding of human Y chromosomes - MVR-PCR at Y-specific minisatellites. Cytogenet. Cell Genet. 67:390

Jobling MA, Heyer E, Dieltjes P, de Knijff P (1999) Y-chromosome-specific microsatellite mutation rates re-examined using a minisatellite, MSY1. Hum. Mol. Genet. in press

Jobling MA, Pandya A, Tyler-Smith C (1997) The Y chromosome in forensic analysis and paternity testing. Int. J. Legal Med.

Jobling MA, Samara V, Pandya A, Fretwell N, Bernasconi B, Mitchell RJ, Gerelsaikhan T, et al (1996) Recurrent duplication and deletion polymorphisms on the long arm of the Y chromosome in normal males. Hum. Mol. Genet. 5:1767-1775

Jobling MA, Tyler-Smith C (1995) Fathers and sons: the Y chromosome and human evolution. Trends Genet. 11:449-456

Jobling MA, Williams G, Schiebel K, Pandya A, McElreavey K, Salas L, Rappold GA, et al (1998) A selective difference between human Y-chromosomal DNA haplotypes. Curr. Biol. 8:1391-1394

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

Kalaydjieva L, Hallmayer J, Chandler D, Savov A, Nikolova A, Angelicheva A, King RHH, et al (1996) Gene mapping in Gypsies identifies a novel demyelinating neuropathy on chromosome 8q24. Nature Genet. 14:214-217

Karafet T, Zegura SL, VuturoBrady J, Posukh O, Osipova L, Wiebe V, Romero F, et al (1997) Y chromosome markers and trans-Bering Strait dispersals. American Journal Of Physical Anthropology 102:301-314

Karafet TM, Zegura SL, Posukh O, Osipova L, Bergen A, Long J, Goldman D, et al (1999) Ancestral Asian source(s) of New World Y-chromosome founder haplotypes. Am. J. Hum. Genet. 64:817-831

Karafet TM, Zegura SL, Posukh O, Osipova L, Bergen A, Long J, Goldman D, et al (1999) Ancestral Asian source(s) of New World Y-chromosome founder haplotypes. American Journal Of Human Genetics 64:817-831

Kayser M, Caglià A, Corach D, Fretwell N, Gehrig C, Graziosi G, Heidorn F, et al (1997) Evaluation of Y-chromosomal STRs: a multicenter study. Int. J. Legal Med. 110:125-133

Kayser M, Caglià A, Corach D, Fretwell N, Gehrig C, Graziosi G, Heidorn F, et al (1997) Evaluation of Y chromosomal STRs: a multicenter study. Int. J. Legal Med. This issue.

Kermouni A, Van Roost E, Arden KC, Vermeesch JR, Weiss S, Godelaine D, Flint J, et al (1995) The IL-9 receptor gene (IL9R): genomic structure, chromosomal localization in the pseudoautosomal region of the long arm of the sex chromosomes, and identification of the IL9R pseudogenes at 9qter, 10pter, 16pter, and 18pter. Genomics 29:371-382

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

Kirch PV, Green RC (1992) History, Phylogeny and Evolution in Polynesia. Current Anthropology 33:161-186

Kittles RA, Long JC, Bergen AW, Eggbert M, Virkkunen M, Linnoila M, Goldman D (1999) Cladistic association analysis of Y chromosome effects on alcohol dependance and related personality traits. PNAS 96:4204-4209

Krawczak M, Ball EV, Cooper DN (1998) Neighbouring nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. American Journal of Human Genetics 63:474-488

Krings M, Stone A, Schmitz RW, Krainitzki H, Stoneking M, Paabo S (1997) Neanderthal DNA sequences and the origin of modern humans. Cell 90:19-30

Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nature Genetics 22:139-144

Kunkel TA (1990) Misalignment-mediated DNA synthesis errors. Biochemistry 29:8003-8011

Kuroki Y, Iwamoto T, Lee JW, Yoshiike M, Nozawa S, Nishida T, Ewis AA, et al (1999) Spermatogenic ability is different among males in different Y chromosome lineage. Journal Of Human Genetics 44:289-292

Kvaløy K, Galvagni F, Brown WRA (1994) The sequence organization of the long arm pseudoautosomal region of the human sex-chromosomes. Hum. Mol. Genet. 3:771-778

Kwok C, Tyler-Smith C, Medonca BB, Hughes I, Berkovitz GD, Goodfellow PN, Hawkins JR (1996) Mutation analysis of 2kb 5′ to SRY in XY females and XX intersex subjects. J. Med. Genet. 33:465-468

Laan M, Pääbo S (1997) Demographic history and linkage disequilibrium in human populations. Nature Genet. 17:435-438

Lahermo P, Savontaus ML, Sistonen P, Beres J, deKnijff P, Aula P, Sajantila A (1999) Y chromosomal polymorphisms reveal founding lineages in the Finns and the Saami. European Journal Of Human Genetics 7:447-458

Lahn BT, Page DC (1997) Functional coherence of the human Y chromosome. Science 278:675-680

Lahr MMF, R. (1994) Multiple dispersals and modern human origins. Evol. Anthropol. 3:48-60

Lam NS-N (1983) Spatial interpolation methods: a review. The American Cartographer 10:129-149

Lasker GW (1985) Surnames and genetic structure. In: Lasker GW, Mascie-Taylor CGN, Roberts DF, Washburn SL (eds) Cambridge studies in biological anthropology. Cambridge University Press, Cambridge

Lau YFC (1999) Gonadoblastoma, testicular, and prostate cancers and the TSPY gene. American Journal of Human Genetics 64

Lewontin RC (1972) The apportionment of human diversity. Evolutionary Biology 6:381-398

Li L, Hamer DH (1995) Recombination and allelic association in the Xq/Yq homology region. Hum. Mol. Genet. 4:2013-2016

Liegeois JP (1994) Roma, Gypsies, Travellers. Council of Europe Press, Strasbourg

Liu W, Smith DI, Rechtzigel KJ, Thibodeau SN, James CD (1998) Denaturing High Performance Liquid Chromatography (DHPLC) used in the detection of germline and somatic mutations. Nucleic Acid Research 26:1396-1400

Lucotte G, Hazout S (1996) Y chromosome DNA haplotypes in Basques. J. Mol. Evol. 42:472-475

Lucotte G, Sriniva KR, Loirat F, Hazout S, Ruffié J (1990) The p49/*Taq*I Y-specific polymorphisms in three groups of Indians. Gene Geography 4:21-28

Lum JK, Cann RL (1998) mtDNA and language support a common origin of Micronesians and Polynesians in Island Southeast Asia. American Journal Of Physical Anthropology 105:109-119

Lum JK, Cann RL, Martinson JJ, Jorde LB (1998) Mitochondrial and nuclear genetic relationships among Pacific Island and Asian populations. American Journal Of Human Genetics 63:613-624

Lum JK, Richards O, Ching C, Cann RL (1994) Polynesian mitochondrial DNAs reveal three deep maternal lineage clusters. Hum. Biol. 66:567-590

Luria SE, Delbruck M (1943) Genetics 28:491-511

Ma K, Inglis JD, Sharkey A, Bickmore WA, Hill RE, Prosser EJ, Speed RM, et al (1993) A Y chromosome gene family with RNA-binding protein homology - candidates for the azoospermia factor AZF controlling human spermatogenesis. Cell 75:1287-1295

Malaspina P, Cruciani F, Ciminelli BM, Terrenato L, Santolamazza P, Alonso A, Banyko J, et al (1998) Network analyses of Y-chromosomal types in Europe, Northern Africa, and Western Asia reveal specific patterns of geographic distribution. American Journal Of Human Genetics 63:847-860

Malaspina P, Persichetti F, Novelletto A, Iodice C, Terrenato L, Wolfe J, Ferraro M, et al (1990) The human Y chromosome shows a low level of DNA polymorphism. Ann. Hum. Genet. 54:297-305

Manly BFJ (1991) Randomization methods and monte carlo methods in biology. Chapman and Hall, London

Mantel N (1967) The detection of disease clustering and a generalised regression approach. Cancer Research 27:209-220

Manz E, Schnieders F, Brechlin AM, Schmidtke J (1993) TSPY-related sequences represent a microheterogeneous gene family organized as constitutive elements in DYZ5 tandem repeat units on the humna Y chromosome. 17:726-731

Marshall Graves JA (1995) The origin and function of the mammalian Y chromosome and Y-borne genes - an evolving understanding. BioEssays 17:311-321

Martinson JJ, Harding RM, Philippon G, Saintemarie FF, Roux J, Boyce AJ, Clegg JB (1993) Demographic reductions and genetic bottlenecks in humans - minisatellite allele distributions in Oceania. Hum. Genet. 91:445-450

Matheron G (1971) The theory of regionalised variables and its application. Les cahiers du centre de morphologie mathematique de Fontainebleau, Fontainebleau

Mathias N, Bayés M, Tyler-Smith C (1994) Highly informative compound haplotypes for the human Y chromosome. Hum. Mol. Genet. 3:115-123

Matisoo-Smith E, Roberts RM, Irwin GJ, Allen JS, Penny D, Lambert DM (1998) Patterns of prehistoric human mobility in Polynesia indicated by mtDNA from the Pacific rat. Proceedings Of the National Academy Of Sciences Of the United States Of America 95:15145-15150

Maude HE (1981) Slavers in Paradise: the Peruvian labour trade in Polynesia, 1862-1864. Institute of Pacific studies, University of the South Pacific, Suva

McElreavey K, Krausz C (1999) Male Infertility and the Y chromosome. American Journal of Human Genetics 64

Melton T, Clifford S, Martinson J, Batzer M, Stoneking M (1998) Genetic Evidence for the Proto-Austronesian Homeland in Asia: mtDNA and Nuclear DNA Variation in Taiwanese Aboriginal Tribes. American Journal of Human Genetics 63:1807-1823

Melton T, Peterson R, Redd AJ, Saha N, Sofro ASM, Martinson J, Stoneking M (1995) Polynesian genetic affinities with Southeast Asian populations as identified by mtDNA analysis. Am. J. Hum. Genet. 57:403-414

Michalakis Y, Excoffier L (1996) A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. Genetics 142:1061-1064

Milatovich A, Kitamura T, Miyajima A, Francke U (1993) Gene for the alpha-subunit of the human Interleukin-3 receptor (IL3RA) localized to the X-Y pesudoautosomal region. Am. J. Hum. Genet. 53:1146-1153

Mitchell RJ, Hammer MF (1996) Human evolution and the Y chromosome. Curr. Opin. Genet. Dev. 6:737-742

Monckton DG, Neumann R, Guram T, Fretwell N, Tamaki K, MacLeod A, Jeffreys AJ (1994) Minisatellite mutation-rate variation associated with a flanking DNA sequence polymorphism. Nature Genet. 8:162-170

Murray-McIntosh RP, Scrimshaw BJ, Hatfield PJ, Penny D (1998) Testing migration patterns and estimating founding population size in Polynesia by using human mtDNA sequences. PNAS 95:9047-9052

Nei M (1987) Molecular Evolutionary Genetics. Columbia University Press, New York

Ngo KY, Vergnaud G, Johnsson C, Lucotte G, Weissenbach J (1986) A DNA probe detecting multiple haplotypes of the human Y chromosome. Am. J. Hum. Genet. 38:407-418

Nickerson DA, Whitehurst C, Boysen C, Charmley P, Kaiser R, Hood L (1992) Identification of clusters of biallelic polymorphic sequence-tagged sites (pSTSs) that generate highly informative and automatable markers for genetic-linkage mapping. Genomics 12:377-387

Ohno S (1967) Sex chromosomes and sex-linked genes. Springer-Verlag, Berlin

Ophoff RA, Terwindt GM, Vergouwe MN, van Eijk R, Oefner PJ, Hoffman SMG, Lamerdin JE, et al (1996) Familial hemiplegic migraine and episodic ataxia type-2 are caused by mutations in the Ca2+ channel gene CACNL1A4. Cell 87:543-552

Parsons TJ, Muniec DS, Sullivan K, Woodyatt N, Alliston-Greiner R, Wilson MR, Berry DL, et al (1997) A high observed substitution rate in the human mitochondrial DNA control region. Nature Genet. 15:363-367

Pawley A, Ross M (1993) Austronesian historical linguistics and culture history. Annual Review of Anthropology 22:425-459

Penny D, Watson EE, Steel MA (1993) Trees From Languages and Genes Are Very Similar. Systematic Biology 42:382-384

Perez-Lezaun A, Calafell F, Mateu E, Comas D, Ruiz-Pacheco R, Bertranpetit J (1997) Microsatellite variation and the differentiation of modern humans. Hum. Genet. 99:1-7

Perez-Lezaun A, Calafell F, Comas D, Mateu E, Bosch E, MartinezArias R, Clarimon J, et al (1999) Sex-specific migration patterns in central Asian populations, revealed by analysis of Y-chromosome short tandem repeats and mtDNA. American Journal Of Human Genetics 65:208-219

Petit RJ, ElMousadik A, Pons O (1998) Identifying populations for conservation on the basis of genetic markers. Conservation Biology 12:844-855

Piccolo F, Jeanpierre M, Leturcq F, Dode C, Azibi K, Toutain A, Merlini L, et al (1996) A founder mutation in the gamma-sarcoglycan gene of Gypsies possibly predating their migration out of India. Human Molecular Genetics 5:2019-2022

Piertney SB, MacColl ADC, Bacon PJ, Dallas JF (1998) Local genetic structure in red grouse (Lagopus lagopus scoticus): evidence from microsatellite DNA markers. Molecular Ecology 7:1645-1654

Poloni ES, Semino O, Passarino G, Santachiara-Benerecetti AS, Dupanloup L, Langaney A, Excoffier L (1997) Human genetic affinities for Y-chromosome P49a,f/*Taq*I haplotypes show strong correspondence with linguistics. Am. J. Hum. Genet. 61:1015-1035

Previdere C, Stuppia L, Gatta V, Fattorini P, Palka G, TylerSmith C (1999) Y-chromosomal DNA haplotype differences in control and infertile Italian subpopulations. European Journal Of Human Genetics 7:733-736

Rao E, Weiss B, Fukami M, Rump A, Niesler B, Mertz A, Muroya K, et al (1997) Pseudoautosomal deletions encompassing a novel homeobox gene cause growth failure in idiopathic short stature and Turner syndrome. Nature Genet. 16:54-63

Redd AJ, Takezaki N, Sherry ST, McGarvey ST, Sofro ASM, Stoneking M (1995) Evolutionary history of the COII/tRNAlys intergenic 9 base pair deletion in human mitochondrial DNAs from the Pacific. Mol. Biol. Evol. 12:604-615

Renfrew C (1989) The origins of Indo-European languages. Sci. Amer. 261:106-114

Renfrew C (1994) World Linguistic Diversity. Scientific American January:104-110

Renfrew C, Nettles D (1999) Nostratic: examining a linguistic macrofamily. McDonald Institute for Archaeological Research, Cambridge

Richards M, Côrte-Real H, Forster P, Macaulay V, Wilkinson-Herbots H, Demaine A, Papiha S, et al (1996) Paleolithic and neolithic lineages in the European mitochondrial gene pool. Am. J. Hum. Genet. 59:185-203

Richards M, Macaulay V, Sykes B, Pettitt P, Hedges R, Forster P, Bandelt H-J (1997) Reply to Cavalli-Sforza and Minch. American Journal of Human Genetics 61:251-254

Richards M, Oppenheimer S, Sykes B (1998) MtDNA suggests Polynesian origins in eastern Indonesia. American Journal Of Human Genetics 63:1234-1236

Ried K, Rao E, Schiebel K, Rappold GA (1998) Gene duplications as a recurrent theme in the evolution of the human pseudoautosomal region 1: isolation of the gene ASTML. Human Molecular Genetics 7:1771-1778

Roewer L, Arnemann J, Spurr NK, Grzeschik KH, Epplen JT (1992) Simple repeat sequences on the human Y chromosome are equally polymorphic as their autosomal counterparts. Hum. Genet. 89:389-394

Roewer L, Kayser M, Dieltjes P, Nagy M, Bakker E, Krawczak M, de Knijff P (1996) Analysis of molecular variance (AMOVA) of Y-chromosome-specific microsatellites in two closely related human populations. Hum. Mol. Genet. 5:1029-1033

Roff DA, Bentzen P (1989) The Statistical-Analysis Of Mitochondrial-Dna Polymorphisms - Chi-2 and the Problem Of Small Samples. Molecular Biology and Evolution 6:539-545

Rogers AR, Harpending H (1992) Population growth makes waves in the distribution of pairwise genetic differences. Mol. Biol. Evol. 9:552-569

Roses AD, Saunders AM, Strittmatter WJ, Pericakvance MA, Schmechel D (1993) Association Of Apolipoprotein-E Allele-E4 With Late-Onset Familial and Sporadic Alzheimers-Disease. Neurology 43:A192

Ruhlen M (1991) A guide to the world's languages. Edward Arnold, London

Sajantila A, Lahermo P, Anttinen T, Lukka M, Sistonen P, Savontaus M-L, Aula P, et al (1995) Genes and languages in Europe: an analysis of mitochondrial lineages. Genome Res. 5:42-52

Sajantila A, Salem A-H, Savolainen P, Bauer K, Gehrig C, Pääbo S (1996) Paternal and maternal DNA lineages reveal a bottleneck in the founding of the Finnish population. Proc. Natl. Acad. Sci. USA in press

Santos FR, Pandya A, TylerSmith C, Pena SDJ, Schanfield M, Leonard WR, Osipova L, et al (1999) The central Siberian origin for Native American Y chromosomes. American Journal Of Human Genetics 64:619-628

Santos FR, Pena SDJ, Tyler-Smith C (1995) PCR haplotypes for the human Y chromosome based on alphoid satellite DNA variants and heteroduplex analysis. Gene 165:191-198

Santos FR, Tyler-Smith C (1996) Reading the human Y chromosome: the emerging DNA markers and human genetic history. Braz. J. Hum. Genet. 19:665-670

Sarich VM, Wilson AC (1967) Immunological timescale for hominid evolution. Science 158:1200-1203

Saxena R, Brown LG, Hawkins T, Alagappan RK, Skaletsky H, Reeve MP, Reijo R, et al (1996) The *DAZ* gene cluster on the human Y chromosome arose from an autosomal gene that was transposed, repeatedly amplified and pruned. Nature Genet. 14:292-299

Schiebel K, Weiss B, Wohrle D, Rappold G (1993) A human pseudoautosomal gene, ADP ATP translocase, escapes X-inactivation whereas a homolog on Xq is subject to X-inactivation. Nature Genetics 3:82-87

Schmitt K, Goodfellow PN (1994) Predicting the future. Nature Genetics 7:219-219

Schneider S, Kueffer J-M, Roessli D, Excoffier L (1997) Arlequin ver. 1.1: A software for population genetic data analysis. Genetics and Biometry Laboratory, University of Geneva, Geneva, Switzerland

Seielstad MT, Hebert JM, Lin AA, Underhill PA, Ibrahim M, Vollrath D, Cavalli-Sforza LL (1994) Construction of human Y-chromosomal haplotypes using a new polymorphic A to G transition. Hum. Mol. Genet. 3:2159-2161

Seielstad MT, Minch E, Cavalli-Sforza LL (1998) Genetic evidence for a higher female migration rate in humans. Nature Genet. 20:278-280

Semino O, Passarino G, Brega A, Fellous M, Santachiara-Benerecetti AS (1996) A view of the Neolithic demic diffusion in Europe through two Y chromosome-specific markers. Am. J. Hum. Genet. 59:964-968

Sheffield VC, Stone EM, Carmi R (1998) Use of isolated inbred human populations for identification of disease genes. Trends In Genetics 14:391-396

Shinka T, Tomita K, Toda T, Kotliarova SE, Lee J, Kuroki Y, Jin DK, et al (1999) Genetic variations on the Y chromosome in the Japanese population and implications for modern human Y chromosome lineage. Journal Of Human Genetics 44:240-245

Silvestroni E, Bianco I (1975) American Journal of Human Genetics 27:198-212

Sinclair AH, Berta P, Palmer MS, Hawkins JR, Griffiths B, Smith MJ, Foster JW, et al (1990) A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. Nature 346:240-244

Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. Genetics 139:457-462

Smith MJ, Goodfellow PN (1994) MIC2R - a transcribed MIC2-related sequence associated with a CpG island in the human pseudoautosomal region. Human Molecular Genetics 3:1575-1582

Smouse PE, Long JC, Sokal RR (1986) Multiple-Regression and Correlation Extensions Of the Mantel Test Of Matrix Correspondence. Systematic Zoology 35:627-632

Society CSS (1997) DHPLC workshop Second symposium on gene expression and mutation analysis, Millbrae, California, pp 32-43

Sokal RR, Harding RM, Lasker GW, Mascietaylor CGN (1992) A Spatial-Analysis Of 100 Surnames In England and Wales. Annals Of Human Biology 19:445-476

Sokal RR, Jacquez GM, Oden NL, Digiovanni D, Falsetti AB, McGee E, Thomson BA (1993) Genetic-Relationships Of European Populations Reflect Their Ethnohistorical Affinities. American Journal Of Physical Anthropology 91:55-70

Sokal RR, Oden NL, Rosenberg MS, DiGiovanni D (1997) Ethnohistory, genetics, and cancer mortality in Europeans. Proceedings Of the National Academy Of Sciences Of the United States Of America 94:12728-12731

Sokal RR, Oden NL, Walker J, DiGiovanni D, Thomson BA (1996) Historical population movements in Europe influence genetic relationships in modern samples. Human Biology 68:873-898

Sokal RR, Oden NL, Wilson C (1991) Genetic evidence for the spread of agriculture in Europe by demic diffusion. Nature 351:143-145

Sokal RR, Rohlf JF (1994) Biometry. W.H. Freeman and Co

Spriggs M, Anderson A (1993) Late colonisation of East Polynesia. Antiquity 67:200-217

Spurdle AB, Jenkins T (1992) The search for Y chromosome polymorphism is extended to negroids. Hum. Mol. Gent. 1:169-170

Spurdle AB, Woodfield DG, Hammer MF, Jenkins T (1994) The genetic affinity of Polynesians - evidence from Y chromosome polymorphisms. Annals Of Human Genetics 58:251-263

Stoneking M (1993) DNA and recent human evolution. Evol. Anthropol. 2:60-73

Stoneking M (1998) Women on the move. Nature Genet. 20:219-220

Stoneking M, Soodyall H (1996) Human evolution and the mitochondrial genome. Curr. Op. Genet. Devel. 6:731-736

Sykes B, Leiboff A, Low-Beer J, Tetzner S, Richards M (1995) The origins of the Polynesians: an interpretation from mitochondrial lineage analysis. Am. J. Hum. Genet. 57:1463-1475

Templeton AR (1993) The "Eve" hypothesis: a genetic critique and reanalysis. Am. Anthropol. 95:51-72

Templeton AR (1997) Out of Africa? What do genes tell us? Current Opinion In Genetics & Development 7:841-847

Templeton AR (1998) Nested clade analyses of phylogeographic data: testing hypotheses about gene flow and population history. Molecular Ecology 7:381-397

Templeton AR, Routman E, Phillips CA (1995) Separating population structure from population history: a cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the tiger salamander, *Ambystoma tigrinum*. Genet. 140:767-782

Terrell J (1986) Prehistory in the Pacific islands. Cambridge University Press, Cambridge
Terrell JE (1997) Colonisation of the Pacific Islands Society for American Archaeology meetings, Nashville
Thomas MG, Skorecki K, Ben-Ami H, Parfitt T, Bradman N, Goldstein DB (1998) Origins of Old Testament priests. Nature 384:138-140
Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Kidd JR, Cheung K, Bonné-Tamir B, et al (1996) Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271:1380-1387
Tishkoff SA, Goldman A, Calafell F, Speed WC, Deinard AS, BonneTamir B, Kidd JR, et al (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. American Journal Of Human Genetics 62:1389-1402
Torroni A, Lott MT, Cabell MF, Chen YS, Lavergne L, Wallace DC (1994) mtDNA and the origin of Caucasians - identification of ancient Caucasian-specific haplogroups, one of which is prone to a recurrent somatic duplication in the D-loop region. American Journal Of Human Genetics 55:760-776
Underhill PA, Jin L, Lin AA, Mehdi SQ, Jenkins T, Vollrath D, Davis RW, et al (1997) Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. Genome Research 7:996-1005
Underhill PA, Jin L, Zemans R, Oefner PJ, Cavalli-Sforza LL (1996) A pre-Columbian Y chromosome-specific transition and its implications for human evolutionary history. Proc. Natl. Acad. Sci. USA 93:196-200
Urquhart A, Oldroyd NJ, Kimpton CP, Gill P (1995) Highly discriminating heptaplex short tandem repeat PCR system for forensic identification. Biotechniques 18:116
Veitia R, Ion A, Barbaux S, Jobling MA, Souleyreau N, Ennis K, Ostrer H, et al (1997) Mutations and sequence variants in the testis-determining region of the Y chromosome in individuals with a 46,XY female phenotype. Hum. Genet. 99:648-652
Verma RS, Dosik H, Scharf T, Lubs HA (1978) Length heteromorphisms of fluorescent (f) and non-fluorescent (nf) segments of human Y chromosome: classification, frequencies, and incidence in normal Caucasians. Journal Of Medical Genetics 15:277-281
Vidalpuig A, Moller DE (1994) Comparative sensitivity of alternative single-strand conformational polymorphism methods. Biotechniques 17:490
Vogt PH, Affara N, Davey P, Hammer M, Jobling MA, Lau YF-C, Mitchell M, et al (1997) Report of the third international workshop on Y chromosome mapping 1997. Cytogenet. Cell Genet. 79:2-16

Vogt PH, Edelmann A, Kirsch S, Henegariu O, Hirschmann P, Kiesewetter F, Köhn FM, et al (1996) Human Y chromosome azoospermia factors (AZF) mapped to different subregions in Yq11. Hum. Mol. Genet. 5:933-943

Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, et al (1998) Large-scale identification, mapping, and genotyping of single- nucleotide polymorphisms in the human genome. Science 280:1077-1082

Weber JL, Wong C (1993) Mutation of human short tandem repeats. Hum. Mol. Genet. 3:1123-1128

Weisler MI, Kirch PV (1996) Interisland and interarchipelago transfer of stone tools in prehistoric Polynesia. PNAS 93:1381-1385

Weisler MI, Woodhead JD (1995) Basalt Pb isotope analysis and the prehistoric settlement of Polynesia. PNAS 92:1881-1885

Weiss K (1973) American Antiquity 38:1-186

White PS, Tatum OL, Deaven LL, Longmire JL (1999) New, male-specific microsatellite markers from the human Y chromosome. Genomics 57:433-437

Whitfield LS, Sulston JE, Goodfellow PN (1995) Sequence variation of the human Y chromosome. Nature 378:379-380

Wilson IJ, Balding DJ (1998) Genealogical inference from microsatellite data. Genetics 150:499-510

Womble WH (1951) Differential systematics. Science 114:315-322

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

Wright S (1969) Evolution and the genetics of populations. University of Chicago Press, Chicago

Yi HF, Donohoe SJ, Klein DC, McBride OW (1993) Localization of the hydroxyindole-O-methyltransferase gene to the pseudoautosaomal region. Hum. Mol. Genet. 2:127-131

Zerjal T, Dashnyam B, Pandya A, Kayser M, Roewer L, Santos FR, Schiefenhövel W, et al (1997) Genetic relationships of Asians and northern Europeans revealed by Y-chromosomal DNA analysis. Am. J. Hum. Genet. 60:1174-1183

Zerjal T, Dashnyam B, Pandya A, Kayser M, Roewer L, Santos FR, Schiefenhövel W, et al (1997) Genetic relationships of Asians and northern Europeans, revealed by Y-chromosomal DNA analysis. Am. J. Hum. Genet. 60:1174-1183

Zhivotovsky LA, Feldman MW (1995) Microsatellite variability and genetic distances. Proc. Natl. Acad. Sci. USA 92:11549-11552

Appendix A

Sequence of the 50f2 (P) amplicon containing the MEH1 polymorphism

TEKB primer ---15bp----→TGGTAACTTT TAGTGATTTC TGATGCTATG TTTATTTCTA GTAGGCACTG TCTTTTCTGG GAGCAACCTG CATTTTGGAT GCTAGTGCAC ACATGCGTTT CGATTTTACC TCTGAGAGGA GGCAAAATTT ATTCTCAATG CAGTAGATGC CCATCCATAA AGTAAAAGTT CTAATTGAGT TGCATGTGGG TAGGCTTGTG TTCCAGTTGC TGACAGGCCT CTCTCTCCCA CACTACCCCC AAGTGACATG CAGAAGGAAC GCTGCCAATG CCATTTTCCA GTAGAATAA($^{C}/_{G}$) GAGATCATGT TTCACAGTTC TGAGTTTATA TGGAGGTTTT GTGCCTTTCA ACTTATGAGA ACCTCATCTT GCACCCTTAG TTTGGAATCA GAACCCATGA TATCACCCTG CAAAGTGTAA ATTTGTGTTG AAAAACAGTA AGTCCACCTT GGCTTCAGCT CTAGCTTACT GCATGGGCTT TATGATCCTG CTTCATATCC CATGTGTGAT TTTTTTGAGT TCTATTAAAA CTATTAATAT TTAATAGTTT TCATTTACAC AGCATCCTTA GGTTTTTGGA GGAGAAGTTC ACTCTGCTGT GTCAAATGGA CATGTTTACA TTTCCAGGGT CTTCC***CAAAT CTGTGGGA-end of the complementary sequence to the TEKA primer.***

Both alleles of the polymorphic base is indicated in bold

Appendix B

Excel worksheets used for dating lineages using intra-lineage microsatellite and MSY1 diversity.

Description of the 6 worksheets that constitute the workbook

1. Input data and calculate differences at each locus for each haplotype compared to a common reference haplotype given at the top of the table

2. Pairwise distance matrix between all haplotypes, including the sum the mutational distance between a single haplotype and all others within the lineage, for each haplotype.

3. Calculation of the lineage age by the 'proportion of mutants' method, the minimum sum of pairs value from sheet 2 is chosen automatically.

4. Calculation of the lineage age by the 'variance' method, variance is calcualted from sheet 1.

5. Calculation of the lineage age by the 'average squared distance' method. The ASD value is inputed manually having been calcuated using the Microsat program.

6. Summary of the ages obtained with confidence limits.

NB. For MSY1 dating the workbook is modified such that each repeat block is analogous to a microsatellite locus. It also includes weighting for repeat number for individual block mutation rates.

Size of Allele 0	15	16	14	22	10	11	12
Chromosomes	*DYS19*	*DYS389A*	*DYS389B*	*DYS390*	*DYS391*	*DYS392*	*DYS393*
1	0	0	0	0	0	0	0
2	0	0	0	1	0	0	0
3	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0
11	-1	0	0	0	0	0	0
12	-1	0	0	0	0	0	0
13	0	0	0	0	0	0	0
14	-1	0	0	0	0	0	0
16	-1	0	0	0	0	0	0
17	0	0	0	0	0	0	0
19	-1	0	0	0	0	0	0
20	-1	0	0	0	0	0	0
23	-1	0	0	0	0	0	0
28	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0
33	0	0	0	0	0	0	0
36	0	0	0	0	0	0	0
41	0	0	0	1	0	0	0
45	0	0	0	0	0	0	0
47	0	0	0	0	0	0	0
48	0	0	0	0	0	0	0
49	0	0	0	0	0	0	0
51	0	0	0	0	0	0	0
52	0	0	0	-1	0	0	0

DYS19	*DYS389A*	*DYS389B*	*DYS390*	*DYS391*	*DYS392*	*DYS393*
15	16	14	22	10	11	12
15	16	14	23	10	11	12
15	16	14	22	10	11	12
15	16	14	22	10	11	12
15	16	14	22	10	11	12
15	16	14	22	10	11	12
15	16	14	22	10	11	12
15	16	14	22	10	11	12
14	16	14	22	10	11	12
14	16	14	22	10	11	12
15	16	14	22	10	11	12
14	16	14	22	10	11	12
14	16	14	22	10	11	12
15	16	14	22	10	11	12
14	16	14	22	10	11	12
14	16	14	22	10	11	12
14	16	14	22	10	11	12
15	16	14	22	10	11	12
15	16	14	22	10	11	12
15	16	14	22	10	11	12
15	16	14	22	10	11	12
15	16	14	22	10	11	12
15	16	14	22	10	11	12
15	16	14	23	10	11	12
15	16	14	22	10	11	12
15	16	14	22	10	11	12
15	16	14	22	10	11	12
15	16	14	22	10	11	12
15	16	14	22	10	11	12
15	16	14	21	10	11	12

Worksheet 2

	1	2	3	4	5	6	7	8	11	12	13	14	16	17	19	20	23	28	29	31	32	33	36	41	45	47	48	49	51	52
1																														
2	1																													
3	0	1																												
4	0	1	0																											
5	0	1	0	0																										
6	0	1	0	0	0																									
7	0	1	0	0	0	0																								
8	0	1	0	0	0	0	0																							
11	1	2	1	1	1	1	1	1																						
12	1	2	1	1	1	1	1	1	0																					
13	0	1	0	0	0	0	0	0	1	1																				
14	1	2	1	1	1	1	1	1	0	0	1																			
16	1	2	1	1	1	1	1	1	0	0	1	0																		
17	0	1	0	0	0	0	0	0	1	1	0	1	1																	
19	1	2	1	1	1	1	1	1	0	0	1	0	0	1																
20	1	2	1	1	1	1	1	1	0	0	1	0	0	1	0															
23	1	2	1	1	1	1	1	1	0	0	1	0	0	1	0	0														
28	0	1	0	0	0	0	0	0	1	1	0	1	1	0	1	1	1													
29	0	1	0	0	0	0	0	0	1	1	0	1	1	0	1	1	1	0												
31	0	1	0	0	0	0	0	0	1	1	0	1	1	0	1	1	1	0	0											
32	0	1	0	0	0	0	0	0	1	1	0	1	1	0	1	1	1	0	0	0										
33	0	1	0	0	0	0	0	0	1	1	0	1	1	0	1	1	1	0	0	0	0									
36	0	1	0	0	0	0	0	0	1	1	0	1	1	0	1	1	1	0	0	0	0	0								
41	1	0	1	1	1	1	1	1	2	2	1	2	2	1	2	2	2	1	1	1	1	1	1							
45	0	1	0	0	0	0	0	0	1	1	0	1	1	0	1	1	1	0	0	0	0	0	0	1						
47	0	1	0	0	0	0	0	0	1	1	0	1	1	0	1	1	1	0	0	0	0	0	0	1	0					
48	0	1	0	0	0	0	0	0	1	1	0	1	1	0	1	1	1	0	0	0	0	0	0	1	0	0				
49	0	1	0	0	0	0	0	0	1	1	0	1	1	0	1	1	1	0	0	0	0	0	0	1	0	0	0			
51	0	1	0	0	0	0	0	0	1	1	0	1	1	0	1	1	1	0	0	0	0	0	0	1	0	0	0	0		
52	1	2	1	1	1	1	1	1	2	2	1	2	2	1	2	2	2	1	1	1	1	1	1	2	1	1	1	1	1	
sum of pairs	10	36	10	10	10	10	10	10	26	26	10	26	26	10	26	26	26	10	10	10	10	10	10	36	10	10	10	10	10	38

The 'proportion of mutants' method for dating

Based on the average number of accumulated mutational steps from an ancestral haplotype

λ/μ = t (in generations) μ = mutation rate (Haplotype= usat haplotype = MSY1 MVR code)

where λ is the sum of the number of mutations between the ancestral haplotype and the haplotype on every other chromosome in the data set divided by the number of loci assayed which equals the number of loci per chromosome multiplied by the number of chromosomes.

The root is chosen from the pairwise differences table as the haplotype that has the least number of mutational steps from all other chromosomes. Alternatively an ancestral haplotype can be constructed from the mode allele length at each locus and the closest haplotype to this used as a root. An outgroup can be chosen from an ancestral haplogroup, however there are likely to be many equally parsimonious links between the two and any single most parsimonious link is likely to be purely a function of the sample size.

root pairwise sum (X)	μ	μ lower 95% limit	μ upper 95% limit	No. of loci (l)	No. of chr. (n)	λ (x/nl)	generation time
10	**0.0021**	**0.0006**	**0.0049**	7	30	0.04761905	20

(Mutation rates from Heyer et al. 1997)

Time to MRCA	22.675737 generations	454 years
upper 95% limit	79.3650794 generations	1587 years
lower 95% limit	9.71817298 generations	194 years

Variance dating

Based on the time required to accumulate the observed variance from a time when the variance was zero, in other words when a haplogroup was founded by a single chromosome.

Firstly the possibility that the diversity is at mutation drift equilibrium ($V\hat{}$) must be discounted by seeing if the observed diversity ($V_r(t)$)is significantly different from that expected at mutation drift equilibrium by using the equations for 95% confidence limits (V_1 and V_2).

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

$V_r(t)=V_0(1-1/N_e)t+(N_e-1)[1-(1-1/N_e)t]\mu$ which rearranges to $t= \ln(1-V_r(t)/V\hat{})/\ln(1-1/N_e)$ (when $V_0=0$)

$V_1=\exp\{\ln V_r(t)-V_g(t)/(2IV_r(t)^2)-(2/V_r(t))*(V_g(t)/I)^{0.5}\}$

$V_2=\exp\{\ln V_r(t)-V_g(t)/(2IV_r(t)^2)+(2/V_r(t))*(V_g(t)/I)^{0.5}\}$

$V_g(t^*)=1/5[(V\hat{}+12V\hat{}^2)(1-e^{-t^*})-1/6(V\hat{}+2V\hat{}^2)(1-e^{-6t^*})+2/5V\hat{}2e^{-t^*}(1-e^{-5t^*})-12V\hat{}^2t^*e^{-t^*}]-V\hat{}^2(1-e^{-t^*})^2$
where $t^*= t/N$

Dating

$V_r(t) = (\Sigma V_l)/I$

Calculate the average variance from the variance of each microsatellite divided by the number of microsatellites

μsat, l	DYS19	DYS389I	DYS389II	DYS390	DYS391	DYS392	DSY393	$V_r(t)$
Variance, V_l	0.18506	0	0	0.1022989	0	0	0	0.041050903

N_e	μ	μ upper limit	μ lower limit	I	$V\hat{}$	$V\hat{}_1$	$V\hat{}_2$	$V_g(t^*)$ for $V\hat{}$	G_t
4900	0.0021	0.0049	0.0006	7	10.2879	5.332949248	18.08099653	69.02911477	20

To calculate the 95% confidence limits around $V\hat{}$ use the time to the MRCA of all Y chromosomes as 200000 years

therefore $t^*(V\hat{})=$ 2.040816327

If $V_r(t)$ is outside the 95% confidence limits of $V\hat{}$ ($V\hat{}_1$ and $V\hat{}_2$) then proceed with dating.

Time to MRCA (t) equals 19.589153 generations 391.78305 years

Confidence Limits

To calculate the confidence limits we must factor the mutation rate limits into the equations for V_1 and V_2

For upper limit:

μ=	0.0006						
$V\hat{}$=	2.9394						
t=	68.90745425	and t^*=	0.014062746				
$V_g(t^*)$=	0.000297						
$V_{1=}$	0.029514311	therefore	t=	49.44418906	generations		
$V_{2=}$	0.055677318	therefore	t=	**93.695203**	generations	**1873.90406**	years

For lower limit:

μ=	0.0049						
$V\hat{}$=	24.0051						
t=	8.3857626	and t^*=	0.00171138				
$V_g(t^*)$=	3.69339E-05						
V_1	0.036647191	therefore	t=	**7.4854971**	generations	**149.7099419**	years
V_2	0.045840038	therefore	t=	9.365009139	generations		

The age of this haplogroup is 392 years 95% limits 150 to 1874

ASD Dating

The Average Squared Distance (ASD) is calculated between a population of chromosomes belonging to a given haplogroup and a root haplotype using the Microsat program from Eric Minch's web site.

The root haplotype is assigned by the same rules as in the proportion of mutants dating. In other words by considering if the haplotype given by the mode/median alleles at each locus is the same as that which gives the smallest sum of pairs value.

The age of the haplogroup is related to the ASD by the simple relationship below

$$ASD=\mu t$$

Therefore:

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

ASD Value	μ	μ lower 95% limit	μ upper 95% limit	generation time
0.048	**0.0021**	**0.0006**	**0.0049**	20

Time to MRCA	22.85714286 generations	457 years
upper 95% limit	80 generations	1600 years
lower 95% limit	9.795918367 generations	196 years

Summary of microsatellite dating calculations for the age of Gypsy group 2.1

Method	Age (years)	95% Range		
ASD	**457**	196	to	1600
Prop. of mutants	**454**	194	to	1587
Variance	**392**	150	to	1874

Appendix C

Papers published from work done in the completion of this thesis.

1. Hurles *et al.*, 1998

2. Hurles *et al.*, 1999

Am. J. Hum. Genet. 63:1793–1806, 1998

European Y-Chromosomal Lineages in Polynesians: A Contrast to the Population Structure Revealed by mtDNA

Matthew E. Hurles,[1] Catherine Irven,[2] Jayne Nicholson,[2] Paul G. Taylor,[1] Fabricio R. Santos,[3,*] John Loughlin,[2] Mark A. Jobling,[1] and Bryan C. Sykes[2]

[1]Department of Genetics, University of Leicester, Leicester; and [2]Cellular Genetics Group, Institute of Molecular Medicine, and [3]CRC Chromosome Molecular Biology Group, Department of Biochemistry, University of Oxford, Oxford

Summary

We have used Y-chromosomal polymorphisms to trace paternal lineages in Polynesians by use of samples previously typed for mtDNA variants. A genealogical approach utilizing hierarchical analysis of eight rare-event biallelic polymorphisms, seven microsatellite loci, and internal structural analysis of the hypervariable minisatellite, MSY1, has been used to define three major paternal-lineage clusters in Polynesians. Two of these clusters, both defined by novel MSY1 modular structures and representing 55% of the Polynesians studied, are also found in coastal Papua New Guinea. Reduced Polynesian diversity, relative to that in Melanesians, is illustrated by the presence of several examples of identical MSY1 codes and microsatellite haplotypes within these lineage clusters in Polynesians. The complete lack of Y chromosomes having the M4 base substitution in Polynesians, despite their prevalence (64%) in Melanesians, may also be a result of the multiple bottleneck events during the colonization of this region of the world. The origin of the M4 mutation has been dated by use of two independent methods based on microsatellite-haplotype and minisatellite-code diversity. Because of the wide confidence limits on the mutation rates of these loci, the M4 mutation cannot be conclusively dated relative to the colonization of Polynesia, 3,000 years ago. The other major lineage cluster found in Polynesians, defined by a base substitution at the 92R7 locus, represents 27% of the Polynesians studied and, most probably, originates in Europe. This is the first Y-chromosomal evidence of major European admixture with indigenous Polynesian populations and contrasts sharply with the picture given by mtDNA evidence.

Received May 18, 1998; accepted for publication September 22, 1998; electronically published December 2, 1998.

Address for correspondence and reprints: Dr. Mark A. Jobling, Department of Genetics, University of Leicester, University Road, Leicester, LE1 7RH, UK. E-mail: maj4@le.ac.uk

* Present affiliation: Departimento de Biologia Geral, Instituto Ciências Biologicas/Universidade Federal de Minas Gerais, Minas Gerais, Belo Horizonte, Brazil.

© 1998 by The American Society of Human Genetics. All rights reserved.
0002-9297/98/6306-0026$02.00

Introduction

The Polynesians were skillful ocean-going navigators at a time when the Greeks and Romans were little more than coastal sailors. Prior to 1500 A.D. they constituted the most geographically widespread people on Earth (Bellwood 1987); their remarkable settlement of the remote Pacific islands has been extensively studied from the viewpoint of the archaeological record and cultural and linguistic affiliations (Jennings 1979). The Pacific islands have been traditionally classified into three geographical areas. Melanesia includes Papua New Guinea and islands to the east, including the Solomon Islands and Vanuatu. To the north of Papua New Guinea, the dispersed archipelagos of Micronesia stretch from Palau in the west to Kiribati in the east. The islands to the east of Fiji constitute Polynesia, a vast ocean triangle with sides ~6,500 km in length and apices at Hawaii, Aotearoa (New Zealand), and Rapanui (Easter Island).

Archaeology, anthropology, and linguistics (Bellwood 1987, 1989, 1991) have been used to construct the following picture of the region's prehistory. Approximately 30,000–50,000 years before the present (YBP), australoid hunter-gatherers moved through Southeast Asia into the Sahul landmass, which comprised the present islands of Papua New Guinea and Australia. These peoples are thought to be ancestral to the extant populations of native Australians and Papuan highlanders. The second major prehistoric migration into the region is associated with the expansion of a mongoloid Austronesian-speaking population, which began 5,000–6,000 YBP in Taiwan and coastal Southeast China. Over a relatively short period, their highly developed navigational skills allowed them to settle the islands of Melanesia, beyond the Solomon Islands, that had not previously been colonized by the australoid population. By 3,000 YBP, Fiji in eastern Melanesia and Samoa and Tonga in western Polynesia had been reached. This initial colonization of island Melanesia is associated with the Lapita culture, characterized by its distinctive pottery, and was quickly followed by the migration of other Melanesian populations that had acquired ocean-going skills. Colonization of remote Oceania followed a lengthy period of

adaptation in western Polynesia, during which a distinct Polynesian culture developed. The Marquesas were reached by 300 A.D., followed by Hawaii (by 500 A.D.), Rapanui (by 800 A.D.), and, finally, Aotearoa (by 1200 A.D.).

Linking of the proto-Polynesians to the Lapita culture, whose sites are found in island Melanesia, suggests relatively little genetic contribution of australoid peoples to Polynesians. Rather, it has been thought that the Polynesians had their origins in the relatively rapid maritime expansion from Southeast Asia (Bellwood 1979). An alternative model postulates that the proto-Polynesians evolved within Melanesia, from the resident population that had been there for ≥30,000 years (Terrell 1986). A third model proposes a substantial colonization of Polynesia from the Americas (Heyerdahl 1950). In addition, during the past 300 years there has been substantial contact with Europeans, who may have contributed genetically to the extant population. Analysis of classic nuclear-encoded markers in Polynesians has weakly supported a Southeast Asian origin with perhaps some Melanesian admixture (Hill and Serjeantson 1989), a contribution that has been further supported by the discovery, in Polynesians, of a specific thalassemia allele of Melanesian origin (Hill et al. 1985). There has been no strong support for an American origin, although data from classic loci were never able to exclude this possibility because Native Americans themselves have an origin in Asia.

In contrast to these uncertainties, studies utilizing mtDNA have been particularly informative (Lum et al. 1994; Melton et al. 1995; Redd et al. 1995; Sykes et al. 1995). A common lineage cluster comprising three or four characteristic base substitutions in the control region and a 9-bp deletion elsewhere in the mitochondrial genome has been found in 94% of all Polynesian mtDNA samples. It has also been found at moderate frequency in coastal Papua New Guinea. Ancestral haplotypes have been traced to Indonesia, the Philippines, and Taiwan (Melton et al. 1995; Sykes et al. 1995). Other maternal lineages identified in Polynesians (Lum et al. 1994) have been confirmed as showing a Melanesian australoid admixture of ~4% (Sykes et al. 1995). A common finding in all genetic studies has been a striking lack of diversity in Polynesians, compared with source populations in Melanesia and Southeast Asia (Flint et al. 1989; Lum et al. 1994; Sykes et al. 1995). Together with a cline of diversity within mitochondrial lineage groups, from high in western Polynesia to low in eastern Polynesia (Sykes et al. 1995), this suggests that there have been, not surprisingly, severe population bottlenecks during the colonization of Polynesia. There is no mtDNA evidence for a substantial input from either the Americas or Europe.

Most of the paternally inherited Y chromosome is haploid and thus escapes recombination. Mutations on this portion of the chromosome represent a simple record of its evolution, which can be used to address questions of human population structure (Jobling and Tyler-Smith 1995). Modern Y chromosomes coalesce back to a common ancestor who has been dated, in independent studies, to 188,000 YBP (Hammer 1995), 37,000–49,000 YBP (Whitfield et al. 1995), or ~170,000 YBP (Underhill et al. 1997). Mating practices, the cultural phenomenon of patrilocality, and the small effective population size of the Y chromosome result in a high degree of geographical differentiation, which has been utilized to investigate prehistoric migration events (Underhill et al. 1997; Zerjal et al. 1997). It is the presence of different Y-chromosomal lineage clusters in Southeast Asia, the Americas, and Europe that allows us to investigate the origins of Polynesian Y chromosomes. The Y chromosome is likely to be more sensitive than other loci to certain kinds of admixture—for example, recent male-dominated admixture between populations normally separated by geographical distance.

The Y chromosome contains a wealth of different polymorphic systems with different mutational mechanisms and rates that vary from ~5×10^{-7}/locus/generation, for base substitutions (Hammer 1995), to a few percent per generation, for the minisatellite MSY1 (Jobling et al. 1998). Initial attempts to use the Y chromosome for regional evolutionary studies were hampered by a lack of well-characterized polymorphisms (Spurdle and Jenkins 1992; Spurdle et al. 1994). Recently, however, a number of studies have successfully addressed questions of regional prehistory by use of the Y chromosome (Underhill et al. 1997; Zerjal et al. 1997). It has been suggested that the best way to utilize the information content from the different polymorphic systems is to adopt a genealogical approach based on a hierarchical analysis of different marker systems (Jobling and Tyler-Smith 1995; Santos and Tyler-Smith 1996; de Knijff et al. 1997). This entails subdivision of sets of chromosomes into distinct lineage clusters defined by compound haplotypes (haplogroups) of rare-event biallelic polymorphisms, followed by assaying of diversity within such haplogroups by use of more variable loci such as mini- and microsatellites. This provides information about the demographic history of haplogroups while minimizing the effect of recurrent mutation in the multiallelic polymorphic systems.

This study complements the mtDNA analysis, by examining the paternal lineages of two populations, one from coastal Papua New Guinea and the other from Polynesia (Cook Islands), by use of Y-chromosomal markers that can be assayed by PCR. Thirty of the 33 Cook Islands samples have the common Polynesian mtDNA haplotype characterized by control-region transitions at positions 16189, 16217, 16247, and 16261.

Seven base substitutions and one insertion/deletion were used to distinguish 10 possible Y-chromosome haplogroups, and diversity within observed haplogroups was assayed by use of seven microsatellite loci and the minisatellite MSY1.

A well-documented and well-dated history suggesting simple population movements in Oceania has been drawn from studies of achaeology, linguistics, and maternal lineages; Oceania thus represents an ideal region for studies of paternal lineages. This is the first Y-chromosomal study to reveal substantial European admixture within the Polynesian Y-chromosomal pool.

Subjects and Methods

Subjects

The DNA samples used in this study were provided by 91 individuals from two locations in the Pacific, all of whom had agreed to take part in a genetic survey. Polynesian samples came from Rarotonga in the Cook Islands in central Polynesia. Melanesian samples were from Port Moresby in Papua New Guinea.

Biallelic-Polymorphism Typing

All of these polymorphisms can be typed by use of similar PCR protocols. Although the cycling programs differ, the reaction volume (10 μl) and composition are the same. Samples (10–20 ng) of genomic DNA were added to a PCR buffer, as described by Jeffreys et al. (1990), with the addition of 1 μM of each primer and 0.5 U of *Taq* polymerase (Advanced Biotechnologies).

All PCR reactions were done in 96-well Thermowell M microtiter plates in an MJR PTC-200 machine. For RFLP analysis, 1 U of the appropriate restriction enzyme in 10 μl of 2 × digestion buffer was added directly to the PCR reaction and incubated at the appropriate temperature for 2 h. All 96 digest/PCR products were run out on a single agarose gel (1%–3%; Seakem agarose [FMC]), in 1 × Tris-borate EDTA (TBE), by means of gel tanks designed and made in house. All pipetting operations were performed with 12-channel pipettes.

A Y *Alu* polymorphism (YAP) was typed according to the procedure of Hammer and Horai (1995); SRY-1532 (identical to SRY 10,831 of Whitfield et al. 1995) was typed according to the procedure of Kwok et al. (1996); and SRY-2627 was typed according to the procedure of Veitia et al. (1997), who referred to it as "SRY-2628." *DYS199* was typed by use of the PCR primers 5′-TAATCAGTCTCCTCCCAGCA-3′ and 5′-AGGTACCAGCTCTTCCCAATT-3′; a cycling program of 94°C for 20 s, 59°C for 20 s, and 72°C for 30 s, repeated 36 times; and the restriction enzyme *Mfe*I. SRY-3225 (identical to SRY 9,138 of Whitfield et al. 1995) was typed by use of the primers 5′-CAACTGTTGAGAAATAGTCATC-3′ and 5′-CCCAGATGCATATATTACAGG-3′; a cycling program of 94°C for 30 s, 58°C for 30 s, and 72°C for 60 s, repeated 34 times; and the restriction enzyme *Hae*III. 92R7 (Mathias et al. 1994) was also typed by use of a PCR-RFLP assay (M. E. Hurles and C. Tyler-Smith, unpublished data). The M4 amplicon (*DYS234*), which has been shown to contain a site polymorphic in Oceanic people (Underhill et al. 1997), was typed by PCR-RFLP analysis, with PCR conditions of 94°C for 30 s, 60°C for 30 s, and 72°C for 45 s, repeated 34 times, and *Nde*I digestion. M9 was amplified by use of the conditions described by Underhill et al. (1997), and the polymorphic site was assayed by digestion with *Hin*fI.

The polymorphism SRY-2627 was typed only on those chromosomes that had been shown to carry the 92R7 (1) allele, since this polymorphism has been shown to occur only on this chromosomal background (M. E. Hurles, unpublished data). Polymorphisms known to occur only on a YAP^+ background were not typed in the sample set, since it contained no YAP^+ chromosomes.

Y chromosomes having the Tat polymorphism (Zerjal et al. 1997) constitute a subset of 50f2/C-deletion chromosomes (Jobling et al. 1996). 50f2/C-deletion chromosomes are identified by the absence of the 203-bp MSY1 homologue, which is coamplified in the MSY1-flanking PCR prior to minisatellite variant repeat–PCR (MVR-PCR) (Jobling et al. 1998). No such deletion chromosomes were detected in the MSY1-flanking reactions, and, therefore, the Tat polymorphism itself was not typed in these samples.

The allelic states of each haplogroup are as follows (in the order *DYS199*, YAP, SRY-3225, SRY-1532, 92R7, M4, SRY-2627, and M9): haplogroup 1 chromosomes have the compound haplotype 00011001, those of haplogroup 2 have 00010000, those of haplogroup 3 have 00001001, those of haplogroup 24 have 00010101, and those of haplogroup 26 have 00010001. Binary nomenclature indicates ancestral (0) and derived (1) forms of the polymorphism: 92R7, YAP (Jobling and Tyler-Smith 1995), SRY-1532, SRY-3225 (Whitfield et al. 1995), *DYS199* (Underhill et al. 1996), M4, M9 (Underhill et al. 1997), and SRY-2627 (Veitia et al. 1997).

Microsatellite Typing

The following seven highly polymorphic Y chromosome–specific microsatellites were analyzed: *DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392,* and *DYS393;* all have tetranucleotide-repeat units, except for *DYS392,* which has a trinucleotide-repeat unit. Primer sequences are as described elsewhere (de Knijff et al. 1997), except for the forward primer of *DYS389II,* which was 5′-TCCAACTCTCATCTGTATTATCTATGTG-3′. PCR reactions were performed in a 15-μl re-

action volume containing 6 pmol of each primer, 200 μM of each dNTP, 1× GeneAmp PCR buffer II, 1.5 mM $MgCl_2$, 0.375 U of Ampli*Taq* Gold (PE Applied Biosystems), and 50 ng of DNA. The forward primer was 5′-labeled with tetrachlorofluoroscein (*DYS19, DYS389I,* and *DYS389II*), carboxyfluoroscein (*DYS390, DYS391,* and *DYS393*), or 4,7,2′,4′,5′,7′-hexachloro-6-carboxyfluoroscein (*DYS392*). The general PCR profile was as follows: 94°C for 14 min, annealing temperature (AT) for 1 min, 72°C for 0.5 min, and 94°C for 1 min, with the latter three steps repeated 35 times; AT for 1 min; and 72°C for 10 min. Annealing temperatures were as follows: 54°C for *DYS19,* 51°C for *DYS389I,* 55°C for *DYS389II,* 56°C for *DYS390,* 55°C for *DYS391,* 53.5°C for *DYS392,* and 56°C for *DYS393.*

The PCR products from all seven individual reactions were diluted and pooled together with a size-standard GeneScan-500 TAMRA, so that an individual's seven-locus microsatellite haplotype could be read from a single gel track (Reed et al. 1994). The pooled products were run on denaturing 6% acrylamide gels in 1 × TBE on an Applied Biosystems 373A DNA sequencer. Gels were analyzed by ABI PRISM GeneScan Analysis 2.0.2, and samples were genotyped by Genotyper 1.1 (both from PE Applied Biosystems). By means of GAS (genetic-analysis system [A. Young, personal communication]), the discrete distributions of allele lengths were then put into allele bins, which were assigned numbers according to their sizes, with the smallest allele being denoted "1." To account for the *DYS389I* variable array being contained inside the *DYS389II* amplicon (Cooper et al. 1996), the length of the *DYS389I* product was subtracted from that of *DYS389II,* prior to placement in allele bins. The correspondence between our allele definitions and the repeat-unit number of previously published sources (de Knijff et al. 1997) is as follows: allele 1 corresponds to *DYS19,* repeat-unit number 11; *DYS389I,* repeat-unit number 9; *DYS389II,* repeat-unit number 15; *DYS390,* repeat-unit number 20; *DYS391,* repeat-unit number 9; *DYS392,* repeat-unit number 11; and *DYS393,* repeat-unit number 12.

Minisatellite Coding

Three-state MSY1 MVR-PCR of repeat types 1, 3, and 4 was performed according to the method of Jobling et al. (1998). A code—for example, (1)20(3)35(4)20—represents the minisatellite array as blocks of different repeat-unit variants, in this case 20 type 1 repeats are followed by a block of 35 type 3 repeats followed by a block of 20 type 4 repeats. Modular-structure nomenclature of, for example, the form (1,3,4) refers to a block of type 1 repeats followed by a block of type 3 repeats followed by a block of type 4 repeats.

Sequence Analysis of MSY1 Null Repeats

MSY1 alleles containing null repeats were analyzed by fluorescent automated sequencing of PCR products generated by means of the primers $Y1A^+$ and $Y1B^+$.

Phylogenetic Analysis

The haplogroup tree was based on that of Jobling and Tyler-Smith (1995) and used only those polymorphisms that are presently typable by PCR and incorporated the new polymorphisms *DYS199,* SRY-2627, SRY-3225, M9, and M4, by typing them on a fully representative panel of chromosomes previously typed for all polymorphisms in the original tree. Minimum spanning microsatellite networks were constructed for chromosomes belonging to individual haplogroups, under the assumption of a single-step mutation process, by means of pairwise comparisons between all haplotypes. Initially all haplotypes differing by single mutational steps are linked, and then most-parsimonious steps of greater mutational distance are considered, in turn, until all haplotypes are included in the network. All most-parsimonious steps linking any given haplotype into the network are included (Zerjal et al. 1997).

The root for the maximum-parsimony tree used in dating haplogroup 24 (Bertranpetit and Calafell 1996) was chosen by combining the mode allele length of each individual locus into a seven-locus haplotype. This haplotype is identical to haplotype 44, which is itself represented more than once in the data set. The network gives additional support for this haplotype being a plausible root because the latter occupies a central position within the network and has the greatest number of links to other haplotypes.

Results

A total of 33 unrelated Cook Islander and 58 unrelated Papua New Guinean samples were assayed for all polymorphisms. Of the 10 haplogroups defined by the rare-event biallelic polymorphisms, only 5 (i.e., haplogroups 1-3, 24, and 26) were observed in this sample set. Haplogroup 1 and haplogroup 3 chromosomes are found in Polynesians alone, haplogroup 2 and haplogroup 26 chromosomes are found in both Polynesians and Melanesians, and haplogroup 24 chromosomes are found only in Melanesians. Haplogroup 2 is the most ancestral. These data are summarized in figure 1.

A total of 91 samples gave full biallelic-polymorphism data, and 79 of them also had full seven-locus microsatellite haplotypes; of these, 77 gave MSY1 codes. MSY1 subtypes were defined initially on the basis of modular structures of arrays of different repeat types that are represented more than once in the sample set. Three subtypes were defined: (1), (1,3,4); (2), (3,1,3,4);

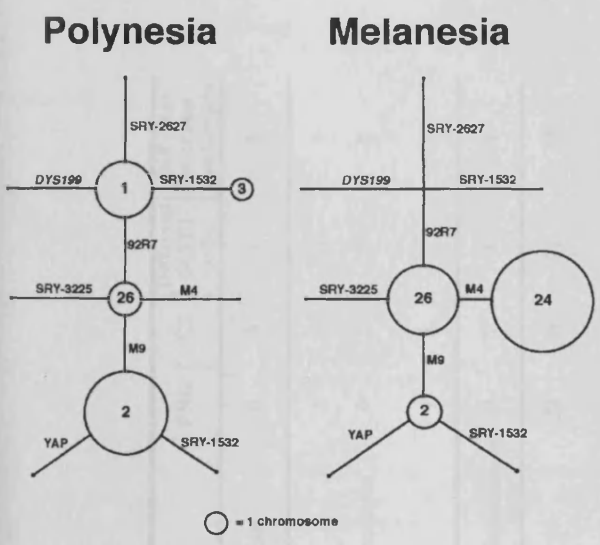


Figure 1 Haplogroup distributions within Polynesians and Melanesians: unrooted trees based on that in the work of Jobling and Tyler-Smith (1995), with new mutations added (see Subjects and Methods). Circles represent haplogroups, and lines represent single mutational steps between them. Lines are labeled with the polymorphisms that they represent. Circle area is proportional to the number of chromosomes within a given haplogroup. Dots at the ends of the lines denote haplogroups not represented in that population. Numbers within circles indicate haplogroup number. All 91 chromosomes with full rare-event biallelic-polymorphism data are included, irrespective of the status of their microsatellite or minisatellite data.

and (3), a group of MSY1 codes that all end with the modular structure (...4,0,4). An additional subtype ($3,1,3^+,4^-$) was defined when nonoverlapping block-size ranges within the (3,1,3,4) chromosomes were shown to be statistically significant ($P < .001$; Student's t-test). An example of a code of the former MSY1 subtype is (3)3(1)13(3)61(4)8, and an example of the latter MSY1 subtype is (3)1(1)13(3)39(4)16. These typing data are summarized in figure 2 and are represented diagrammatically in figure 3. Note that the percentages of haplogroup frequencies given in the text refer to all 91 chromosomes studied, whereas the pie charts in figure 3 consider only those chromosomes for which MSY1 codes were available. Sample CI185 is classified as a (...4,0,4) chromosome by virtue of its microsatellite haplotype, the characteristic null at the start of the array, and the similarity between its repeat-block sizes and those in the other (...4,0,4) chromosomes. The null repeats appear to have been converted back to type 4 repeats in this chromosome, perhaps by a repeat-homogenization process (Bouzekri et al. 1998). Complete typing results, including microsatellite haplotypes, are tabulated in the Appendix (table A1).

Minimum spanning microsatellite networks for each haplogroup were constructed with the 79 samples for which there were full microsatellite- and biallelic-polymorphism data. MSY1 subtypes were then mapped onto these networks. The final networks are shown in figure 4.

It is important to distinguish between MSY1 subtypes that represent very rare mutational events and those that may have independently evolved multiple times. Subtypes (3,1,3,4) and (1,3,4) are common to haplogroups 24 and 26. They have also both been observed in other haplogroups, which were not found in the present study (Jobling et al. 1998). Given that a haplogroup defined by a unique mutational event must have been founded by a single chromosome, these MSY1 subtypes have probably arisen independently several times. In addition, there is, in the microsatellite networks, no tight clustering of chromosomes having these MSY1 subtypes. By contrast, the (...4,0,4) and ($3,1,3^+,4^-$) MSY1 subtypes have been found only in haplogroup 2 and haplogroup 26 chromosomes, respectively, and only in this region of the world, suggesting that each has a unique origin. In addition, the tight clustering of these chromosomes within the microsatellite networks further supports this conclusion.

Haplogroup 1 Chromosomes

Haplogroup 1 chromosomes were found only in Polynesians, in whom they comprised 9 (27%) of 33 chromosomes in the sample. The nine haplogroup 1 chromosomes have eight different microsatellite haplotypes, and all belong to the same MSY1 subtype, (1,3,4).

Haplogroup 2 Chromosomes

Haplogroup 2 chromosomes are found in both Melanesians and Polynesians, in whom they comprise 3 (5%) of 58 and 19 (58%) of 33 of the Y chromosomes, respectively. Although this haplogroup is found in both populations, it can also be found in most areas of the world (C. Tyler-Smith and A. Pandya, personal communication) and thus, on its own, cannot be taken as evidence of recent common ancestry for the two populations. There are 11 different microsatellite haplotypes in the 22 chromosomes analyzed here, and two different MSY1 subtypes—(...4,0,4) and (3,1,3,4)—are represented more than once within this haplogroup. Only one of these MSY1 subtypes, the (...4,0,4) subtype, is found in both Polynesians and Melanesians.

The (...4,0,4) chromosomes found in Melanesians have modular structures different than those found in Polynesian (...4,0,4) chromosomes. The Polynesian chromosomes all have the modular structure (0,1,3,4,0,4), whereas the two (...4,0,4) chromosomes in Melanesians have the modular structures (3,1,3,4,0,4) and (0,3,1,3,4,0,4,0,4). To determine whether these

Haplogroup	MSY1 subtype	MSY1 code range	PNG	CI	Different MSY1 codes	Different microsat haplotypes
1	(1,3,4)	(1)15-17 (3)36-38 (4)18-21	0	9	7	8
2	(...4,0,4)	Variable modular structure	2	15	11	7
	(3,1,3,4)	(3)3 (1)13,14 (3)37 (4)23,24	0	2	2	2
	Others (singletons)	Variable	1	2	3	3
3	(1,3,4) singleton	(1)20 (3)51 (4)18	0	1	1	1
24	(3,1,3,4)	(3)1-2 (1)11-14 (3)31-40 (4)8-19	21	0	21	18
	(1,3,4)	(1)11-14 (3)37-40 (4)14	4	0	3	3
	Others (singletons)	Variable	3	0	3	3
26	(3,1,3,4)	(3)1-3 (1)10-14 (3)37-44 (4)13-16	5	0	5	5
	(3,1,3^{+},4^{-})	(3)3 (1)12-14 (3)60-63 (4)6-8	4	3	5	4
	(1,3,4)	(1)12-15 (3)25-48 (4)12-24	4	0	4	4
	Others (singletons)	Variable	0	1	1	1
		Total	**44**	**33**	**66**	**59**

Figure 2 MSY1 structural diversity within haplogroups defined by rare-event biallelic polymorphisms. Each circle represents a single repeat unit: black circles represent type 3 repeats, white circles represent type 1 repeats, and gray circles represent type 4 repeats. Only the 77 chromosomes with full rare-event, biallelic-polymorphism, microsatellite, and minisatellite data are included. (PNG = Papua New Guineans; CI = Cook Islanders).

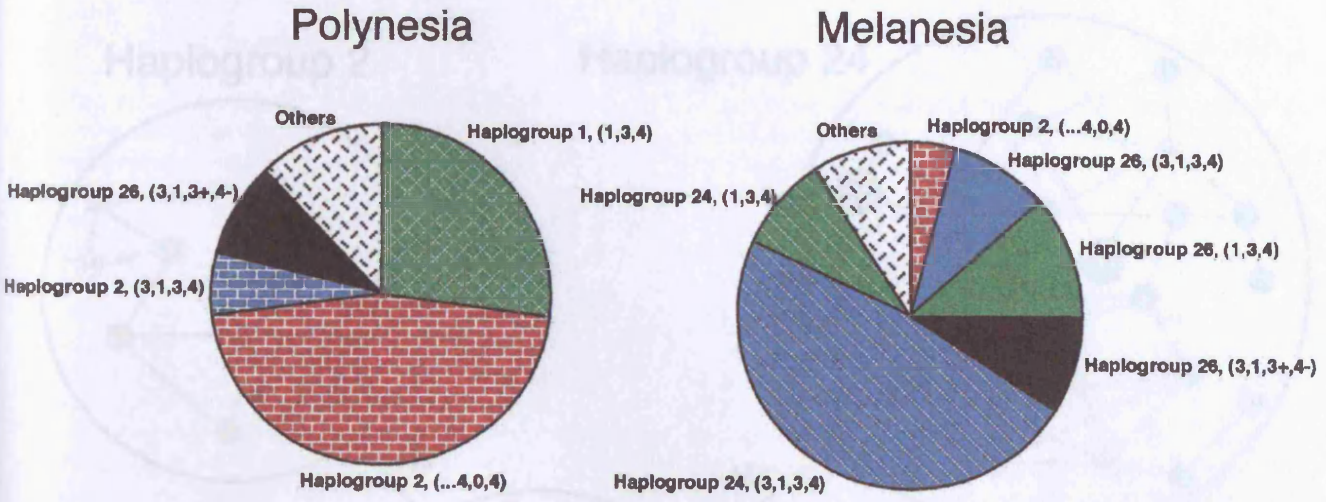


Figure 3 Distribution of MSY1 subtypes within Polynesians and Melanesians. Only the 77 chromosomes for which there are full biallelic-polymorphism and minisatellite-typing data are included. The background pattern represents the haplogroup, and color represents MSY1 subtype. Only those MSY1 subtypes that are represented more than once are displayed. "Others" includes all singleton MSY1 modular structures.

chromosomes do, indeed, belong to the same lineage cluster, the null repeats from a Cook Islander chromosome (CI140) and from a Papua New Guinean chromosome (7092) were sequenced and shown to be identical. The null repeats at the start of these arrays were both shown to be type 3 repeats modified by a T→A transversion at position 21 within the 25-bp repeat unit. The null repeats near the end of the array were found to be type 2 repeats, which are characterized by having a C→G transversion at position 13 within the repeat, relative to type 4 repeats (Jobling et al. 1998). Additional support for the common ancestry of these chromosomes comes from their association with a rare short *DYS390* microsatellite allele that is found almost exclusively in this region of the world (Forster et al. 1998).

The other MSY1 subtype found more than once within haplogroup 2 chromosomes is the (3,1,3,4) subtype, which is found in only two chromosomes (6%) in Polynesians. These two chromosomes, along with a third Polynesian haplogroup 2 chromosome that has neither the (3,1,3,4) nor the (...4,0,4) MSY1 modular structure, cluster together in the microsatellite network, perhaps suggesting a recent common ancestry.

Haplogroup 3 Chromosomes

Only a single haplogroup 3 chromosome has been found in this study—in Polynesians, in whom it is found in 1 (3%) of 33 chromosomes in the sample. Haplogroup 3 chromosomes are found in both Europe and Asia (as haplotype "1D" in the work of Hammer et al. [1998]) and have also been found at low frequency in Indonesians (M. E. Hurles, unpublished data).

Haplogroup 24 Chromosomes

This group of Y chromosomes was found in 37 (64%) of 58 chromosomes in the Papua New Guinean sample but not at all in the Polynesian sample. Previously, haplogroup 24 chromosomes had been found in Papua New Guineans and at low frequency in Indonesians (Underhill et al. 1997; M. E. Hurles, unpublished data). There are 24 different microsatellite haplotypes in the 30 chromosomes that have full microsatellite haplotypes. Two different MSY1 subtypes are found in these chromosomes. By far the most frequent MSY1 subtype is (3,1,3,4), found in 21 of 28 chromosomes with identified MSY1 structure. The (1,3,4) MSY1 subtype is found in four of the remaining seven chromosomes.

The question arises as to whether the mutation defining haplogroup 24 occurred prior to or after the colonization of Polynesia. The absence of such chromosomes in Polynesians might constitute significant evidence of reduced diversity. Consequently, microsatellite diversity was used to date the origin of these chromosomes, so as to compare this date with that of the initial migration into Polynesia, which, on the basis of archaeological evidence, has been deduced to be ~3,000 YBP. Two different dating methods were used.

In the first approach (Bertranpetit and Calafell 1996), a rooted maximum-parsimony tree (not shown) was constructed from the pairwise comparisons between compound microsatellite haplotypes. The root of the tree was placed at haplotype 44, on the basis of the mode-allele length at each locus (see Subjects and Methods section). Haplogroup 26 is ancestral, but it could not be used as an outgroup for the purpose of rooting, because

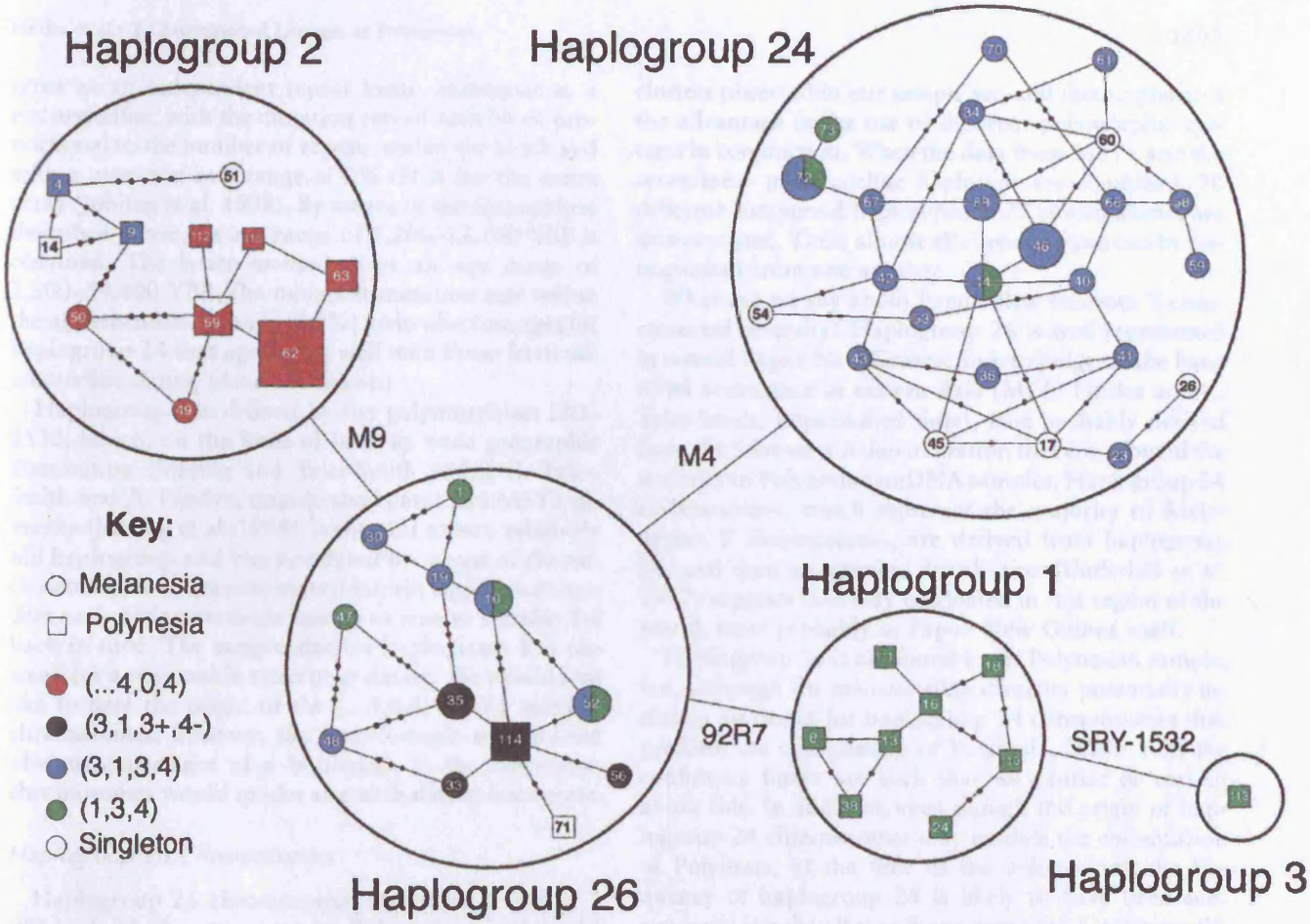


Figure 4 Microsatellite networks for individual haplogroups. Numbered circles and squares represent compound microsatellite haplotypes; small black circles represent intermediate haplotypes not observed in this sample set. Color indicates MSY1 subtype. All 79 chromosomes for which there are complete microsatellite and biallelic-polymorphism data are included.

there are multiple equally parsimonious links between the networks of the two haplogroups. The average number of mutations per locus is calculated from the root to every haplotype in the tree, with consideration of the number of times that that haplotype is represented in the sample. A generation time of 25 years (Thomas et al. 1998) and a mutation rate of 2.1×10^{-3}/locus/generation for Y-chromosome microsatellites (Heyer et al. 1997) was used to convert the calculated figure of 0.38 mutations/locus/generation to a time of origin of ~4,400 YBP for this base substitution. Factoring in the 95% confidence limits, $0.6–4.9 \times 10^{-3}$, of the microsatellite mutation rate gives an interval of 1,900–15,300 YBP (Heyer et al. 1997).

The second approach (Goldstein et al. 1996) relies on the variance of linked microsatellite loci and assumes a single-step–mutation model. As long as the observed variance can be shown to be significantly different from that expected at mutation/drift equilibrium, the time taken for the variance to accumulate since its initial value of zero can be calculated. Given a mutation rate of 2.1×10^{-3}/locus/generation and an effective population size of 4,500 (Goldstein et al. 1996), the variance expected at mutation/drift equilibrium is 9.4, with 95% confidence limits of 4.2–18.4. The variance observed in our sample is 0.45, with 95% confidence limits of 0.36–0.56. This gives a time to the origin of this haplogroup (i.e., when variance was zero) of ~5,500 YBP, with 95% confidence limits of 3,500–5,450 YBP. If the 95% confidence limits for the mutation rate are factored into the equations, an expanded interval of 2,000–30,400 years is obtained. Despite the difference in the various estimates of haplogroup age, it seems likely, on the basis of microsatellite haplotype diversity, that this mutation predates the migration into Polynesia, 3,000 YBP.

Preliminary attempts were also made to use MSY1 diversity for dating, by treating each block of repeat

types as an independent repeat locus, analogous to a microsatellite, with the mutation rate of each block proportional to the number of repeats within the block and with a mutation rate range of 2%–11% for the entire array (Jobling et al. 1998). By means of the first method described above, an age range of 1,200–12,700 YBP is obtained. The latter method gives an age range of 2,300–54,400 YBP. The midpoint mutation rate within the aforementioned range (6.5%) gives absolute ages for haplogroup 24 that agree very well with those from microsatellite dating (data not shown).

Haplogroup 2 is defined by the polymorphism SRY-1532, which, on the basis of both its wide geographic distribution (Jobling and Tyler-Smith 1995; C. Tyler-Smith and A. Pandya, unpublished data) and MSY1 diversity (Jobling et al. 1998), is thought to be a relatively old haplogroup and was not dated by means of the microsatellite data, because mutation/drift equilibrium renders such dating methods unable to resolve suitably far back in time. The sample size for haplogroup 1 is too small for a reasonable attempt at dating. We would also like to date the origin of the (...4,0,4) MSY1 subtype chromosomes; however, the small sample size and the obvious occurrence of a bottleneck in the Polynesian chromosomes would render any such dating inaccurate.

Haplogroup 26 Chromosomes

Haplogroup 26 chromosomes have been found in 3 (9%) of 33 chromosomes in Polynesians and in 13 (22%) of 58 chromosomes in Melanesians. Within this haplogroup, there are three main subgroups—(3,1,3,4), (1,3,4), and ($3,1,3^+,4^-$)—defined by different MSY1 subtypes. Of these subgroups, only the last is shared between the Melanesian and Polynesian populations; the others are found solely in Melanesians. As discussed above, the ($3,1,3^+,4^-$) subtype defines a unique Y-chromosomal lineage. There are three different microsatellite haplotypes among the four ($3,1,3^+,4^-$) chromosomes in Melanesians, whereas all three ($3,1,3^+,4^-$) chromosomes in Polynesians share the same microsatellite haplotype.

Discussion

The present study illustrates the informative capacity of a genealogical approach to Y-chromosomal analysis and is also the first application of MVR-PCR coding of the minisatellite MSY1 to address a specific issue in human evolution. If we consider all 77 chromosomes for which we have obtained complete typing data, there are 59 different microsatellite haplotypes and 66 different MSY1 codes. Although MSY1 modular structures can be good predictors of lineage clusters (Jobling et al. 1998; P. G. Taylor, unpublished data), no single polymorphic system can distinguish between all the lineage clusters observed in our sample set, and this emphasizes the advantage in the use of different polymorphic systems in conjunction. When the data from MSY1 and the seven-locus microsatellite haplotype are combined, 70 different compound haplotypes of 77 chromosomes are differentiated. Thus, almost all chromosomes can be distinguished from one another.

What can we say about Papua New Guinean Y-chromosomal diversity? Haplogroup 26 is well represented in coastal Papua New Guinea, and, to judge on the basis of its abundance in eastern Asia (M. E. Hurles and C. Tyler-Smith, unpublished data), it is probably derived from the Southeast Asian migration that contributed the majority of Polynesian mtDNA samples. Haplogroup 24 chromosomes, which represent the majority of Melanesian Y chromosomes, are derived from haplogroup 26, and their geographic distribution (Underhill et al. 1997) suggests that they originated in this region of the world, most probably in Papua New Guinea itself.

Haplogroup 24 is not found in the Polynesian sample, but, although the microsatellite diversity potentially indicates an origin for haplogroup 24 chromosomes that predates the colonization of Polynesia, 3,000 YBP, the confidence limits are such that we cannot be certain about this. In addition, even though the origin of haplogroup 24 chromosomes may predate the colonization of Polynesia, at the time of the colonization the frequency of haplogroup 24 is likely to have been substantially less than it is at the present time. Consequently, the absence of haplogroup 24 chromosomes from Polynesians cannot be taken as evidence of bottleneck events. However, there is a reduction of diversity, within both the (...4,0,4) and ($3,1,3^+,4^-$) MSY1 subtypes, between Papua New Guineans and Polynesians, and this is most probably due to the multiple population bottleneck events that accompanied the colonization of Polynesia. This picture of reduced diversity in Polynesians compared with Melanesians is common to nearly all genetic studies of the region, independent of the locus studied (Flint et al. 1989; Sykes et al. 1995).

If the blocks of MSY1 repeat-unit variants are considered to be independent loci with mutation rates proportional to their sizes, the use of MSY1 diversity to date haplogroup 24 generates ranges of ages that agree well with those derived from microsatellite data. In the future, greater empirical knowledge about the mutation dynamics of this locus will allow more-sophisticated dating analyses.

What are the possible origins of the three major lineage clusters found in Polynesians? Clearly, the haplogroup 2 chromosomes with the (...4,0,4) MSY1 subtype and the haplogroup 26 chromosomes with the ($3,1,3^+,4^-$) MSY1 subtype found in Polynesians share a recent common origin with those found in Melanesians.

Together, these account for 55% of the Polynesian Y chromosomes in this study.

Haplogroup 1 chromosomes, which comprise 27% of the Polynesian Y chromosomes in this study, are not found in our Melanesian sample. From where have these Y chromosomes come? Haplogroup 1 chromosomes are found at high frequency in Europeans and at low frequency in Asians and peoples of the Americas (Santos and Tyler-Smith 1996; C. Tyler-Smith and A. Pandya, personal communication). Within haplogroup 1, the MSY1 subtype (1,3,4) is found at appreciable frequencies only in Europeans and peoples of the Americas, with Asian chromosomes belonging almost exclusively to the (3,1,3,4) subtype (Jobling et al. 1998). In addition, in a microsatellite network of haplogroup 1 Y chromosomes, the Polynesian Y chromosomes cluster closely with the European and not with the Asian Y chromosomes (M. E. Hurles, unpublished data). Thus, we can discount an Asian origin for these haplogroup 1 chromosomes. The majority (90%) of indigenous Y chromosomes in South Americans have the derived form of the *DYS199* polymorphism (Underhill et al. 1996). Haplogroup 1 chromosomes do not carry this base substitution and, consequently, represent only a small minority of South American Y chromosomes. In contrast, haplogroup 1 chromosomes of the (1,3,4) MSY1 subtype are the most common type of Y chromosome in western Europeans (Jobling et al. 1998), representing approximately two-thirds of all Y chromosomes within this region (Santos and Tyler-Smith 1996; M. E. Hurles, unpublished data). Within our Polynesian sample, we do not observe any Y chromosomes that have the derived form of *DYS199*. Thus, in this study there is no evidence for a Native American contribution to the Polynesian Y-chromosomal pool. It therefore seems likely that the haplogroup 1 chromosomes found in the Polynesians have a recent European origin.

If the haplogroup 1 chromosomes are indeed European in origin, then we should also expect to see in Polynesians some other haplogroups that are found at appreciable frequencies in western European populations. Haplogroup 2 chromosomes belonging to the (3,1,3,4) subtype are the other major type of Y chromosome found in western Europeans, representing approximately one-quarter of all Y chromosomes from this population (Jobling et al. 1998). The Polynesian sample contains two of these chromosomes, and the Melanesian sample contains none.

Thus, the Polynesian sample contains two types of chromosomes that are neither found within the Melanesian sample nor known to be common in Southeast Asians in general (M. E. Hurles, C. Tyler-Smith, and A. Pandya, unpublished data). These two types of Y chromosome represent the most common types found in western Europe. Indeed, in Polynesians these chromosomes are found in approximately the same ratio at which they are present in western European populations, providing additional evidence for recent European admixture. In summary, 55% of Polynesian Y chromosomes can be traced to Melanesians and have Southeast Asian origins, 33% (i.e., haplogroup 1 and [3,1,3,4] haplogroup 2 chromosomes]) appear to be European in origin, and 12% remain of indeterminate origin.

This study illustrates the power of the phylogeographic approach to population-structure analysis using the Y chromosome. The use of microsatellite data alone in Y-chromosome studies of this region of the world will not differentiate between European and genuine Polynesian Y chromosomes. Attempts that investigate paternal relationships within this region of the world (Lum et al. 1998) but that do not take into account this substantial European contribution run the risk of obscuring, with the "noise" of recent admixture, the real patterns of prehistoric population movements and, thus, of potentially drawing spurious conclusions.

This study is also a dramatic example of the advantages of combining the Y-chromosomal results with mtDNA data from the same samples. Comparisons of Y-chromosomal and mtDNA data have previously been used to characterize ethnic introgression within Native American (Bianchi et al. 1997) and African-derived (Bravi et al. 1997) populations. In the Polynesian population studied here, although almost all maternal lineages are derived from native Polynesian ancestors, at least one-third of Y chromosomes are probably of recent European origin. Studies of human leukocyte antigen (HLA) haplotypes in Oceania have identified 5%–10% as being of recent caucasoid admixture (Serjeantson 1989). Analysis of autosomal loci, such as HLA, will always reflect an average of maternal and paternal contributions. By contrast, the current study, in combination with the earlier mtDNA analysis, vividly demonstrates the differential input from males and females. In the case of Polynesians, the predominantly paternal route for European admixture can be explained by the exclusively male composition of the postcontact groups, which included sailors, traders, whalers, and missionaries.

Acknowledgments

We thank Chris Tyler-Smith and Arpita Pandya, for unpublished data on biallelic polymorphisms; Manfred Kayser, for useful discussions; Stuart Bayliss, for automated DNA sequencing; and the MRC Molecular Haematology Unit and John Mitchell, for access to DNA collections. We are grateful to the Prime Minister's Office (Tere Tangiiti) and Health Department (Dr. George Koteka and Vaevae Pare) of the Government of the Cook Islands. This work was supported by the Wellcome Trust and the Medical Research Council. M.A.J. is a Wellcome Trust Career Development Fellow supported by grant 044910.

Appendix

Table A1

Rare-Event Biallelic-, Microsatellite-, and MSY1-Polymorphism Data for All Samples, Classified by Haplogroup and MSY1 Modular Structure

Population and Sample[a]	MSY1 Code	Microsatellite Haplotype	Haplotype Number
Haplogroup 1 (1,3,4):			
CI75A	(1)17(3)36(4)21	2236334	24
CI115	(1)15(3)37(4)20	2225332	16
CI183	(1)16(3)38(4)20	2214232	8
CI188	(1)17(3)38(4)18	2226342	18
CI194	(1)15(3)36(4)21	2224332	13
CI196	(1)16(3)38(4)20	3224232	38
CI120	(1)15(3)39(4)21	2236232	110
CI139	(1)17(3)38(4)18	2226342	18
CI149	(1)15(3)39(4)18	2225331	111
Haplogroup 2 (...4,0,4):			
CI140	(0)1(1)13(3)33(4)6(0)3(4)16	4131223	62
CI147	(0)1(1)13(3)33(4)5(0)4(4)15	4131323	63
CI156	(0)1(1)12(3)35(4)5(0)3(4)17	4231223	69
CI175	(0)1(1)13(3)33(4)6(0)3(4)16	4131223	62
CI180	(0)1(1)13(3)32(4)5(0)4(4)14	4131223	62
CI181	(0)1(1)13(3)33(4)6(0)3(4)16	4131223	62
CI198	(0)1(1)13(3)34(4)5(0)3(4)14	4131223	62
CI206	(0)1(1)13(3)33(4)5(0)4(4)15	4131323	63
CI153	(0)1(1)13(3)33(4)5(0)4(4)16	4131223	62
CI155	(0)1(1)12(3)34(4)5(0)3(4)16	4231223	69
CI186	(0)1(1)13(3)33(4)6(0)3(4)16	4131223	62
CI123	(0)1(1)13(3)33(4)6(0)3(4)16	4131223	62
CI138	(0)1(1)14(3)32(4)6(0)3(4)16	4231323	115
CI167	(0)1(1)13(3) >11 (4)5(0)3(4)16	4232223	112
CI185	(0)1.(1)13.(3)33.(4)25	4131223	62
6591	(3)2(1)12(3)29(4)4(0)5(4)24	3251213	49
7092	(0)1(3)1(1)11(3)29(4)7(0)3(4)9(0)3(4)11	3321212	50
Haplogroup 2 (3,1,3,4):			
CI142	(3)3(1)13(3)37(4)24	2125212	4
CI191	(3)3(1)14(3)37(4)23	2223212	9
Haplogroup 2 (others):			
CI145	(0)1(1)12(3)38(4)11	4231223	69
CI192	(3)3(1)1(1/3?)1(1)5(1/3?)2(1)4(3)28(4)24	2225211	14
14791	(1)14(3/4)21(0)27	3324331	51
Haplogroup 3 (1,3,4):			
CI151	(1)20(3)51(4)18	3234213	113
Haplogroup 24 (3,1,3,4):			
2592	(3)1(1)13(3)36(4)12	4225332	66
2792	(3)1(1)12(3)34(4)15	3335322	53
2892	(3)1(1)13(3)34(4)17	3128232	34
2992	(3)1(1)13(3)40(4)12	4226232	67
3091	(3)1(1)12(3)31(4)16	2236333	23
3092	(3)1(1)14(3)36(4)8	4125332	58
3392	(3)2(1)13(3)33(4)15	3226233	43
3792	(3)2(1)11(3)39(4)15	3215233	36
4292	(3)1(1)13(3)32(4)12	3225333	41
4992	(3)1(1)13(3)35(4)17	4129332	61
5092	(3)1(1)13(3)36(4)19	4226332	68
5492	(3)1(1)12(3)34(4)13	3227332	46
6092	(3)1(1)12(3)37(4)16	3227332	46
8092	(3)1(1)13(3)37(4)12	4125432	59
8392	(3)1(1)13(3)38(4)13	3226232	42

(*continued*)

Table A1 (continued)

Population and Sample[a]	MSY1 Code	Microsatellite Haplotype	Haplotype Number
9391	(3)1(1)11(3)36(4)13	5226232	72
09921	(3)1(1)12(3)34(4)14	4226332	68
12191	(3)2(1)11(3)37(4)15	3226332	44
12591	(3)1(1)12(3)36(4)16	3225332	40
12991	(3)2(1)11(3)35(4)16	4233232	70
13191	(3)1(1)12(3)33(4)13	3227332	46
Haplogroup 24 (1,3,4):			
14991	(1)14(3)37(4)14	3226332	44
6991	(1)12(3)38(4)14	6226232	73
7292	(1)11(3)40(4)14	5226232	72
9192	(1)12(3)38(4)14	5226232	72
Haplogroup 24 (others):			
6192	(3)1(1)13(3)35(4)1(3)1(4)15	4127332	60
8692	(3)1(1)11(0)1(3)32(4)14	2226333	17
14591	(3)1(1)12(3/1)2(3)34(4)16	3336232	54
Haplogroup 24 (MSY1 undetermined):			
9092	x	2245233	26
3691	x	3226334	45
Haplogroup 24 (microsatellites incomplete):			
1892	(3)1(1)12(3)>30(4)11	33x6333	103
4092	(3)1(1)9(3)34(4)16	42x6332	106
4492	(3)1(1)13(3)34(4)16	42x5342	105
4692	(3)1(1)12(3)>27(4)14	42x6343	107
5292	(3)1(1)9(3)36(4)14	32x6232	99
6292	(3)1(0)1(1)11(0/3)1(3)38(4)15	32x5233	96
7992	(3)1(1)13(3)34(4)17	41x9332	104
Haplogroup 26 (1,3,4):			
1992	(1)12(3)25(4)24	3235162	47
5192	(1)15(3)47(4)12	3325232	52
9791	(1)14(3)48(4)14	1333332	1
13591	(1)14(3)46(4)15	2223232	10
Haplogroup 26 (3,1,3^+,4^-):			
CI128	(3)2(1)13(3)63(4)6	3125232	114
CI135	(3)3(1)14(3)59(4)8	3125232	114
CI166	(3)2(1)13(3)63(4)6	3125232	114
2692	(3)3(1)13(3)61(4)8	3135232	35
9191	(3)3(1)13(3)61(4)8	4115132	56
5692	(3)3(1)12(3)63(4)6	3126232	33
13991	(3)3(1)14(3)60(4)8	3135232	35
Haplogroup 26 (3,1,3,4):			
4591	(3)1(1)13(3)39(4)14	2233232	19
5791	(3)1(1)12(3)39(4)14	2223232	10
7792	(3)3(1)10(3)37(4)16	3246242	48
8792	(3)1(1)14(3)44(4)13	3325232	52
14391	(3)1(1)13(3)39(4)15	2433232	30
Haplogroup 26 (others):			
CI190	(3)1(1)11(0)1(1)1(3)51(4)15	5125231	71
Haplogroup 26 (microsatellites incomplete):			
3192	(3)1(1)14(3)>25(4)15	43x5232	109
5392	(3)1(1)13(3)>25(4)12	23x5253	86
5592	(3)1(1)14(3)42(4)14	33x5232	101
6692	(3)3(1)13(3)>33(4)8	31x6232	93
8492	(3)5(1)11(3)>10(4)14	22x5242	85

[a] Sample numbers with the prefix "CI" are from the Cook Islands; all others are from Papua New Guinea.

References

Bellwood PS (1979) Man's conquest of the Pacific: the prehistory of Southeast Asia and Oceania. Oxford University Press, Oxford

——— (1987) The Polynesians. Thames & Hudson, London

——— (1989) The colonization of the Pacific: some current hypotheses. In: Hill AVS, Serjeantson SW (eds) The colonization of the Pacific: a genetic trail. Clarendon Press, Oxford, pp 1–60

——— (1991) The Austronesian dispersal and the origin of languages. Sci Am (July): 70–75

Bertranpetit J, Calafell F (1996) Genetic and geographic variability in cystic fibrosis: evolutionary considerations. In: Chadwick D, Cardew G (eds) Variation in the human genome. Wiley & Sons, Chichester, pp 97–118

Bianchi NO, Bailliet G, Bravi CM, Carnese RF, Rothhammer F, Martínez-Marignac VL, Pena SDJ (1997) Origin of Amerindian Y-chromosomes as inferred by the analysis of six polymorphic markers. Am J Phys Anthropol 102:79–89

Bouzekri N, Taylor PG, Hammer MF, Jobling MA (1998) Novel mutation processes in the evolution of a haploid minisatellite, MSY1: array homogenization without homogenization. Hum Mol Genet 7:655–659

Bravi CM, Sans M, Bailliet G, Martinez-Marignac VL, Portas M, Barreto I, Bonilla C, et al (1997) Characterization of mitochondrial DNA and Y-chromosome haplotypes in a Uruguayan population of African ancestry. Hum Biol 69: 641–652

Cooper G, Amos W, Hoffman D, Rubinsztein DC (1996) Network analysis of human Y microsatellite haplotypes. Hum Mol Genet 5:1759–1766

de Knijff P, Kayser M, Caglià A, Corach D, Fretwell N, Gehrig C, Graziosi G, et al (1997) Chromosome Y microsatellites: population genetics and evolutionary aspects. Int J Legal Med 110:134–140

Flint J, Boyce AJ, Martinson JJ, Clegg JB (1989) Population bottlenecks in Polynesia revealed by minisatellites. Hum Genet 83:257–263

Forster P, Kayser M, Meyer E, Roewer L, Pfeiffer H, Benkmann H, Brinkmann B (1998) Phylogenetic resolution of complex mutational features at Y-STR DYS390 in Aboriginal Australians and Papuans. Mol Biol Evol 15:1108–1114

Goldstein DB, Zhivotovsky LA, Nayar K, Linares AR, Cavalli-Sforza LL, Feldman MW (1996) Statistical properties of the variation at linked microsatellite loci: implications for the history of human Y chromosomes. Mol Biol Evol 13: 1213–1218

Hammer MF (1995) A recent common ancestry for human Y chromosomes. Nature 378:376–378

Hammer MF, Horai S (1995) Y chromosomal DNA variation and the peopling of Japan. Am J Hum Genet 56:951–962

Hammer MF, Karafet T, Rasanayagam A, Wood ET, Altheide TK, Jenkins T, Griffiths RC, et al (1998) Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. Mol Biol Evol 15:427–441

Heyer E, Puymirat J, Dieltjes P, Bakker E, de Knijff P (1997) Estimating Y-chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. Hum Mol Genet 6:799–803

Heyerdahl T (1950) Kontiki: across the Pacific by raft. Rand McNally, Chicago

Hill AVS, Gentile B, Bonnardot JM, Roux J, Weatherall DJ, Clegg JB (1987) Polynesian origins and affinities: globin gene variants in eastern Polynesia. Am J Hum Genet 40:453–463

Hill AVS, Serjeantson SW (1989) The colonization of the Pacific: a genetic trail. Clarendon Press, Oxford

Jennings JD (1979) The prehistory of Polynesia. Harvard University Press, Cambridge, MA

Jeffreys AJ, Neumann R, Wilson V (1990) Repeat unit sequence variation in minisatellites: a novel source of DNA polymorphism for studying variation and mutation by single molecule analysis. Cell 60:473–485

Jobling MA, Bouzekri N, Taylor PG (1998) Hypervariable digital DNA codes for human paternal lineages: MVR-PCR at the Y-specific minisatellite, MSY1 (*DYF155S1*). Hum Mol Genet 7:643–653

Jobling MA, Samara V, Pandya A, Fretwell N, Bernasconi B, Mitchell RJ, Gerelsaikhan T, et al (1996) Recurrent duplication and deletion polymorphisms on the long arm of the Y chromosome in normal males. Hum Mol Genet 5: 1767–1775

Jobling MA, Tyler-Smith C (1995) Fathers and sons: the Y chromosome and human evolution. Trends Genet 11: 449–456

Kwok C, Tyler-Smith C, Medonca BB, Hughes I, Berkovitz GD, Goodfellow PN, Hawkins JR (1996) Mutation analysis of 2kb 5 to SRY in XY females and XY intersex subjects. J Med Genet 33:465–468

Lum JK, Cann RL, Martinson JJ, Jorde LB (1998) Mitochondrial and nuclear genetic relationships among Pacific Island and Asian populations. Am J Hum Genet 63:613–624

Lum JK, Richards O, Ching C, Cann RL (1994) Polynesian mitochondrial DNAs reveal three deep maternal lineage clusters. Hum Biol 66:567–590

Mathias N, Bayés M, Tyler-Smith C (1994) Highly informative compound haplotypes for the human Y chromosome. Hum Mol Genet 3:115–123

Melton T, Peterson R, Redd AJ, Saha N, Sofro ASM, Martinson J, Stoneking M (1995) Polynesian genetic affinities with Southeast Asian populations as identified by mtDNA analysis. Am J Hum Genet 57:403–414

Redd AJ, Takezaki N, Sherry ST, McGarvey ST, Sofro ASM, Stoneking M (1995) Evolutionary history of the COII/tRNAlys intergenic 9 base pair deletion in human mitochondrial DNAs from the Pacific. Mol Biol Evol 12:604–615

Reed PW, Davies JL, Copeman JB, Bennett ST, Palmer SM, Pritchard LE, Gough SCL, et al (1994) Chromosome-specific microsatellite sets for fluorescence-based, semi-automated genome mapping. Nat Genet 7:390–395

Santos FR, Tyler-Smith C (1996) Reading the human Y chromosome: the emerging DNA markers and human genetic history. Braz J Genet 19:665–670

Serjeantson SW (1989) HLA genes and antigens. In: Hill AVS, Serjeantson SW (eds) The colonization of the Pacific: a genetic trail. Clarendon Press, Oxford, pp 120–173

Spurdle AB, Jenkins T (1992) The search for Y-chromosome polymorphism is extended to negroids. Hum Mol Genet 1: 169–170

Spurdle AB, Woodfield DG, Hammer MF, Jenkins T (1994)

The genetic affinity of Polynesians: evidence from Y-chromosome polymorphisms. Ann Hum Genet 58:251–263

Sykes B, Leiboff A, Low-Beer J, Tetzner S, Richards M (1995) The origins of the Polynesians: an interpretation from mitochondrial lineage analysis. Am J Hum Genet 57: 1463–1475

Terrell J (1986) Prehistory in the Pacific islands. Cambridge University Press, Cambridge

Thomas MG, Skorecki K, Ben-Ami H, Parfitt T, Bradman N, Goldstein DB (1998) Origins of Old Testament priests. Nature 394:138–139

Underhill PA, Jin L, Lin AA, Mehdi SQ, Jenkins T, Vollrath D, Davis RW, et al (1997) Detection of numerous Y-chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. Genome Res 7:996–1005

Underhill PA, Jin L, Zemans R, Oefner PJ, Cavalli-Sforza LL (1996) A pre-Columbian Y-chromosome–specific transition and its implications for human evolutionary history. Proc Natl Acad Sci USA 93:196–200

Veitia R, Ion A, Barbaux S, Jobling MA, Souleyreau N, Ennis K, Ostrer H, et al (1997) Mutations and sequence variants in the testis-determining region of the Y chromosome in individuals with a 46, XY female phenotype. Hum Genet 99:648–652

Whitfield LS, Sulston JE, Goodfellow PN (1995) Sequence variation of the human Y chromosome. Nature 378:379–380

Zerjal T, Dashnyam B, Pandya A, Kayser M, Roewer L, Santos FR, Schiefenhövel W, et al (1997) Genetic relationships of Asians and northern Europeans, revealed by Y-chromosomal DNA analysis. Am J Hum Genet 60:1174–1183

Am. J. Hum. Genet. 65:1437–1448, 1999

Recent Male-Mediated Gene Flow over a Linguistic Barrier in Iberia, Suggested by Analysis of a Y-Chromosomal DNA Polymorphism

Matthew E. Hurles,[1,*] Reiner Veitia,[2,*] Eduardo Arroyo,[3] Manuel Armenteros,[4] Jaume Bertranpetit,[5,†] Anna Pérez-Lezaun,[5,†] Elena Bosch,[5,†] Maria Shlumukova,[1,‡] Anne Cambon-Thomsen,[6] Ken McElreavey,[2] Adolfo López de Munain,[7] Arne Röhl,[8] Ian J. Wilson,[9,§] Lalji Singh,[10] Arpita Pandya,[11] Fabrício R. Santos,[11,‖] Chris Tyler-Smith,[11] and Mark A. Jobling[1]

[1]Department of Genetics, University of Leicester, University Road, Leicester, United Kingdom; [2]Unité d'Immunogénétique Humaine, Institut Pasteur, Paris; [3]Universidad Complutense de Madrid, Facultad de Medicina, and [4]Centro de Investigacion y Criminalistica, Laboratorio de ADN, Policia Judicial, Guardia Civil, Madrid; [5]Unitat d'Anthropologia, Departimento de Biologia Animal, Universitat de Barcelona, Barcelona; [6]INSERM U518, Faculté de Médecine, Toulouse; [7]Servicio de Neurologia, Arantzazuko Amaren Ospitalea, San Sebastian, Spain; [8]Mathematisches Seminar, Universität Hamburg, Hamburg; [9] School of Biological Sciences, Queen Mary and Westfield College, London; [10]Centre for Cellular and Molecular Biology, Uppal Road, Hyderabad; India; and [11]CRC Chromosome Molecular Biology Group, Department of Biochemistry, University of Oxford, Oxford

Summary

We have examined the worldwide distribution of a Y-chromosomal base-substitution polymorphism, the T/C transition at SRY-2627, where the T allele defines haplogroup 22; sequencing of primate homologues shows that the ancestral state cannot be determined unambiguously but is probably the C allele. Of 1,191 human Y chromosomes analyzed, 33 belong to haplogroup 22. Twenty-nine come from Iberia, and the highest frequencies are in Basques (11%; $n = 117$) and Catalans (22%; $n = 32$). Microsatellite and minisatellite (MSY1) diversity analysis shows that non-Iberian haplogroup-22 chromosomes are not significantly different from Iberian ones. The simplest interpretation of these data is that haplogroup 22 arose in Iberia and that non-Iberian cases reflect Iberian emigrants. Several different methods were used to date the origin of the polymorphism: microsatellite data gave ages of 1,650, 2,700, 3,100, or 3,450 years, and MSY1 gave ages of 1,000, 2,300, or 2,650 years, although 95% confidence intervals on all of these figures are wide. The age of the split between Basque and Catalan haplogroup-22 chromosomes was calculated as only 20% of the age of the lineage as a whole. This study thus provides evidence for direct or indirect gene flow over the substantial linguistic barrier between the Indo-European and non–Indo-European–speaking populations of the Catalans and the Basques, during the past few thousand years.

Received April 14, 1999; accepted for publication August 17, 1999; electronically published October 8, 1999.

Address for correspondence and reprints: Dr. Mark A. Jobling, Department of Genetics, University of Leicester, University Road, Leicester LE1 7RH, United Kingdom. E-mail: maj4@leicester.ac.uk

© 1999 by The American Society of Human Genetics. All rights reserved.
0002-9297/1999/6505-0028$02.00

Introduction

Most regions of sharp genetic change within Europe correspond to linguistic boundaries (Barbujani and Sokal 1990). The Basques speak a non–Indo-European language with no close affinities to any other extant language (Ruhlen 1991), and this linguistic uniqueness has led to the idea that the Basques may represent a Mesolithic relict population, isolated, by linguistic and geographic barriers, from cultural and genetic exchange. Although specific archaeological evidence for such a picture is lacking (Collins 1986), genetic analysis certainly lends some support to the view of the Basques as an isolate, in light of unusual frequencies of alleles in blood groups such as rhesus and ABO (Mourant 1947, 1983) and of disease alleles in the calpain-3 gene that are responsible for limb-girdle muscular dystrophy (Urtasun et al. 1998). Albeit to a lesser extent, mtDNA sequences (Bertranpetit et al. 1995; Côrte-Real et al. 1996; Comas et al. 1997) and HLA types (Comas et al. 1998) also support this view. The use of multiple autosomal loci in principal-component analysis (Bertranpetit and Cavalli-Sforza 1991; Calafell and Bertranpetit 1994*a*, 1994*b*), in the calculation of various genetic-distance measures (Calafell and Bertranpetit 1994*b*) and in a method de-

[*] The first two authors contributed equally to this work and are listed alphabetically.

[†] Present affiliation: Unitat de Biologia Evolutiva, Facultat de Ciències de la Salut i de la Vida, Universitat Pompeu Fabra, Barcelona.

[‡] Present affiliation: School of Biomedical Sciences, Queen's Medical Centre, Nottingham.

[§] Present affiliation: Department of Mathematical Sciences, King's College, Aberdeen.

[‖] Present affiliation: Departamento de Biologia Geral, Instituto Ciências Biologicas/Universidade Federal de Minas Gerais, Minas Gerais, Brazil.

signed to detect genetic boundaries (Barbujani 1991), clearly distinguishes the Basque region from the rest of the Iberian peninsula.

The majority of the Y chromosome is nonrecombining, and so mutations on it represent a record of its evolutionary past that can be used in the reconstruction of population histories (Jobling and Tyler-Smith 1995; Mitchell and Hammer 1996). The presence of different polymorphic systems with different mutation rates and processes is a particular strength of the Y chromosome and allows us to use Y-chromosome markers "genealogically," defining lineages ("haplogroups") with slowly mutating biallelic polymorphisms such as base substitutions, which can be regarded as unique events in human evolution, and then examining diversity within these haplogroups, using polymorphisms that mutate more rapidly, such as microsatellites (Kayser et al. 1997) and the minisatellite, MSY1 (Jobling et al. 1998). This alleviates the problem of recurrent mutation at these loci and allows attempts to be made to date haplogroup origins.

A comparison of the correlation of languages with Y-chromosomal haplotypes (defined by the marker 49f [Ngo et al. 1986]) and with mtDNA haplotypes has suggested that the passing on of language from generation to generation is governed more by patrilineage than by matrilineage (Poloni et al. 1997). This certainly appears to be so for the Uralic-speaking Finns, who share most of their mtDNA lineages with Indo-European speakers (Lahermo et al. 1996) but, in contrast, approximately half of their Y-chromosome lineages with Central Asian Uralic speakers (Zerjal et al. 1997). In the case of the Basques, with their unique linguistic heritage, it is of particular interest to know whether their Y chromosomes are distinct from those of surrounding populations. Although a statistically significant difference has been shown between Basques and other populations, including Catalans, in studies using 49f (Lucotte and Hazout 1996; Poloni et al. 1997), a study using Y-chromosomal microsatellites (Pérez-Lezaun et al. 1997) finds no such difference. Here, we show that a specific Y-chromosomal lineage, which has a recent origin and is rare or absent in most parts of the world, is shared at high frequency between Basques and Catalans. This constitutes evidence for substantial recent male-mediated gene flow over a major linguistic barrier.

Subjects and Methods

Subjects

Gifts (providers) of DNA samples from autochthonous males, defined in most cases on the basis of grandpaternal birthplace, were as follows: Castilians (Santos Alonso and John Armour), Galicians (Marisol Rodriguez-Calvo), León (Carlos Polanco), Belarusians (Yuri Dubrova), Germans (Manfred Kayser), and other DNA samples (our own collections). The Catalan samples (from Girona) and 51 of the Basque samples (from Guipúzcoa [and denoted by the suffix "v" intable 2]) have been described elsewhere (Pérez-Lezaun et al. 1997); Basque samples denoted as "m336"–"m365" are from Zumaya in Guipúzcoa, and the remaining Basques are from Pyrenées Atlantiques. All samples were taken with appropriate informed consent.

DNA Sequencing

Direct sequencing of 1.2-kb PCR products amplified from human and primate DNAs by use of the SRY-2627 primers R1 and F1 (Veitia et al. 1997) was performed by use of F1 as sequencing primer and with BigDye technology (Perkin-Elmer) on an ABI377 sequencer (Applied Biosystems).

Typing of SRY-2627 and 92R7

SRY-2627 (previously referred to as "SRY-2628" [Veitia et al. 1997]) was typed by PCR using the R1 and F1 primers (Veitia et al. 1997), followed by *Bsi*HKAI digestion (fig. 1*A*). Apparent T-allele chromosomes were verified by use of *Ban*I. A 709-bp amplicon containing the 92R7 (Mathias et al. 1994) polymorphism was amplified by use of the primers 5-GAC CCG CTG TAG ACC TGA CT-3′ and 5′-GCC TAT CTA CTT CAG TGA TTT CT-3′, in an MJR PTC-200 thermal cycler (33 cycles of 94°C for 30 s, 62°C for 30 s, and 72°C for 60 s). Then, typing by *Hin*dIII digestion was done, to give 197- and 512-bp fragments from the C allele; the 709-bp product remains, since there is more than one copy of the locus on the Y chromosome, and since only a subset contains the polymorphic site. In this study, haplogroup 1 is defined by the 92R7 T allele in the presence of the SRY-1532 G allele (Hurles et al. 1998).

Microsatellite and MSY1 Haplotyping

Seven Y-specific microsatellites (*DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392,* and *DYS393*) were typed. Primer sequences are those given by Kayser et al. (1997).

PCR reactions (94°C for 10 min; 30 cycles of 94°C for 30 s, 55°C for 30 s, and 72°C for 30 s; and then 72°C for 10 min) were performed in the presence of a fluorescently labeled dUTP derivative (R6G; Applied Biosystems) diluted 1:600 with respect to dTTP. Products were electrophoresed on an ABI373A sequencer, and data were analyzed by GeneScan software (Applied Biosystems). Three-state microsatellite variant repeat (MVR)-PCR at the minisatellite MSY1 was performed as described elsewhere (Jobling et al. 1998).

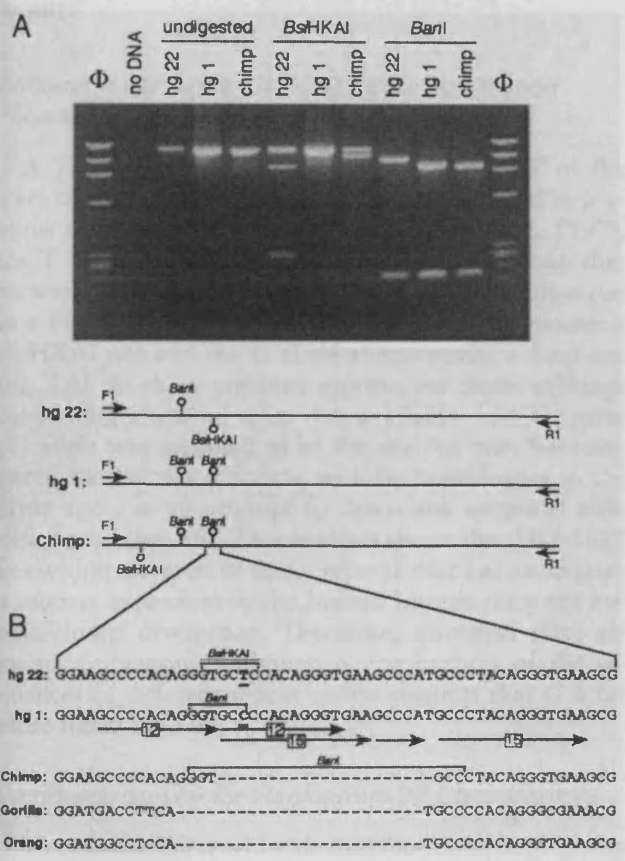


Figure 1 SRY-2627 polymorphism. *A*, Detection by PCR-RFLP analysis, and restriction map of PCR fragment. Digestion of the 1,242-bp product by *Bst*HKAI in the T allele (haplogroup 22) chromosomes gives fragments of 945 and 297 bp (digests shown are partial). Products of 946, 210, and 86 bp are obtained by digestion of the C allele (here, haplogroup 1) chromosomes with *Ban*I. The fragment generated from chimp DNA contains a *Ban*I site at the polymorphic position and therefore resembles the human C allele. hg = Haplogroup; Φ = *Hae*III-digested ΦX174 phage DNA. Short arrows indicate PCR primers. *B*, Sequences of the SRY-2627 C and T alleles and of the homologous regions in DNAs of great apes. Arrows under the haplogroup 1 sequence indicate complete and partial copies of 12- and 15-bp direct repeats. Because of these, several alignments are possible, and the ancestral state cannot be unambiguously deduced; however, the polymorphism is at the third base of one of the 15-bp repeats in human DNA, and all other copies of the repeat, in both humans and apes, share a C at this position, suggesting that C is ancestral. In the alignment shown, the *Ban*I site in chimp DNA is not homologous to that in DNA of the human C allele.

Microsatellite and Minisatellite Network Construction and Dating

Median joining networks were constructed by Network 1.1 (Bandelt et al. 1999); *DYS389I*-allele lengths (tables 2and 3) were subtracted from *DYS389II*-allele lengths prior to analysis, since the former is contained within the latter (Cooper et al. 1996). Dating was done separately, by use of both microsatellite and MSY1 data, by three different methods; for a description, see the Age of the SRY-2627 Mutation subsection. For two of the methods (Bertranpetit and Calafell 1996; Thomas et al. 1998), the root of a haplotype tree (microsatellite haplotype 21; MSY1 haplotype 6 [see table 2]) was chosen from pairwise differences, as that having the smallest number of mutational steps from all other chromosomes; this is identical to a haplotype constructed from the modal allele lengths. Note that uncertainty in root assignment has only a minimal effect on dating (data not shown). The average squared distance (ASD [Thomas et al. 1998]) was calculated, by Microsat 1.5d (Minch 1997), between a population of chromosomes and the root haplotype (this method was also used to date the divergence between Basque and Catalan haplogroup-22 chromosomes and to estimate diversity differences between haplogroups 1 and 22 and between European and Asian haplogroup-1 chromosomes). For all methods, we have assumed a generation time of 25 years and mutation rates of 2%–11% per generation (Jobling et al. [1998]; we also use the midpoint of this range, 6.5%), for MSY1, and 2.1×10^{-3} (95% confidence limits [95% CI] $0.6–4.9 \times 10^{-3}$ [Heyer et al. 1997]), for microsatellites; for the third method (Goldstein et al. 1996), we consider N_e to be 4,900 (Hammer 1995). Throughout, mutation in each MSY1 repeat block is weighted for repeat number (Hurles et al. 1998), on the assumption that each repeat unit has an equal probability of mutating.

A fourth, coalescent-based method also was used, for microsatellite data only; it differs from a published one (Wilson and Balding 1998), by allowing for exponential growth in the population of Y chromosomes. This method uses a Markov-chain Monte Carlo simulation algorithm to generate simulated trees consistent with the observed haplotype data, sampling 10,000 of these trees at a rate proportional to their probability under a coalescent-with-exponential-growth model. No prior assumption is made about population size, but mutations are assumed to be stepwise, and the prior mutation rate is the same as that used in the other methods (Heyer et al. 1997). The output from this method includes probability distributions for T (tree height) and N (population size), from which a probability distribution for the time to the most recent common ancestor can be derived, again by use of a 25-year generation time. The standard coalescent model assumes that haplotypes are sampled at random from the whole population. However, when a population is growing rapidly, the coalescent-with-exponential-growth will be a good approximation to the genealogy of a haplogroup. Readers interested in this method are asked to contact I.J.W. Pairwise R_{ST} analysis was performed by use of the Arlequin package (Schneider et al. 1997).

Results

Ancestral State of the SRY-2627 Base-Substitution Polymorphism

A T/C transition polymorphism 2,627 bp 5′ of the start codon of the *SRY* gene has been described in previous reports (Bianchi et al. 1997; Veitia et al. 1997); the T allele defines a Y-chromosomal haplogroup that we term "haplogroup 22" and can be conveniently typed in a PCR-RFLP assay, since the T allele alone creates a *Bsi*HKAI site and the C allele alone creates a *Ban*I site (fig. 1*A*). In these previous reports, no direct evidence concerning ancestral state was available, and the rarer (T) allele was assumed to be the derived one. We compared the human sequence with its homologues in the great apes, in an attempt to determine ancestral state definitively (fig. 1*B*). This analysis shows that SRY-2627 lies within a region of direct repeats that has undergone a modest expansion in the human lineage since the human-chimp divergence. Therefore, ancestral state remains ambiguous, although a comparison of the sequences of different repeat copies suggests that C is far more likely than T.

Worldwide Survey for Haplogroup-22 Chromosomes

To examine the worldwide distribution of haplogroup 22, we performed an initial survey of 752 Y chromosomes. We found 10 haplogroup-22 chromosomes: singletons were found in England, in Germany, in a general French sample, and in a sample from a southwestern French population, the Béarnais, but the remaining six chromosomes were all found in Basques. In previous studies, SRY-2627/T-allele chromosomes had been found in France (Veitia et al. 1997) and also in South America (Bianchi et al. 1997), where they were at highest frequency in nonindigenous groups, who are likely to have an Iberian origin. We therefore intensified our survey within Iberia itself; practical difficulties in obtaining French DNA samples precluded a detailed survey of this region.

Haplogroup-22 Chromosomes in Iberia and France

We typed a further 439 Y chromosomes from Iberia, for SRY-2627, making a total of 469 Iberian chromosomes and 1,191 chromosomes worldwide. This survey yielded a further 23 chromosomes from haplogroup 22. The global distribution of this haplogroup is shown in table 1 and figure 2*A*, and the distribution within Iberia is shown in more detail in figure 2*B*. When the data are summed (and the very small Valencian sample is excluded), the populations in which we find haplogroup 22 at its highest frequency are the Basques (11%) and the Catalans (22%), whereas in other parts of Iberia that were sampled these chromosomes are rare or absent.

Table 1

Populations Tested for SRY-2627, and Summary of Results

Continent and Population	No. of Alleles C	T
Europe:		
Iberia:		
Basque	104	13
Catalan	25	7
Galician	46	2
Andalucian	83	1
Madrid (urban)	60	2
Castilla la Mancha	23	0
Castilla y León	47	0
Extremaduran	18	0
Valencian	5	1
Other Spanish (not Basque or Catalan)	58	3
Other:		
Béarnais	13	1
French	33	1
British	63	1
German	48	1
Italian	39	0
Greek	19	0
Hungarian	34	0
Icelandic	27	0
Norwegian	46	0
Belarusian	23	0
Caucasus	16	0
Other	8	0
Asia:		
Chinese	40	0
Japanese	25	0
Indian	86	0
Gujarati	61	0
Sri Lankan	13	0
Tibetan	3	0
Mongolian	23	0
Other	5	0
Africa:		
Algerian	27	0
San	6	0
Biaka	4	0
Kenyan	7	0
Other	3	0
Oceania:		
Cook Islander	6	0
Australian	4	0
Melanesian	2	0
Americas:		
Various	5	0
Total	1,158	33

Diversity Assessed by Use of Y-Specific Microsatellites and a Minisatellite

To illuminate the geographic and temporal origin of these haplogroup-22 chromosomes, we first examined their diversity, using seven highly polymorphic Y-specific

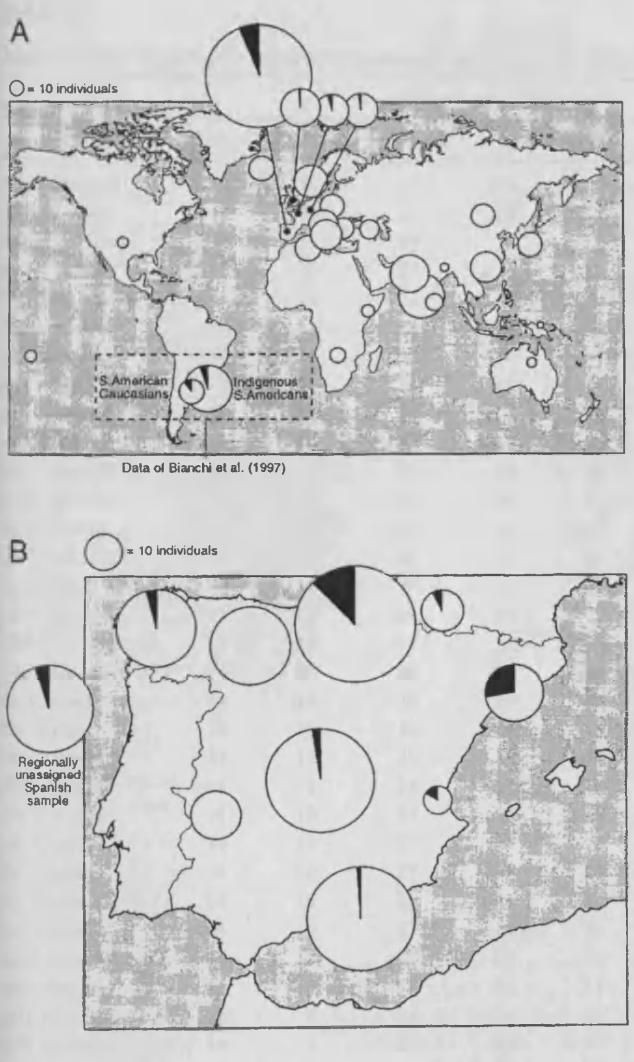


Figure 2 Geographical distribution of haplogroup-22 chromosomes. *A*, Worldwide distribution (see table 1). Data for South America are from Bianchi et al. (1997), as follows: pooled indigenous South Americans, 5/93 SRY-2627/T allele chromosomes; La Plata nonindigenous groups, 3/26 SRY-2627/T allele chromosomes. *B*, Distribution within Iberia and the Béarnais of southern France. Data on Madrid and Castilla la Mancha have been pooled.

microsatellites. Haplotypes were determined (table 2), and a median-joining network was constructed (fig. 3*A*). In this analysis, we included 2 SRY-2627/T-allele chromosomes of French origin that previously had been identified within the pedigrees catalogued by the Centre d'Étude du Polymorphisme Humain (CEPH) (Bianchi et al. 1997) and 1 SRY-2627/T-allele chromosome from the study by Veitia et al. (1997), together with the 33 identified in the present study, making a total of 36.

In addition, we determined the microsatellite diversity of 50 Asian and European Y chromosomes belonging to haplogroup 1 (table 3 and fig. 3*B*). This haplogroup is distinguished from haplogroup 22 only by the SRY-2627 mutation and—if it is assumed that the ancestral state of SRY-2627 is indeed the C allele (Bianchi et al. 1997; Veitia et al. 1997)—is the ancestral haplogroup. When we compare the two microsatellite networks (fig. 3*A* and *B*), it is clear that haplogroup 1 is substantially more diverse; calculated ASD values are 0.290 for haplogroup 22, compared with 1.063 for haplogroup 1. Together with the much wider geographic distribution of haplogroup 1 (Santos and Tyler-Smith 1996), this is additional evidence that the ancestral state of SRY-2627 is the C allele.

We also determined "MSY1 codes," by MVR-PCR for the haplogroup-22 chromosomes (table 2); in this technique, the positions of three different classes of variant 25-bp repeat units along the MSY1 minisatellite array are mapped by use of discriminator primers specific to individual repeat types. Compared with other haplogroups analyzed (Jobling et al. 1998), this haplogroup has low diversity: four pairs, one set of three, and one set of seven males have identical MSY1 codes. Of the 35 chromosomes analyzed, 32 have MSY1 codes with the same modular structure (the order of blocks of different repeat types along the array), "1,3,4." If a single-step mutation model is assumed, these codes can also be assembled into a compact network (fig. 3*C*). Three chromosomes (all Iberian) have codes with structures that are more complex, and they are omitted from the network and from the dating calculations described below.

We can envisage two scenarios to explain the current geographic distribution of haplogroup 22: either the SRY-2627 mutation occurred outside Iberia, and individuals carrying it migrated into Iberia, where subsequent drift led to the high frequency in this region; or, alternatively, the origin was in Iberia, and the non-Iberian cases are explained by emigration. If the first explanation is correct, then we might expect that non-Iberian haplogroup-22 chromosomes would have higher haplotype diversity than is present in Iberian haplogroup-22 chromosomes and that these non-Iberian cases would include haplotypes at the peripheries of the networks, with the Iberian chromosomes forming a tighter cluster. This kind of partitioning is vividly displayed in the haplogroup-1 microsatellite network (fig. 3*B*), where Asian chromosomes lie at the network's periphery, consistent with an origin of this haplogroup outside Europe (see Karafet et al. [1999], who refer to the equivalent class of chromosomes as "haplotype 1C"), and show much more diversity than European chromosomes: diversity, measured in terms of ASD, is 1.762 for Asians, whereas it is only 0.359 for Europeans, a difference that is also supported by principal-components analysis (data not shown). In this scenario, we would also expect to see sharing of Asian and European haplotypes in the core of the network, and the absence of such sharing

Table 2

Microsatellite Haplotypes and MSY1 Codes of Haplogroup-22 Chromosomes

Individual (Population)	No. of Repeats Units in[a]							Microsatellite Haplotype[b]	MSY1	
	DYS19	*DYS389I*	*DYS389II*	*DYS390*	*DYS391*	*DYS392*	*DYS393*		Code[c]	Haplotype[d]
m337 (Basque)	14	10	27	24	10	13	13	14	$(1)_{16}(3)_{40}(4)_{18}$[e]	10
4301 (Basque)	14	10	27	24	10	13	13	14	$(1)_{16}(3)_{38}(4)_{19}$	7
46205 (Basque)	14	10	27	24	10	13	13	14	$(1)_{17}(3)_{38}(4)_{14}$	18
35v (Basque)	14	10	27	24	10	13	13	14	$(1)_{16}(3)_{39}(4)_{19}$	9
67c (Catalan)	14	10	27	24	10	13	13	14	$(1)_{15}(3)_{38}(4)_2(3)_1(4)_{15}$	22
m354 (Basque)	14	11	28	24	11	13	13	32	$(1)_{16}(3)_{38}(4)_{19}$[e]	7
m362 (Basque)	14	11	28	24	11	13	13	32	$(1)_{16}(3)_{38}(4)_{19}$[e]	7
m363 (Basque)	14	11	28	24	11	13	13	32	$(1)_{17}(3)_{37}(4)_{19}$[e]	17
ma19 (Madrid)	14	11	28	24	11	11	13	30	$(1)_{18}(3)_{39}(4)_{19}$	20
m62 (British)	14	10	28	24	11	13	13	33	$(1)_{15}(3)_{38}(4)_{20}$	3
m147 (French)[f]	14	10	29	24	10	13	13	34	$(1)_{16}(3)_{40}(4)_{19}$	11
m348 (Basque)	14	10	27	24	11	13	13	21	$(1)_{16}(3)_{41}(4)_{17}$[e]	14
m95 (French)	14	10	27	24	11	13	13	21	$(1)_{16}(3)_{39}(4)_{18}$[e]	8
ga29 (Galician)	14	10	27	24	11	13	13	21	$(1)_{14}(3)_{38}(4)_{18}$	1
7c (Catalan)	14	10	27	24	11	13	13	21	$(1)_{16}(3)_{38}(4)_{18}$	6
m341 (Basque)	14	10	27	24	12	13	13	35	$(1)_{14}(3)_{39}(4)_{19}$[e]	2
CEPH201 (French)[g]	14	10	27	24	12	13	13	35	$(1)_{16}(3)_{38}(4)_{19}$	7
6201 (Béarnais)	14	10	26	24	10	13	13	17	$(1)_{16}(3)_{38}(4)_{17}$	5
8v (Basque)	14	11	30	24	10	13	13	36	$(1)_{15}(3)_{39}(4)_{18}$	4
32v (Basque)	14	10	26	24	11	13	13	15	$(1)_{16}(3)_{41}(4)_{16}$	13
98v (Basque)	14	11	28	24	10	13	13	37	$(1)_{16}(3)_{38}(4)_{19}$	7
101v (Basque)	14	11	28	24	11	11	13	30	$(1)_{16}(3)_{38}(4)_{19}$	7
41c (Catalan)	14	10	27	23	10	13	13	38	$(1)_{15}(3)_{37}(4)_1(3)_2(4)_{16}$	21
45c (Catalan)	14	10	27	24	10	11	13	39	$(1)_{16}(3)_{41}(4)_{15}$	12
56c (Catalan)	14	10	27	24	11	11	13	25	$(1)_{16}(3)_{38}(4)_{18}$	6
81c (Catalan)	15	10	28	24	11	13	13	40	$(1)_{16}(3)_{38}(4)_{18}$	6
85c (Catalan)	14	10	27	24	10	11	14	41	$(1)_{16}(3)_{38}(4)_{17}$	5
ma20 (Madrid)	14	10	27	24	11	13	14	42	$(1)_{16}(3)_{44}(4)_{14}$	15
sp21 (Spanish)	15	10	26	24	11	13	13	43	$(1)_{16}(3)_{41}(4)_{16}$	13
ga22 (Galician)	14	9	26	24	11	13	13	23	$(1)_{18}(3)_{38}(4)_{17}$	19
sp77 (Spanish)	14	9	26	24	11	13	13	23	$(1)_{16}(3)_2(1)_1(3)_{40}(4)_1(3)_2(4)_{15}$	23
sp79 (Spanish)	15	10	27	24	10	13	13	29	$(1)_{17}(3)_{37}(4)_{19}$	17
sp123 (Valencian)	14	11	28	23	11	13	13	44	$(1)_{16}(3)_{39}(4)_{18}$	8
CEPH3501 (French)[g]	14	10	28	23	11	13	13	45	$(1)_{16}(3)_{38}(4)_{19}$	7
ge3202 (German)	14	10	29	24	11	13	13	46	$(1)_{16}(3)_{44}(4)_{16}$	16
alm1 (Andalucian)	14	11	28	24	11	12	13	47	Not done	...

[a] As defined by Kayser et al. (1997).

[b] See figure 3*A*.

[c] For example, "$(1)_{16}(3)_{40}(4)_{18}$" denotes that, 5′→3′, there are 16 type 1 repeats, 40 type 3 repeats, and 18 type 4 repeats (Jobling et al. 1998).

[d] See figure 3*C*.

[e] As determined by P. G. Taylor (Jobling et al. 1998).

[f] Identified by Veitia et al. (1997).

[g] Identified by Bianchi et al. (1997).

here is a function of the relatively small sample size—and limited geographic distribution—of the Asian chromosomes sampled. In contrast to this, all haplogroup-22 chromosomes cluster tightly in both microsatellite and minisatellite networks (fig. 3*A* and *C*), with, for the microsatellites, no pair of connected haplotypes differing by more than one mutation, indicating a probable Iberian origin.

Age of the SRY-2627 Mutation

From the microsatellite and minisatellite haplotype diversity, we can attempt to date the origin of the SRY-2627 mutation (table 4). We have used three different methods, one (Bertranpetit and Calafell 1996) based on the mean number of mutational steps from the network root, a second (Thomas et al. 1998) based on the ASD

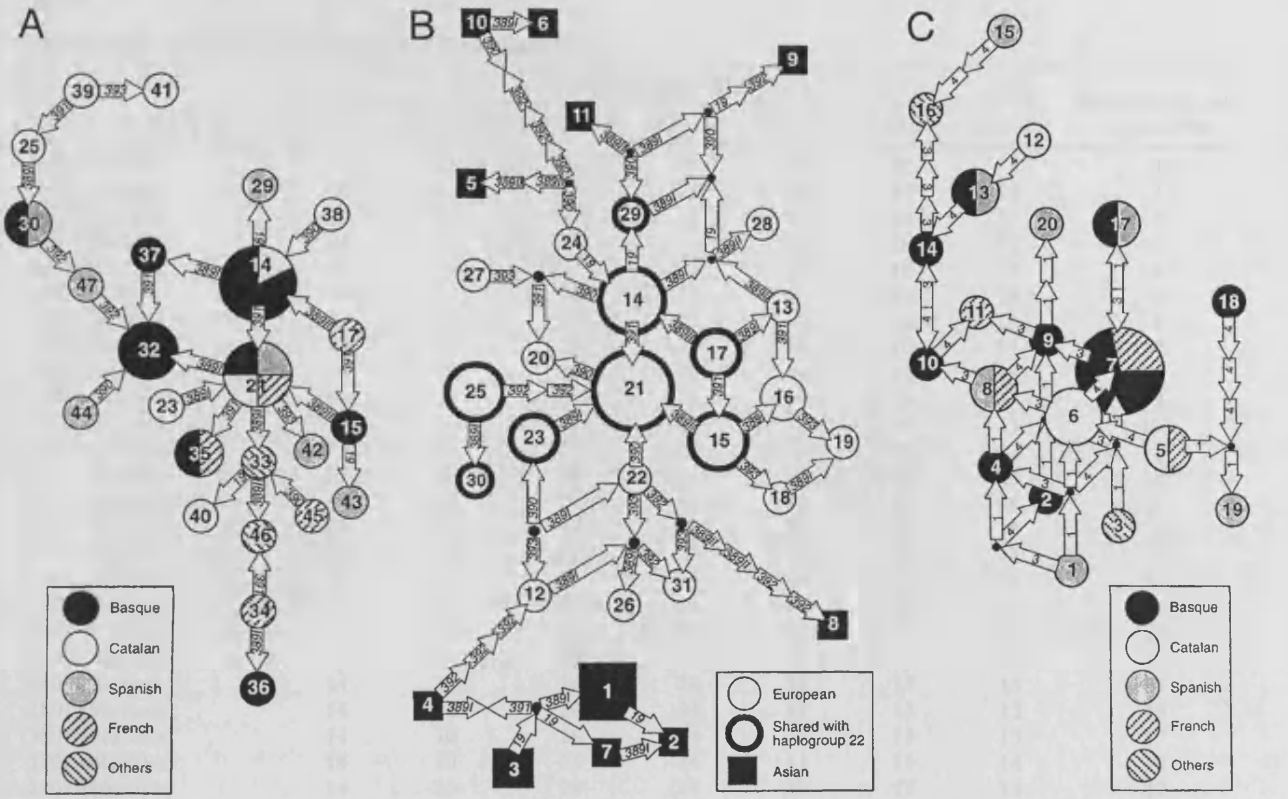


Figure 3 Median joining microsatellite haplotype networks for haplogroup-22 chromosomes (*A*) and haplogroup-1 chromosomes (*B*) and (*C*) MSY1 network for haplogroup 22 chromosomes. A microsatellite or MSY1 haplotype is represented by a circle, with its area proportional to the number of individuals having that haplotype; haplotype number (see tables 2and 3) is given within each circle. A single arrow between circles represents a single mutational step in a haplotype, and its direction indicates an increase in allele length; the mutating microsatellite (e.g., "*390*," representing *DYS390*) or MSY1 repeat block (e.g., "4," representing the block of type 4 repeats) is indicated on the arrow. Arrows do not imply evolutionary pathways. A small blackened circle on a line indicates an intermediate, unobserved haplotype used as a node (i.e., median vector). Thick edges to circles highlight haplotypes shared by both haplogroups (haplogroup 1 only).

from the root, and a third (Goldstein et al. 1996) based on the extent of variance accumulated since the base substitution occurred on a single haplotype. We also have used a fourth, coalescent-based approach, for microsatellite data alone, that is an extension of a published method (Wilson and Balding 1998) and that has been described in the Subjects and Methods section.

Although the 95% CIs are wide, agreement between the different methods and systems is good (with five of the seven ages within the 2,300–3,500-year range) and suggests that the origin of the SRY-2627 polymorphism occurred a few thousand years before the present. Ages calculated from MSY1 data are consistently younger than those calculated from microsatellite data; this may be due to the omission of the more complex MSY1 alleles from the dating and also to the probable inadequacy of the single-step model for MSY1 mutation. The structure of the MSY1 network in itself provides information on possible deviations from this simple model. There are six connections between haplotypes that involve more than one repeat unit: of these, four are confined to a single repeat block, and two involve a single-step increase in one repeat block, accompanied by a single-step decrease in an adjacent block. One interpretation of these observations is that multistep mutations can occur within blocks and, possibly, that mutations can occur that simultaneously expand one block and contract a neighboring block, perhaps by switching a boundary repeat from one type to another. Forthcoming direct data on mutation at MSY1 should show whether such events really occur and should allow us to use this locus in a more sophisticated way in the future.

The microsatellite networks for haplogroups 1 and 22 overlap substantially, with eight shared haplotypes; the two most common haplotypes in haplogroup 22 (frequency 9/36 when considered together) are also common (frequency 10/50) in haplogroup 1. This is consistent with an origin of haplogroup 22 on a haplogroup-1 background, followed by much parallel mutation, and insufficient time for substantial divergence of haplo-

Table 3

Microsatellite Haplotypes of Haplogroup-1 Chromosomes

Individual (Population)	No. of Repeat Units in[a]							Microsatellite Haplotype
	DYS19	*DYS389I*	*DYS 389II*	*DYS390*	*DYS391*	*DYS392*	*DYS393*	
m419 (Indian)	14	11	27	23	10	10	14	1
m432 (Indian)	14	11	27	23	10	10	14	1
m456 (Indian)	15	11	27	23	10	10	14	2
m464 (Indian)	13	10	27	23	10	10	14	3
m467 (Indian)	14	9	27	23	11	10	14	4
m503 (Indian)	13	10	29	23	10	13	13	5
m507 (Indian)	13	11	27	22	10	16	13	6
m620 (Indian)	15	10	27	23	10	10	14	7
m245 (Mongolian)	14	10	29	23	11	16	13	8
m283 (Mongolian)	16	11	27	23	10	14	13	9
m418 (Indian)	13	10	27	22	10	16	13	10
m434 (Indian)	14	11	27	23	10	10	14	1
m437 (Indian)	15	10	28	23	10	13	13	11
m469 (Indian)	14	11	27	23	10	10	14	1
m470 (Indian)	13	10	27	23	10	10	14	3
2001 (Basque)	14	9	27	23	11	13	14	12
2401 (Basque)	14	11	26	24	10	13	13	13
3101 (Basque)	14	10	27	24	10	13	13	14
3301 (Basque)	14	10	26	24	11	13	13	15
5501 (Basque)	14	11	26	24	11	13	13	16
6103 (Basque)	14	10	26	24	11	13	13	15
2801 (Béarnais)	14	11	26	24	11	13	13	16
3001 (Béarnais)	14	10	26	24	10	13	13	17
3701 (Béarnais)	14	10	26	24	11	13	14	18
3901 (Béarnais)	14	10	26	24	10	13	13	17
4604 (Béarnais)	14	10	26	24	11	13	13	15
4901 (Béarnais)	14	11	26	24	11	13	14	19
m256 (Irish)	14	10	27	25	11	13	13	20
m285 (Norwegian)	14	10	27	24	10	13	13	14
m288 (Norwegian)	14	10	27	24	11	13	13	21
m293 (Italian)	14	10	27	24	11	13	13	21
m366 (Italian)	14	10	27	23	11	13	13	22
7v (Basque)	14	9	27	24	11	13	13	23
16v (Basque)	14	9	27	24	11	13	13	23
21v (Basque)	13	10	27	24	10	13	13	24
23v (Basque)	14	10	27	24	11	11	13	25
24v (Basque)	14	10	27	24	11	13	13	21
25v (Basque)	14	11	27	23	11	13	14	26
30v (Basque)	14	10	27	25	10	13	12	27
41v (Basque)	14	11	28	24	10	13	13	28
44v (Basque)	14	10	27	24	10	13	13	14
8c (Catalan)	14	10	27	24	10	13	13	14
15c (Catalan)	14	10	27	24	11	13	13	21
17c (Catalan)	14	10	27	24	11	13	13	21
20c (Catalan)	14	10	27	24	11	11	13	25
26c (Catalan)	14	10	27	24	11	11	13	25
30c (Catalan)	14	10	27	24	11	13	13	21
32c (Catalan)	15	10	27	24	10	13	13	29
36c (Catalan)	14	11	27	24	11	11	13	30
43c (Catalan)	14	10	27	23	11	14	14	31

[a] As defined by Kayser et al. (1997).

group-22 microsatellite haplotypes from those in haplogroup 1. The same picture is also evident in MSY1 code diversity (Jobling et al. [1998], and data not shown).

Discussion

The Y chromosome has several properties that make it useful for evolutionary studies and that should make

Table 4

Estimates of Age of SRY-2627 Mutation

Method	Microsatellite Age (95% CI) (years)	MSY1 Age (Range) (years)
Mean mutations from root[a]	2,693 (1,154–9,425)	1,107 (604–3,320)
ASD[b]	3,452 (1,480–12,083)	2,632 (1,555–8,554)
Accumulated variance[c]	3,116 (1,166–16,001)	2,328 (1,217–10,143)
Coalescent[d]	1,650 (1,044–8,248)	Not done

[a] From Bertranpetit and Calafell (1996).
[b] From Goldstein et al. (1995).
[c] From Thomas et al. (1998).
[d] From Wilson and Balding (1998).

it simpler to analyze than the "grande dame" of molecular evolutionary biology, mtDNA. One of these properties is the Y chromosome's comparatively low base-substitution mutation rate: in the case of mtDNA, the rate is so high that many polymorphic bases have been multiply substituted since the human-chimp divergence, and trees cannot easily be rooted; on the Y chromosome, in contrast, unambiguous ancestral-state information should be obtained by analysis of the DNAs of other primates. Here, however, we have shown that this is not always straightforward.

Haplogroup-22 chromosomes are rare or absent in most of the world's populations and are most common in Iberia or in populations with substantial Iberian ancestry (Bianchi et al. 1997); within Iberia, the highest frequencies are found in Basques and Catalans, who speak languages belonging to different language families. Either the SRY-2627 mutation was present in a population that was ancestral to both populations and that spoke a single language, or it has occurred since linguistic divergence, implying gene flow over a linguistic barrier.

Contemporaneous evidence on linguistic prehistory does not exist, and, indeed, written records of Basque date back only 900 years (Collins 1986). However, its lack of linguistic relatives strongly suggests that the Basque language is ancient. Theories of Basque origins are many and varied. One theory, "Vasco-Iberism" (Lafon 1972), sees Basque as the last remnant of a language, Ibero, spoken in much of Iberia, including the northern part of modern Catalonia, before the Roman conquest; if this were true, then the linguistic divergence between Basques and Catalans might date back only 2 millennia, and our findings might then be taken to support the hypothesis. However, alleged similarities between Basque and Ibero rest on the scanty evidence of a few inscriptions and place names and are not supported by modern linguists (de Hoz 1995); Vasco-Iberism also seems inconsistent with information from sources such as Greek and Roman geographers (Collins 1986). Alternatively, it might be thought that contraction of Basque from a previously greater territory could have resulted from the arrival of Indo-European speakers (Barbujani et al. 1994) during the Neolithic period, 4,000–6,000 years ago (Menozzi et al. 1978; Renfrew 1989)—dates that are included in our wide confidence intervals. However, the influence of Indo-European languages here was probably minor, with the non-Basque territory remaining non–Indo-European speaking until the arrival of the Romans.

To explore this issue further, we used microsatellite diversity to calculate ASD between the Basque and Catalan samples within haplogroup 22 and so to estimate the time of divergence of these two populations of haplogroup-22 chromosomes. ASD between all haplogroup-22 chromosomes and the root haplotype is 0.290 (equivalent to μt, where μ is the mutation rate and t the time in generations), and ASD between Basque and Catalan chromosomes, with correction for intrapopulation variance (equivalent to $2\mu t$, since we are no longer considering distance to a root) is 0.115. The age of divergence, as a percentage of the age of haplogroup 22, can then be calculated as the ASD between Basque and Catalan chromosomes, divided by twice the ASD between all haplogroup-22 chromosomes and the root, and is ~20%. Thus, the divergence between these populations of chromosomes is not ancient, and this supports the interpretation that there has been male-mediated gene flow directly between Basques and Catalans since the establishment of the languages now spoken. It also remains possible that haplogroup-22 chromosomes have been contributed to both populations by a third, unsampled population. In either case, genes have flowed over the substantial linguistic barrier that lies between Basque and an Indo-European language.

Can we see evidence of this inferred gene flow in patterns of allele sharing at non–Y-chromosome loci? Published data on mtDNA (Côrte-Real et al. 1996) and HLA (Comas et al. 1998) in Basques and Catalans show no evidence for the sharing of any population-specific alleles or haplotypes. It is, however, striking that, whereas Basque and Catalan samples cluster significantly to-

gether in a neighbor-joining tree based on seven HLA loci (Comas et al. 1998), genetic distances calculated from mtDNA diversity are greatest between Catalans and all other Iberian samples, including Basques (Côrte-Real et al. 1996). This contrast between biparentally and maternally inherited loci may imply that the sharing of Y-chromosomal lineages that we observe is really a result of male-mediated gene flow, with little female-mediated flow and with autosomal markers reflecting an average of the two. Higher-resolution studies of Iberian Y-chromosome diversity, analyzing all available lineages, should further delineate genetic boundaries within this region.

In principle, the direction of gene flow between Basques and Catalans could be addressed by examination of the population distribution of root haplotypes; however, this has not been done here, because a combination of small sample size and uncertainty about the identification of these roots is likely to make such an analysis inaccurate.

The SRY-2627 polymorphism represents another example of the geographic specificity of Y-chromosome lineages, a phenomenon resulting from patrilocality (Seielstad et al. 1998) and cultural influences on mating practices, as well as from the small effective population size of the chromosome, which make it particularly susceptible to drift. When we find non-Iberian examples of haplogroup-22 chromosomes, they are likely to represent emigrants from Iberia. The finding of "Iberian" lineages in South America is not unexpected; their dates and places of origin are amenable to historical analysis, and they may provide a useful way to estimate the extent of admixture between indigenous people and Iberian colonists. Their occurrence in France, Germany, and England is more difficult to interpret, however. The young age of haplogroup 22 means that they cannot be adduced as support for the hypothesized "out of Iberia" migration 10,000–15,000 years ago, proposed on the basis of the distribution of mtDNA haplogroup V (Torroni et al. 1998). Population pairwise R_{ST} (Slatkin 1995) analysis of microsatellite data for the Iberian versus non-Iberian samples shows a significant difference ($P < .05$) between the two, which may be a sample-size effect but may also tell us that the emigrants are not very recent. In support of this, when information is available on the surnames of these individuals, these are typical of the populations in which they were found (data not shown). Surnames in most European populations came into existence after the 13th century (Hassall 1967), and therefore (if we set aside the complicating factors of nonpaternity and local-surname adoption) this suggests that these emigrants may predate this period. This is no proposal for an early origin for tourism: there are many possible causes of such long-distance gene flow—for instance, it is known that the Roman army recruited cohorts of Basque soldiers, who served as far afield as Hungary, the lower Rhine, and northern England (Collins 1986; Perex Agorreta 1986).

Acknowledgments

We thank Santos Alonso, John Armour, Yuri Dubrova, Manfred Kayser, John Mitchell, and Marisol Rodriguez-Calvo, for DNA samples; Mourad Sahbatou, for information about CEPH pedigrees; Paul Taylor, for MSY1 codes of some of the haplogroup-22 males; and Lluïsa Vilageliu, for assistance. Sample collection was partly funded by multidisciplinary grant PR182/96-6745 from Complutense University (Madrid) and was performed with the help of the Analysis Laboratory of the Spanish Civil Guard. J.B. acknowledges the support of grants PB95-0267-C02-01,from DGICYT (Spain), and 1995SGR00205, from Generalitat de Catalunya. M.E.H. was supported by an MRC studentship, M.S. by a Nuffield Foundation Undergraduate Research Bursary, and C.T.-S. by the CRC. M.A.J. is a Wellcome Senior Research Fellow in Basic Biomedical Science and was supported by a Wellcome Trust Career Development Fellowship (grant 044910).

References

Bandelt H-J, Forster P, Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16:37–48

Barbujani G (1991) What do languages tell us about human microevolution? Trends Ecol Evol 6:151–156

Barbujani G, Pilastro A, de Domenico S, Renfrew C (1994) Genetic variation in North Africa and Eurasia: neolithic demic diffusion vs. paleolithic colonisation. Am J Phys Anthropol 95:137–154

Barbujani G, Sokal RR (1990) Zones of sharp genetic change in Europe are also linguistic boundaries. Proc Natl Acad Sci USA 87:1816–1819

Bertranpetit J, Calafell F (1996) Genetic and geographical variability in cystic fibrosis: evolutionary considerations. In: Chadwick D, Cardew G (eds) Variation in the human genome. John Wiley & Sons, New York, pp 97–118

Bertranpetit J, Cavalli-Sforza LL (1991) A genetic reconstruction of the history of the population of the Iberian Peninsula. Ann Hum Genet 55:51–67

Bertranpetit J, Sala J, Calafell F, Underhill PA, Moral P, Comas D (1995) Human mitochondrial DNA variation and the origin of Basques. Ann Hum Genet 59:63–81

Bianchi NO, Bailliet G, Bravi CM, Carnese RF, Rothhammer F, Martínez-Marignac VL, Pena SDJ (1997) Origin of Amerindian Y-chromosomes as inferred by the analysis of six polymorphic markers. Am J Phys Anthropol 102:79–89

Calafell F, Bertranpetit J (1994*a*) Mountains and genes: population history of the Pyrenees. Hum Biol 66:823–842

——— (1994*b*) Principal component analysis of gene frequencies and the origin of Basques. Am J Phys Anthropol 93:201–215

Collins R (1986) The Basques. Blackwell, Oxford

Comas D, Calafell F, Mateu E, Pérez-Lezaun A, Bosch E, Ber-

tranpetit J (1997) Mitochondrial DNA and the origin of the Europeans. Hum Genet 99:443–449

Comas D, Mateu E, Calafell F, Pérez-Lezaun A, Bosch E, Martínez-Arias R, Bertranpetit J (1998) HLA class I and class II DNA typing and the origin of Basques. Tissue Antigens 51:30–40

Cooper G, Amos W, Hoffman D, Rubinsztein DC (1996) Network analysis of human Y microsatellite haplotypes. Hum Mol Genet 5:1759–1766

Côrte-Real HBSM, Macaulay VA, Richards MB, Hariti G, Issad MS, Cambon-Thomsen A, Papiha S, et al (1996) Genetic diversity in the Iberian Peninsula determined from mitochondrial sequence analysis. Ann Hum Genet 60:331–350

de Hoz J (1995) El poblamiento antiguo de los Pirineos desde el punto de vista lingüístico. In: Bertranpetit J, Vives E (eds) Munanyes i població—el passat dels Pirineus des d'una perspectiva multidisciplinària. Centre de Trobada de les Cultures Pirinenques, Andorra, pp 271–299

Goldstein DB, Zhivotovsky LA, Nayar K, Linares AR, Cavalli-Sforza LL, Feldman MW (1996) Statistical properties of the variation at linked microsatellite loci: implications for the history of human Y chromosomes. Mol Biol Evol 13: 1213–1218

Hammer MF (1995) A recent common ancestry for human Y chromosomes. Nature 378:376–378

Hassall WO (1967) History through surnames. Pergamon Press, Oxford

Heyer E, Puymirat J, Dieltjes P, Bakker E, de Knijff P (1997) Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. Hum Mol Genet 6:799–803

Hurles ME, Irven C, Nicholson J, Taylor PG, Santos FR, Loughlin J, Jobling MA, et al (1998) European Y-chromosomal lineages in Polynesia: a contrast to the population structure revealed by mtDNA. Am J Hum Genet 63: 1793–1806

Jobling MA, Bouzekri N, Taylor PG (1998) Hypervariable digital DNA codes for human paternal lineages: MVR-PCR at the Y-specific minisatellite, MSY1 (*DYF155S1*). Hum Mol Genet 7:643–653

Jobling MA, Tyler-Smith C (1995) Fathers and sons: the Y chromosome and human evolution. Trends Genet 11:449–456

Karafet TM, Zegura SL, Posukh O, Osipova L, Bergen A, Long J, Goldman D, et al (1999) Ancestral Asian source(s) of New World Y-chromosome founder haplotypes. Am J Hum Genet 64:817–831

Kayser M, Caglià A, Corach D, Fretwell N, Gehrig C, Graziosi G, Heidorn F, et al (1997) Evaluation of Y-chromosomal STRs: a multicenter study. Int J Legal Med 110:125–133

Lafon R (1972) Basque: pour la comparison du basque et des langues caucasiques. Bedi Kartlisa 27:7–23

Lahermo P, Sajantila A, Sistonen P, Lukka M, Aula P, Peltonen L, Savontaus M-L (1996) The genetic relationship between the Finns and the Finnish Saami (Lapps): analysis of nuclear DNA and mtDNA. Am J Hum Genet 58:1309–1322

Lucotte G, Hazout S (1996) Y chromosome DNA haplotypes in Basques. J Mol Evol 42:472–475

Mathias N, Bayés M, Tyler-Smith C (1994) Highly informative compound haplotypes for the human Y chromosome. Hum Mol Genet 3:115–123

Menozzi P, Piazza A, Cavalli-Sforza LL (1978) Synthetic maps of human gene frequencies in Europeans. Science 201: 786–792

Minch E (1997) Microsat 1.5d. Department of Genetics, University of Stanford, Stanford

Mitchell RJ, Hammer MF (1996) Human evolution and the Y chromosome. Curr Opin Genet Dev 6:737–742

Mourant AE (1947) The blood groups of the Basques. Nature 160:505

——— (1983) Blood relations. Oxford University Press, Oxford

Ngo KY, Vergnaud G, Johnsson C, Lucotte G, Weissenbach J (1986) A DNA probe detecting multiple haplotypes of the human Y chromosome. Am J Hum Genet 38:407–418

Perex Agorreta MJ (1986) Los vascones: el poblamiento en época romana. Gobierno de Navarra Departamento de Educación y Cultura, Institución Príncipe de Viana, Navarra, pp 63–69

Pérez-Lezaun A, Calafell F, Seielstad M, Mateu E, Comas D, Bosch E, Bertranpetit J (1997) Population genetics of Y-chromosome short tandem repeats in humans. J Mol Evol 45:265–270

Poloni ES, Semino O, Passarino G, Santachiara-Benerecetti AS, Dupanloup I, Langaney A, Excoffier L (1997) Human genetic affinities for Y-chromosome P49a,f/*Taq*I haplotypes show strong correspondence with linguistics. Am J Hum Genet 61:1015–1035

Renfrew C (1989) The origins of Indo-European languages. Sci Am 261:106–114

Ruhlen M (1991) A guide to the world's languages. Edward Arnold, London

Santos FR, Tyler-Smith C (1996) Reading the human Y chromosome: the emerging DNA markers and human genetic history. Braz J Genet 19:665–670

Schneider S, Kueffer J-M, Roessli D, Excoffier L (1997) Arlequin ver 1.1: a software for population genetic data analysis. Genetics and Biometry Laboratory, University of Geneva, Geneva

Seielstad MT, Minch E, Cavalli-Sforza LL (1998) Genetic evidence for a higher female migration rate in humans. Nat Genet 20:278–280

Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. Genetics 139:457–462

Thomas MG, Skorecki K, Ben-Ami H, Parfitt T, Bradman N, Goldstein DB (1998) Origins of Old Testament priests. Nature 394:138–140

Torroni A, Bandelt H-J, D'Urbano L, Lahermo P, Moral P, Sellitto D, Rengo C, et al (1998) mtDNA analysis reveals a major late Paleolithic population expansion from Southwestern to Northeastern Europe. Am J Hum Genet 62:1137–1152

Urtasun M, Sáenz A, Roudaut C, Poza JJ, Urtizberea JA, Cobo AM, Richard I, et al (1998) Limb-girdle muscular dystrophy in Guipúzcoa (Basque Country, Spain). Brain 121: 1735–1747

Veitia R, Ion A, Barbaux S, Jobling MA, Souleyreau N, Ennis K, Ostrer H, et al (1997) Mutations and sequence variants in the testis-determining region of the Y chromosome in

individuals with a 46,XY female phenotype. Hum Genet 99: 648–652

Wilson IJ, Balding DJ (1998) Genealogical inference from microsatellite data. Genetics 150:499–510

Zerjal T, Dashnyam B, Pandya A, Kayser M, Roewer L, Santos FR, Schiefenhövel W, et al (1997) Genetic relationships of Asians and northern Europeans, revealed by Y-chromosomal DNA analysis. Am J Hum Genet 60:1174–1183

Appendix D

Source code of programs written during the course of this thesis. All programs were written in Interactive Data Language 5.1.

1. **EuroBarrierSig**

2. **DisplayBarrierSig**

3. **EuroDelaunay**

4. **OceanicDistances**

EuroBarrierSig

; This permutation program is specific for 8 lineages (haplogroups) at 43 sample sites in Europe

;First section inputs the observed data and calculates the summed derivative from the observed data

```
print,'Start Time',today()
;define data array for eight haplogroups and forty-three sample sites lineage=intarr(8,43)
percent=fltarr(8,43)
percentlin=fltarr(8,43)
derRes=fltarr(100,100,8)
derReslin=fltarr(100,100,8)
dertotall=fltarr(100,100,1000)
rowtotals=fltarr(43)

; input haplogroup data [Hg1,Hg2+9,Hg3,Hg4+21+8,Hg12,Hg16,Hg22,Hg26]
lineage(*,0)=[36,33,11,3,0,0,1,0]; haplogroup data for Dutch, row 1
lineage(*,1)=[20,12,2,4,0,0,2,0]; haplogroup data for French
lineage(*,2)=[4,18,3,10,0,0,0,1]; haplogroup data for Greek
lineage(*,3)=[1,2,3,1,1,10,0,0]; haplogroup data for Finnish
lineage(*,4)=[4,11,3,4,0,0,0,0]; haplogroup data for Bulgarian
lineage(*,5)=[97,53,15,5,0,0,1,1]; haplogroup data for East Anglian
lineage(*,6)=[38,22,12,6,0,0,2,0]; haplogroup data for Bavarian
lineage(*,7)=[10,8,4,1,0,1,0,0]; haplogroup data for Turkish
lineage(*,8)=[1,3,13,0,0,2,0,0]; haplogroup data for Polish
lineage(*,9)=[15,23,26,5,0,0,1,0]; haplogroup data for Slovenian
lineage(*,10)=[4,3,0,3,0,0,0,0]; haplogroup data for Sardinian
lineage(*,11)=[86,21,3,12,0,0,3,1]; haplogroup data for Spanish
lineage(*,12)=[23,18,2,0,0,0,0,1]; haplogroup data for Danish
lineage(*,13)=[4,15,16,4,0,1,0,1]; haplogroup data for Belorussian
lineage(*,14)=[2,5,3,0,0,3,0,3]; haplogroup data for Chuvash
lineage(*,15)=[59,19,5,2,0,0,3,2]; haplogroup data for Madrid
lineage(*,16)=[19,2,0,0,0,0,5,0]; haplogroup data for Basque
lineage(*,17)=[12,7,9,0,0,1,0,1]; haplogroup data for German
```

```
lineage(*,18)=[35,7,3,0,0,0,0,0]; haplogroup data for Scottish
lineage(*,19)=[44,34,2,13,0,0,0,6]; haplogroup data for Italian
lineage(*,20)=[11,11,8,6,0,0,0,0]; haplogroup data for Hungarian
lineage(*,21)=[5,4,14,0,0,11,0,0]; haplogroup data for Latvian
lineage(*,22)=[2,14,10,0,0,13,0,1]; haplogroup data for Estonian
lineage(*,23)=[0,9,11,2,0,4,0,0]; haplogroup data for Russian
lineage(*,24)=[5,5,14,0,8,16,0,0]; haplogroup data for Mari
lineage(*,25)=[2,5,13,0,0,18,0,0]; haplogroup data for Lithuanian
lineage(*,26)=[3,15,10,0,0,20,0,0]; haplogroup data for Saami
lineage(*,27)=[0,11,3,0,0,25,0,0]; haplogroup data for Finnish
lineage(*,28)=[11,23,9,1,0,4,0,0]; haplogroup data for Swedish
lineage(*,29)=[15,18,16,1,0,2,0,0]; haplogroup data for Norwegian
lineage(*,30)=[11,38,10,0,0,4,0,1]; haplogroup data for Gotland
lineage(*,31)=[16,18,46,6,9,62,0,10]; haplogroup data for Estonian
lineage(*,32)=[7,18,46,6,5,13,0,1]; haplogroup data for Russian
lineage(*,33)=[23,54,5,3,0,8,0,2]; haplogroup data for Armenian
lineage(*,34)=[9,22,9,3,0,0,1,1]; haplogroup data for Romanian
lineage(*,35)=[4,34,1,11,0,0,0,0]; haplogroup data for Turkish
lineage(*,36)=[207,41,2,6,0,1,0,0]; haplogroup data for Irish
lineage(*,37)=[5,4,2,1,0,0,0,0]; haplogroup data for Danish
lineage(*,38)=[11,57,16,13,2,0,0,1]; haplogroup data for Yugoslavian
lineage(*,39)=[32,13,1,10,0,0,0,1]; haplogroup data for Portugese HC-R
lineage(*,40)=[203,75,1,35,0,0,6,9]; haplogroup data for Portugese M-JP
lineage(*,41)=[58,26,4,2,0,0,1,1]; haplogroup data for Belgian
lineage(*,42)=[1,13,5,2,3,3,0,0]; haplogroup data for Ukraine


rowtotals=fix(total(lineage,1));total each row and put into an array
columntotals=fix(total(lineage,2));ditto for each column

M=fix(total(lineage));M is the total of the array

; convert lineage samples into percentages
FOR j=0,42 DO BEGIN; loop down for all sample sites

        FOR k=0,7 DO BEGIN; loop across for all haplogroups
```

```
                percentlin(k,j)=lineage(k,j)*100./rowtotals(j)

        ENDFOR

ENDFOR

; input the latitudes of the sample sites in degrees
lat=[52.3,48.9,38,60.1,42.7,52.6,48.1,41,51.7,46.1,39.2,40.4,55.7,53.9,$
55.5,40.4,43.3,52.5,56,41.9,47.5,56.9,59.4,55.8,56.5,54.7,68,60.1,63.7,$
59.9,57.5,59.4,55.8,40.2,44.4,41,53.3,55.7,44.8,38.7,41.2,50.8,50.4]

;input the longitudes of the sample sites in degrees
lon=[4.9,2.3,23.7,25,23.3,1.3,11.6,29,19.5,14.5,9.1,-3.7,12.6,27.5,47,$
-3.7,-2.9,13.4,-3.2,12.5,19.1,24.1,24.7,37.7,48,25.3,22,25,20.3,10.8,$
18.5,24.7,37.7,44.5,26.1,29,-6.3,12.6,20.5,-9.1,-8.6,4.3,30.5]

; convert degrees to radians for sample sites, trigonometric functions work in radians
latr=lat*3.1415/180
lonr=lon*3.1415/180

latgrid=fltarr(100)
longrid=fltarr(100)
distance=fltarr(100,100,43)
distance20=fltarr(100,100,20);distance to the closest 20 sample sites
; for each grid point, haplogroup frequencies of the nearest 20 sample sites
dataset=fltarr(100,100,8,20)
; for each permuted grid point, haplogroup frequencies of the nearest 20 sample sites
newdataset=fltarr(100,100,8,20)
invdistsqr=dblarr(100,100,20)
suminvdistsqr=dblarr(100,100)


FOR i=0,99 DO BEGIN; for each grid point lat and lon
```

; calculate the lat and lon of the grid points to be used, in radians, the number here sets the resolution of the interpolation.

latgrid(i)= (37+0.33*i)*3.1415/180; *(33/resolution) make interpolated points at intervals defined by the resolution required*

longrid(i)= (-10+0.6*i)*3.1415/180; *(60/resolution)*

ENDFOR

values=fltarr(100,100,8)
freqD=dblarr(100,100,8,20)
temp1lin=fltarr(100,100)
temp2lin=fltarr(100,100)
gradlin=fltarr(100,100)
derReslin=fltarr(100,100,8)
derTotlin=fltarr(100,100)
indexstore=intarr(100,100,20)

FOR i=0,99 DO BEGIN

FOR j=0,99 DO BEGIN; *for each grid point lon*

FOR s=0,42 DO BEGIN; *for each sample site*

; calculate using Great Circle Distances the distance between each grid point and each sample point

distance(i,j,s)=3963*acos((sin(latr(s))*sin(latgrid(i)))+cos(latr(s))*cos(latgrid(i))*cos(long rid(j)-lonr(s)))

ENDFOR

; sort these distances together with the haplogroup frequencies and only choose the closest 20 sample sites, from which to calculate the value at each grid point, the number of sample sites used can be varied.

```
index=sort(distance(i,j,*))

indexstore(i,j,*)=index(0:19); store for use in replications

distance20(i,j,*)=distance(i,j,index(0:19))
```

; *for each of the nearest 20 sample sites calculate distance squared*

```
FOR s=0,19 DO BEGIN
```

; *square and inverse the distances*

```
invdistsqr(i,j,s)=(1/((distance20(i,j,s))^2))

ENDFOR

suminvdistsqr(i,j)=total(invdistsqr(i,j,*))
```

;calculate the values at the grid point for each haplogroup surface by inverse distance weighted interpolation

```
FOR h=0,7 DO BEGIN;for each lineage
```

;sort each haplogroup frequency according to distance

```
dataset(i,j,h,*)=percentlin(h,(index(0:19)))
```

; *for each of the nearest 20 sample sites calulate frequency multiplied by the inverse distance squared*

```
FOR s=0,19 DO BEGIN
```

; *calculate the freq divided by d squared at each site*

```
freqD(i,j,h,s)=(dataset(i,j,h,s)*invdistsqr(i,j,s))

ENDFOR
```

```
;next step calculates the value for a given haplogroup at each grid point by dividing the sum of the
frequency/d2 from each sample site by the sum of 1/d2

                values(j,i,h)=(total(total(freqD(i,j,h,*)))/suminvdistsqr(i,j))


            ENDFOR

        ENDFOR

ENDFOR
;calculate summed derivative from the interpolations from the observed data


FOR h=0,7 DO BEGIN

        FOR i=0,99 DO BEGIN

            temp1lin(*,i)=deriv(values(*,i,h))

            temp2lin(i,*)=deriv(values(i,*,h))

; loop to calculate the magnitude of dy/dx
            FOR o=0,99 DO BEGIN

                gradlin(i,o)=sqrt(temp1lin(i,o)^2+temp2lin(i,o)^2)

            ENDFOR

        ENDFOR

; puts the new array into the 3D stack and makes all gradients positive
        derReslin(*,*,h)=gradlin

ENDFOR
```

```
; add together derivatives of all haplogroups
dertotlin=total(derReslin,3)



; plot a picture to make sure that the observed data make sense before proceding with the
replications

;plot observed barriers against geography

window,3,ret=2,xsize=800,ysize=550; make new window and set size
loadct,26; define colour
map_set,limit=[38,-9.1,68,48] ; set region for map

; warp image to fit map observed=MAP_IMAGE(dertotlin,startx,starty,xsize,ysize,$
        latmin=37,latmax=70,lonmin=-10,lonmax=50)

tvscl,observed,startx,starty

map_continents,/coasts,color=218,mlinethick=2; overlay continents

map_continents,/countries,color=218,mlinethick=2; overlay countries

plots,lon,lat, psym=2, color=150, symsize=2; plot sample points


;set parameters for replication

y=intarr(8)

; calculate column totals (which are held constant throughout the permutation procedure)
FOR g=0,7 DO BEGIN

        y(g)=TOTAL(lineage(g,*))
```

```
ENDFOR

random=intarr(2,M);define random array

ones=columntotals(0)-1; column totals in the random array to aid accessing it.
twos=ones+columntotals(1)
threes=twos+columntotals(2)
fours=threes+columntotals(3)
fives=fours+columntotals(4)
sixes=fives+columntotals(5)
sevens=sixes+columntotals(6)
eights=sevens+columntotals(7)


row1=rowtotals(0)-1;ditto for rows
row2=row1+rowtotals(1)
row3=row2+rowtotals(2)
row4=row3+rowtotals(3)
row5=row4+rowtotals(4)
row6=row5+rowtotals(5)
row7=row6+rowtotals(6)
row8=row7+rowtotals(7)
row9=row8+rowtotals(8)
row10=row9+rowtotals(9)
row11=row10+rowtotals(10)
row12=row11+rowtotals(11)
row13=row12+rowtotals(12)
row14=row13+rowtotals(13)
row15=row14+rowtotals(14)
row16=row15+rowtotals(15)
row17=row16+rowtotals(16)
row18=row17+rowtotals(17)
row19=row18+rowtotals(18)
row20=row19+rowtotals(19)
row21=row20+rowtotals(20)
row22=row21+rowtotals(21)
```

```
row23=row22+rowtotals(22)
row24=row23+rowtotals(23)
row25=row24+rowtotals(24)
row26=row25+rowtotals(25)
row27=row26+rowtotals(26)
row28=row27+rowtotals(27)
row29=row28+rowtotals(28)
row30=row29+rowtotals(29)
row31=row30+rowtotals(30)
row32=row31+rowtotals(31)
row33=row32+rowtotals(32)
row34=row33+rowtotals(33)
row35=row34+rowtotals(34)
row36=row35+rowtotals(35)
row37=row36+rowtotals(36)
row38=row37+rowtotals(37)
row39=row38+rowtotals(38)
row40=row39+rowtotals(39)
row41=row40+rowtotals(40)
row42=row41+rowtotals(41)
row43=row42+rowtotals(42)
```

; Second section uses a random permutation algorithm to permute the data set 1000 times keeping sample sizes at each sample site and overall haplogroup frequencies constant

```
;start big loop (set to 1000  replications), to generate random arrays
FOR p=0,999 DO BEGIN
```

```
; to check progress of the program and estimate iteration runtime
print, p, ' replications done'
```

```
;generate random array (numbers between 0 and 10000)
        FOR i=0,M-1 DO BEGIN
```

```
        random(0,i)=fix(randomu(1)*10000)

    ENDFOR


;sets right hand side of random array to ones, twos etc.    random(1,0:ones)=1
    random(1,ones+1:twos)=2
    random(1,twos+1:threes)=3
    random(1,threes+1:fours)=4
    random(1,fours+1:fives)=5
    random(1,fives+1:sixes)=6
    random(1,sixes+1:sevens)=7
    random(1,sevens+1:eights)=8


    random0=random(0,*);just picks out left hand side of array
    random1=random(1,*);and right
;generate a sorting index from the random numbers   index=sort(random0)

    newarr=intarr(8,43);new array to hold new permutation


;reorders right hand side of random array using above index  newrandom=random1(index)

; enables a  loop to generate new array
    place=[-1,row1,row2,row3,row4,row5,row6,row7,row8,row9,row10,row11,row12,
row13,row14,row15,row16,row17,row18,row19,row20,row21,row22,row23,
row24,row25,row26,row27,row28,row29,row30,row31,row32,row33,row34,
row35,row36,row37,row38,row39,row40,row41,row42,row43]


;placing number of ones etc. into the new array, looping downwards    FOR    j=0,42    DO
BEGIN

        FOR K=0,7 DO BEGIN ;looping across
```

```
                temp=where(newrandom(place(J)+1:place(J+1)) eq (K+1),count)

                newarr(K,J)=count

; converts the haplogroup info into percentages as input for the interpolation procedures
        percent(k,j)=newarr(k,j)*100./rowtotals(j)

                ENDFOR

        ENDFOR


newvalues=dblarr(100,100,8)
newfreqD=dblarr(100,100,8,20)

        FOR i=0,99 DO BEGIN; for each grid point lat

                FOR j=0,99 DO BEGIN; for each grid point lon

                        FOR h=0,7 DO BEGIN; for each lineage

; sort each haplogroup frequency according to distance and input only the Hg frequencies for the closest 20
                                newdataset(i,j,h,*)=percent(h,(indexstore(i,j,*)))
; for each of the nearest 20 sample sites calulate frequency multiplied by the inverse distance squared
                                FOR s=0,19 DO BEGIN

; calculate the freq divided by d squared at each site
        newfreqD(i,j,h,s)=(newdataset(i,j,h,s)*invdistsqr(i,j,s))

                                ENDFOR

;next step calculates the value for a given haplogroup at each grid point by dividing the sum of the frequency/d2 from each sample site by the sum of 1/d2
```

```
            newvalues(j,i,h)=(total(total(newfreqD(i,j,h,*)))/suminvdistsqr(i,j))

                ENDFOR

            ENDFOR

        ENDFOR


temp1=dblarr(100,100)
temp2=dblarr(100,100)
grad=dblarr(100,100)
derRes=dblarr(100,100,8)
derTot=dblarr(100,100)


        FOR k=0,7 DO BEGIN; for each haplogroup

            FOR r=0,99 DO BEGIN; calculate df/dx and df/dy

                temp1(*,r)=deriv(newvalues(*,r,k))

                temp2(r,*)=deriv(newvalues(r,*,k))

; loop to calculate the magintude of dy/dx and makes all gradients positive
                FOR o=0,99 DO BEGIN

                    grad(r,o)=sqrt(temp1(r,o)^2+temp2(r,o)^2)

                ENDFOR

            ENDFOR

; puts the new array into the 3D stack
        derRes(*,*,k)=grad
```

```
        ENDFOR

; add together derivatives of all haplogroups
                dertot=total(derRes,3)


; put the summed derivative output for each permuted array into the 3D stack

        dertotall(*,*,p)=dertot

ENDFOR ;end loop do next iteration


;  compare stop time to start time to see how long the progrm takes to run
print,'Stop Time', today()

END
```

DisplayBarrierSig

; convert the output of the EuroBarrierSig program by applying the significance filters and displaying the results on a map

; first define the true top ten percent

```
newNumSig=intarr(100,100)

FOR x=0,99 DO BEGIN; loop for each gridded point

        FOR y=0,99 DO BEGIN

; collate all observed values within the landscape smaller than the observed value at each point
                newbin = where(dertotlin(*,*) Lt dertotlin(x,y))

; count number that are smaller than the value from the observed grid point
                newnumten = N_ELEMENTS(newbin)

                newNumSig(x,y)=newnumten; put the counts into an array

                IF (newNumSig(x,y) Gt 9000) THEN newNumSig(x,y)=1 ELSE
newNumSig(x,y) =0; set significant points to 1 and nonsignificant to 0

        ENDFOR

ENDFOR


;calculate the level of permuted significance

newsig=intarr(100,100)

FOR x=0,99 DO BEGIN; loop for each gridded point
```

```
        FOR y=0,99 DO BEGIN

; collate all replicates smaller than observed values at each point
                bin = where(dertotall(x,y,*) Lt dertotlin(x,y))

; count number that are smaller than the value from the observed data set        num        =
N_ELEMENTS(bin)

                newsig(x,y)=num; put the counts into an array

                IF (newsig(x,y) Gt 950) THEN newsig(x,y)=1 ELSE newsig(x,y) =0; set
significant points to 1 and nonsignificant to 0

        ENDFOR

ENDFOR

;combine top ten per cent with the permuted 95% values

newTopTen=dertotlin*newnumsig


; multiply observed map by map of permuted significance to exclude those values which are not
significant

newfinal=newTopTen*newsig

permonly=dertotlin*newsig


;plot observed barriers against geography

window,4,ret=2,xsize=800,ysize=550; make new window and set size
loadct,26; define colour
map_set,limit=[38,-9.1,68,48] ; set region for map
```

```
; warp image to fit map
barriers=MAP_IMAGE(newfinal,startx,starty,xsize,ysize,$
        latmin=37,latmax=70,lonmin=-10,lonmax=50)

tvscl,barriers,startx,starty

map_continents,/coasts,color=218,mlinethick=2; overlay continents

map_continents,/countries,color=218,mlinethick=2; overlay countries

plots,lon,lat, psym=2, color=150, symsize=2; plot sample points

END
```

EuroDelaunay

```
; plots the sample sites onto a map of europe

; input the latitudes of the sample sites in degrees
lat=[52.3,48.9,38,60.1,42.7,52.6,48.1,41,51.7,46.1,39.2,40.4,55.7,53.9,$
55.5,40.4,43.3,52.5,56,41.9,47.5,56.9,59.4,55.8,56.5,54.7,68,60.1,63.7,$
59.9,57.5,59.4,55.8,40.2,44.4,41,53.3,55.7,44.8,38.7,41.2,50.8,50.4]

; input the longitudes of the sample sites in degrees
lon=[4.9,2.3,23.7,25,23.3,1.3,11.6,29,19.5,14.5,9.1,-3.7,12.6,27.5,47,$
-3.7,-2.9,13.4,-3.2,12.5,19.1,24.1,24.7,37.7,48,25.3,22,25,20.3,10.8,$
18.5,24.7,37.7,44.5,26.1,29,-6.3,12.6,20.5,-9.1,-8.6,4.3,30.5]

window,3,ret=2,xsize=840,ysize=600; make new window and set size
loadct,26; define colour
map_set,limit=[38,-9.1,68,48]; set region for map

map_continents,/coasts,color=218,mlinethick=2; overlay continent outlines

plots,lon,lat, psym=2, color=218, symsize=2; plot sample sites

TRIANGULATE, lon, lat, tr, b        ;Obtain triangulation.
FOR i=0, N_ELEMENTS(tr)/3-1 DO BEGIN & $    ;Show the triangles.
  t = [tr[*,i], tr[0,i]] & $        ;Subscripts of vertices [0,1,2,0].
  PLOTS, lon[t], lat[t], color=218, thick=2 & $      ;Connect triangles.

END
```

OceanicDistances

```
;program to calculate the geographical distances between samples sites in Oceania using Great Circle Distances (GCD).

; input the latitudes of the sample sites in degrees
lat=[1,-21.5,15,1.5,6,-9,-27,-14,-20,-16,24]

; input the longitudes of the sample sites in degrees
lon=[115,-160,121,155,172,147,-144,-172,-175,168,122]

; convert degrees to radians for sample sites, trigonometric functions work in radians
latr=lat*3.1415/180
lonr=lon*3.1415/180

distance=fltarr(11,11)

FOR i=0,10 DO BEGIN; for each sample site

        FOR j=0,10 DO BEGIN; to each sample site

; calculate using Great Circle Distances the distance between each grid point and each sample point

    distance(i,j)=3963*acos((sin(latr(i))*sin(latr(j)))+cos(latr(i))*cos(latr(j))*cos(lonr(j)-
lonr(i)))

        ENDFOR

ENDFOR

print, distance

END
```

Appendix E

1. Entire Oceanic data set: Haplogroup and MSY1 data.

2. Table of haplogroup 26 MSY1 diversity, sorted by modular structure, with number identifiers for reference to the Median-Joining networks in chapter 6.

Sample	Population	Haplogroup	MSY1
KG1	Kapamarangi	26	(3)2.(1)16.(3)29.(4)16
KG10	Kapamarangi	26	(1)11.(3)31.(4)20
KG14	Kapamarangi	26	(1)12.(3)26.(4)24
KG30	Kapamarangi	26	(1)12.(3)23.(4)27
KG31	Kapamarangi	26	(3)2.(1)14.(3)57.(4)14
KG37	Kapamarangi	10	(0)1.(1)13.(3)34.(4)5.(0)3.(4)17
KG39	Kapamarangi	26	(3)3.(1)12.(3)55.(4)13
KG41	Kapamarangi	26	(3)3.(1)12.(3)51.(4)13
KG47	Kapamarangi	26	(3)3.(1)12.(3)58.(4)12
KG51	Kapamarangi	26	(1)11.(3)31.(4)20
KG53	Kapamarangi	10	(0)1.(1)13.(3)34.(4)5.(0)3.(4)17
KG56	Kapamarangi	26	(3)3.(1)12.(3)51.(4)13
KG61	Kapamarangi	26	(1)11.(3)32.(4)19
KG62	Kapamarangi	26	(1)12.(3)26.(4)24
KG72	Kapamarangi	10	(0)1.(1)13.(3)34.(4)5.(0)3.(4)2.(0)2.(4)13
KG73	Kapamarangi	26	(3)3.(1)12.(3)52.(4)12
KG75	Kapamarangi	26	(1)12.(3)26.(4)24
KG76	Kapamarangi	10	(0)1.(1)13.(3)34.(4)5.(0)3.(4)17
KG78	Kapamarangi	26	(1)12.(3)24.(4)26
KG96	Kapamarangi	10	(0)1.(1)13.(3)34.(4)5.(0)3.(4)17
KG97	Kapamarangi	10	(1)13.(3)34.(4)6.(0)3.(4)16
MJ1	Majoro	26	(1)14.(3)39.(4)12
MJ7	Majoro	26	(3)2.(1)13.(3)56.(4)13
MJ12	Majoro	26	(1)15.(3)7.(0)1(3)37.(4)4
MJ13	Majoro	26	(3)3.(1)12.(3)57.(4)14
MJ15	Majoro	26	(1)12.(3)25.(4)24
MJ23	Majoro	26	(1)15.(3)7.(0)1(3)37.(4)4
MJ24	Majoro	26	(3)3.(1)12.(3)56.(4)14
MJ30	Majoro	26	(3)3(1)12(3)57(4)14
MJ36	Majoro	26	(3)3.(1)12.(3)58.(4)12
MJ42	Majoro	2	(3)3.(1)2.(0)1.(1)10.(3)38.(4)23
MJ83	Majoro	26	(3)3.(1)12.(3)55.(4)14
MJ90	Majoro	9	(3)3.(1)15.(3)29.(4)19
T4	Tongan	26	(3)3.(1)14.(3)59.(4)8
T5	Tongan	24	(1)13.(3)39.(4)12
T6	Tongan	10	(0)1.(1)13.(3)34.(4)5.(0)4.(4)16
T8	Tongan	24	(1)13.(3)39.(4)12
T14	Tongan	2	(3)3.(1)13.(3)42.(4)19
T18	Tongan	26	(3)2.(1)12.(3)62.(4)6
T20	Tongan	26	(3)2.(1)12.(3)62.(4)6
T22	Tongan	10	(0)1.(1)13.(3)33.(4)5.(0)4.(4)16
T37	Tongan	10	(0)1.(1)13.(3)35.(4)3.(0)3.(4)16
WS1	Western Samoa	26	(3)4.(1)11.(3)63.(4)7
WS10	Western Samoa	10	(0)1.(1)13.(3)34.(4)5.(0)3.(4)16
WS13	Western Samoa	26	(1)17.(3)39.(4)9
WS15	Western Samoa	10	(0)1.(1)13.(3)33.(4)6.(0)3.(4)17
WS18	Western Samoa	10	(0)1.(3)1.(1)12.(3)33.(4)3.(0)4.(4)29
WS19	Western Samoa	10	(0)1.(1)13.(3)34.(4)5.(0)3.(4)16
WS20	Western Samoa	10	(0)1.(1)13.(3)33.(4)25
WS21	Western Samoa	26	(3)2.(1)13.(3)63.(4)6
WS22	Western Samoa	10	(0)1.(1)13.(3)34.(4)5.(0)3.(4)16
WS24	Western Samoa	10	(0)1.(1)13.(3)34.(4)5.(0)3.(4)15
WS26	Western Samoa	10	(0)1.(1)13.(3)34.(4)6.(0)3.(4)16
WS29	Western Samoa	26	(3)2.(1)13.(3)64.(4)6
WS30	Western Samoa	10	(0)1.(1)13.(3)35.(4)5.(0)4.(4)15
WS31	Western Samoa	26	(3)2.(1)13.(3)66.(4)6
WS36	Western Samoa	26	(3)2.(1)13.(3)64.(4)6
WS40	Western Samoa	26	(3)2.(1)13.(3)64.(4)6
WS43	Western Samoa	10	(0)1.(1)13.(3)35.(4)5.(0)4.(4)15
WS47	Western Samoa	26	(3)2.(1)17.(3)28.(4)17
WS50	Western Samoa	10	(0)1.(1)13.(3)34.(4)5.(0)3.(4)16
WS54	Western Samoa	10	(0)1.(1)13.(3)34.(4)5.(0)3.(4)16
WS57	Western Samoa	26	(3)3.(1)14.(3)59.(4)3
WS64	Western Samoa	10	(0)1.(1)13.(3)35.(4)4.(0)3.(4)16
WS72	Western Samoa	26	(3)3.(1)14.(3)59.(4)3
WS87	Western Samoa	10	(0)1.(1)13.(3)34.(4)5.(0)3.(4)16
WS92	Western Samoa	10	(0)1.(1)13.(3)34.(4)5.(0)3.(4)15
WS97	Western Samoa	26	(1)11.(3)30.(4)22
WS98	Western Samoa	10	(0)1.(1)13.(3)34.(4)5.(0)3.(4)16
2002	Rapa	1	(1)17.(3)43.(4)14
2040	Rapa	18	(3)1.(1)17.(3)37.(4)17
2042	Rapa	18	(3)1.(1)17.(3)36.(4)16
2052	Rapa	26	(1)11.(3)30.(4)22
2085	Rapa	26	(1)11.(3)30.(4)22
2096	Rapa	1	(1)15.(3)37.(4)20
2102	Rapa	26	(1)11.(3)30.(4)22
2103	Rapa	26	(1)11.(3)30.(4)22
2104	Rapa	26	(1)11.(3)30.(4)22
2123	Rapa	10	(0)1.(1)13.(3)36.(4)4.(0)3.(4)15
2143	Rapa	26	(3)3.(1)14.(3)58.(4)8
2162	Rapa	10	(0)1.(1)13.(3)36.(4)4.(0)3.(4)15
2163	Rapa	10	(0)1.(1)13.(3)36.(4)4.(0)3.(4)15
2170	Rapa	18	(3)1.(1)17.(3)35.(4)16
2192	Rapa	10	(0)1.(1)13.(3)33.(4)6.(0)3.(4)16
2200	Rapa	1	(1)16.(3)34.(4)21
2202	Rapa	1	(1)16.(3)34.(4)21
2206	Rapa	1	(1)16.(3)34.(4)21
2231	Rapa	26	(1)11.(3)30.(4)22
2240	Rapa	26	(3)1(1)16(3)1(1)3(3)40(4)9
2255	Rapa	10	(0)1.(1)13.(3)34.(4)5.(0)3.(4)16
2260	Rapa	1	(1)16.(3)34.(4)21
2270	Rapa	1	(1)16.(3)36.(4)19
2280	Rapa	1	(1)17.(3)38.(4)18
2282	Rapa	1	(1)17.(3)38.(4)18
22407	Rapa	10	(0)1.(1)13.(3)36.(4)4.(0)3.(4)15
2310	Rapa	1	(1)16.(3)35.(4)20
TL8	Tongan lysates	26	(3)3.(1)14.(3)58.(4)8
TL14	Tongan lysates	24	(3)1.(1)12.(3)28.(4)19
TL18	Tongan lysates	10	(0)1.(3)1.(1)12.(3)34.(4)4.(0)4.(4)20
TL20	Tongan lysates	26	(3)3.(1)14.(3)59.(4)8
TL27	Tongan lysates	26	(3)3.(1)14.(3)59.(4)8
TL28	Tongan lysates	24	(3)1.(1)12.(3)29.(4)20
TL31	Tongan lysates	26	(3)3.(1)14.(3)59.(4)8
TL32	Tongan lysates	26	(3)3.(1)14.(3)59.(4)8

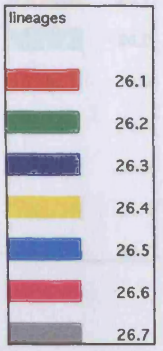
TL33	Tongan lysates	10	(0)1.(1)10.(3)33.(4)5.(0)3.(4)17
TL57	Tongan lysates	10	(1)14.(3)36.(4)3.(0)3.(4)15
TL63	Tongan lysates	26	(1)16.(3)42.(4)5
TL64	Tongan lysates	26	(3)2.(1)14.(3)58.(4)8
TL67	Tongan lysates	26	(1)16.(3)3.(1)2.(3)34.(4)10
TL68	Tongan lysates	24	(1)13.(3)39.(4)12
TL83	Tongan lysates	26	(3)3.(1)14.(3)59.(4)8
TL99	Tongan lysates	26	(3)3.(1)14.(3)60.(4)8
TL108	Tongan lysates	10	(0)1.(1)13.(3)34.(4)4.(0)3.(4)15
TL115	Tongan lysates	26	(3)1.(1)14.(3)63.(4)6
TL116	Tongan lysates	10	(0)1.(1)13.(3)33.(4)5.(0)3.(4)17
TL126	Tongan lysates	26	(3)3.(1)14.(3)59.(4)8
V15	Vav'ua	26	(3)3.(1)14.(3)59.(4)8
V26	Vav'ua	26	(3)2.(1)15.(3)59.(4)8
V34	Vav'ua	10	(0)1.(1)13.(3)34.(4)5.(0)3.(4)16
V36	Vav'ua	26	(3)3.(1)14.(3)59.(4)9
V39	Vav'ua	26	(3)4.(1)13.(3)62.(4)8
V42	Vav'ua	26	(1)11.(3)29.(4)22
PO303	Port Olry	26	(3)1.(1)4.(3)42.(4)22
PO307	Port Olry	26	(3)3.(1)13.(3)41.(4)15
PO314	Port Olry	10	(0)1.(1)13.(3)36.(4)7.(0)2.(4)19
PO315	Port Olry	26	(3)1.(1)4.(3)41.(4)22
PO318	Port Olry	26	(3)1.(1)12.(3)28.(4)20
PO320	Port Olry	1	(1)16.(3)41.(4)17
PO322	Port Olry	26	(3)2.(1)13.(3)40.(4)18
PO324	Port Olry	1	(1)16.(3)41.(4)17
PO326	Port Olry	26	(3)1.(1)4.(3)42.(4)22
PO327	Port Olry	10	(0)1.(1)13.(3)36.(4)7.(0)2.(4)19
PO330	Port Olry	26	(3)2.(1)12.(3)43.(4)13
PO331	Port Olry	26	(1)16.(3)40.(4)6
PO332	Port Olry	10	(0)1.(1)15.(3)30.(4)4.(0)4.(4)19
PO344	Port Olry	26	(3)1.(1)4.(3)41.(4)22
PO346	Port Olry	26	(3)3.(1)13.(3)41.(4)16
PO349	Port Olry	26	(3)1.(1)13.(3)31.(4)26
PO350	Port Olry	26	(1)12.(3)49.(4)14
PO351	Port Olry	26	(1)12.(3)49.(4)14
PO359	Port Olry	26	(3)1.(1)4.(3)42.(4)22
PO360	Port Olry	26	(1)12.(3)3.(1)3.(3)39.(4)11
PO366	Port Olry	26	(3)1.(1)4.(3)42.(4)23
PO370	Port Olry	26	(3)1.(1)13.(3)46.(4)13
PO371	Port Olry	26	(1)6.(3)44.(4)17
PO380	Port Olry	26	(3)1.(1)4.(3)42.(4)22
PO384	Port Olry	26	(3)1.(1)13.(3)46.(4)13
PO388	Port Olry	26	(3)1.(1)13.(3)46.(4)13
PO390	Port Olry	26	(1)10.(3)19.(1)1.(3)8.(4)23
PO392	Port Olry	10	(0)1.(1)15.(3)30.(4)4.(0)4.(4)19
PO393	Port Olry	24	(1)14.(3)32.(4)20
PO397	Port Olry	26	(1)10.(3)3.(1)3.(3)41.(4)8
PO400	Port Olry	10	(0)1.(1)13.(3)36.(4)7.(0)2.(4)19
MF005	Maewo fathers	24	(3)1.(1)12.(3)31.(4)2.(3)2.(4)15
MF019	Maewo fathers	26	(1)14.(3)3.(1)2.(3)34.(4)10
MF022	Maewo fathers	26	(1)11.(3)28.(4)22
MF023	Maewo fathers	26	(3)1.(1)12.(3)29.(4)21
MF025	Maewo fathers	24	(3)1.(1)12.(3)35.(4)19
MF027	Maewo fathers	10	(0)1.(1)13.(3)38.(4)16
MF040	Maewo fathers	26	(1)14.(3)2.(1)3.(3)35.(4)10
MF051	Maewo fathers	26	(3)1.(1)14.(3)47.(4)13
MF054	Maewo fathers	9	(3)6.(1)3.(3)1.(1)5.(3)33.(4)24
MF060	Maewo fathers	26	(1)14.(3)1.(1)6.(3)38.(4)8
MF061	Maewo fathers	26	(3)1.(1)14.(3)39.(4)16
MF066	Maewo fathers	26	(1)10.(3)30.(4)19
MF067	Maewo fathers	26	(1)11.(3)28.(4)22
MF068	Maewo fathers	10	(0)1.(3)2.(1)11.(3)37.(4)23
MF069	Maewo fathers	26	(3)1.(1)12.(3)33.(4)20
MF073	Maewo fathers	26	(3)2.(1)4.(3)43.(4)21
MF082	Maewo fathers	26	(3)3.(1)14.(3)63.(4)6
MF093	Maewo fathers	26	(3)1.(1)12.(3)27.(4)21
MF106	Maewo fathers	26	(1)11.(3)28.(4)22
MF107	Maewo fathers	26	(3)3.(1)14.(3)62.(4)6
MF112	Maewo fathers	26	(3)2.(1)14.(3)62.(4)7
MF113	Maewo fathers	1	(1)15.(3)40.(4)18
MF119	Maewo fathers	10	(0)1.(3)2.(1)11.(3)37.(4)23
MF126	Maewo fathers	26	(3)1.(1)14.(3)42.(4)13
MF145	Maewo Fathers	10	(3)3.(1)9.(3)39.(4)25
MF157	Maewo Fathers	24	(1)11.(3)31.(4)21
MF165	Maewo Fathers	24	(3)1.(1)5.(3)38.(4)23
KK001	Kota Kinabalu	26	(3)1.(1)15.(3)43.(4)9
KK002	Kota Kinabalu	26	(3)3.(1)16.(3)33.(4)18
KK004	Kota Kinabalu	26	(1)17.(3)43.(4)6
KK006	Kota Kinabalu	12	(3)1.(1)18.(3)32.(4)17
KK007	Kota Kinabalu	26	(3)1.(1)18.(3)36.(4)10
KK009	Kota Kinabalu	26	(3)2.(1)14.(3)54.(4)8
KK010	Kota Kinabalu	3	(1)23.(3)48.(4)16
KK012	Kota Kinabalu	26	(3)3.(1)16.(3)38.(4)16
KK014	Kota Kinabalu	26	(3)3.(1)16.(3)48.(4)8
KK020	Kota Kinabalu	26	(1)17.(3)43.(4)6
KK021	Kota Kinabalu	26	(3)1.(1)16.(3)43.(4)10
KK022	Kota Kinabalu	26	(3)1.(1)19.(3)40.(4)8
KK023	Kota Kinabalu	26	(3)1.(1)1.(3)1.(1)16.(3)18.(4)17
KK025	Kota Kinabalu	26	(3)2.(1)17.(3)41.(4)13
KK026	Kota Kinabalu	10	(0)1.(3)1.(1)16.(3)42.(4)22
KK028	Kota Kinabalu	26	(3)3.(1)14.(3)49.(4)5
KK032	Kota Kinabalu	26	(3)1.(1)17.(3)41.(4)9
KK034	Kota Kinabalu	26	(3)3.(1)16.(3)38.(4)16
KK038	Kota Kinabalu	10	(0)1.(3)1.(1)10.(3)37.(4)25
KK040	Kota Kinabalu	10	(0)1.(3)2.(1)11.(3)29.(4)12
KK041	Kota Kinabalu	10	(0)1.(3)1.(1)16.(3)41.(4)22
KK043	Kota Kinabalu	26	(1)15.(3)45.(4)6
KK044	Kota Kinabalu	26	(1)15.(3)45.(4)6
KK046	Kota Kinabalu	26	(1)15.(3)44.(4)4
KK047	Kota Kinabalu	26	(3)2.(1)15.(3)44.(4)5
KK049	Kota Kinabalu	26	(3)3.(1)16.(3)38.(4)16
KK051	Kota Kinabalu	26	(1)14.(3)44.(4)6
KK053	Kota Kinabalu	26	(3)3.(1)15.(3)38.(4)16
KK054	Kota Kinabalu	26	(3)2.(1)17.(3)52.(4)10
KK055	Kota Kinabalu	26	(1)19.(3)39.(4)9
KK058	Kota Kinabalu	26	(3)3.(1)17.(3)37.(4)19

KK059	Kota Kinabalu	1	(1)15.(3)39.(4)21
KK063	Kota Kinabalu	10	(1)16.(3)39.(4)16
KK066	Kota Kinabalu	26	(3)5.(1)14.(3)45.(4)5
KK068	Kota Kinabalu	2	(3)3.(1)13.(3)42.(4)15
KK070	Kota Kinabalu	26	(1)18.(3)40.(4)9
KK075	Kota Kinabalu	26	(3)3.(1)1.(3)1.(1)11.(3)42.(4)10
KK076	Kota Kinabalu	10	(0)1.(3)2.(1)11.(3)30.(4)12
KK077	Kota Kinabalu	26	(3)2.(1)17.(3)53.(4)7
KK079	Kota Kinabalu	3	(1)22.(3)49.(4)17
KK083	Kota Kinabalu	3	(1)22.(3)49.(4)17
KK084	Kota Kinabalu	26	(3)3.(1)14.(3)45.(4)5
KK090	Kota Kinabalu	26	(3)5.(1)14.(3)45.(4)5
KK091	Kota Kinabalu	26	(3)2.(1)12.(3)45.(4)15
KK092	Kota Kinabalu	26	(3)4.(1)14.(3)46.(4)12
KK093	Kota Kinabalu	26	(3)2.(1)2.(3)1.(1)12.(3)43.(4)8
KK094	Kota Kinabalu	26	(1)17.(3)44.(4)9
KK099	Kota Kinabalu	26	(3)3.(1)16.(3)38.(4)16
KK103	Kota Kinabalu	26	(3)3.(1)16.(3)38.(4)16
KK115	Kota Kinabalu	26	(3)1.(1)14.(3)37.(4)15
KK116	Kota Kinabalu	26	(3)3.(1)16.(3)38.(4)15
KK119	Kota Kinabalu	2	5(1)10(3)31(1)11(3)2(1)5(3)7(0)9(3)18(4)16
KK124	Kota Kinabalu	26	(3)3.(1)16.(3)38.(4)16
KK130	Kota Kinabalu	26	(3)2.(1)16.(3)37.(4)16
KK131	Kota Kinabalu	26	(3)1.(1)17.(3)41.(4)8
KK132	Kota Kinabalu	26	(1)14.(3)46.(4)5
KK134	Kota Kinabalu	10	(0)1.(3)1.(1)12.(3)14.(1)1.(3)13.(4)17
KK135	Kota Kinabalu	10	(1)17.(3)39.(4)14
KK136	Kota Kinabalu	26	(3)4.(1)14.(3)58.(4)7
KK144	Kota Kinabalu	26	(3)1.(1)19.(3)40.(4)9
KK145	Kota Kinabalu	26	(3)1.(1)18.(3)38.(4)11
KK147	Kota Kinabalu	10	(3)2.(1)16.(3)41.(4)22
KK150	Kota Kinabalu	26	(3)2.(1)17.(3)37.(4)12
KK151	Kota Kinabalu	26	(1)15.(3)42.(4)10
KK153	Kota Kinabalu	26	(3)3.(1)14.(3)61.(4)8
KK154	Kota Kinabalu	10	(3)2.(1)15.(3)37.(4)30
KK159	Kota Kinabalu	9	(3)6.(1)4.(3)1(1)5.(3)31.(4)22
KK162	Kota Kinabalu	26	(1)15.(3)47.(4)4
KK167	Kota Kinabalu	26	(3)3.(1)16.(3)38.(4)11
KK173	Kota Kinabalu	26	(3)1.(1)19.(3)40.(4)9
KK176	Kota Kinabalu	10	(3)3.(1)11.(3)26.(4)18
KK179	Kota Kinabalu	26	(3)3.(1)16.(3)38.(4)16
BJM001	Banjarmasin	26	(1)19.(3)33.(4)18
BJM005	Banjarmasin	10	(1)16.(3)39.(4)16
BJM010	Banjarmasin	26	(3)2.(1)17.(3)35.(4)17
BJM012	Banjarmasin	26	(3)1.(1)17.(3)44.(4)6
BJM013	Banjarmasin	9	(3)4.(1)6.(3)2.(1)2.(3)2.(1)1.(3)32.(4)24
BJM014	Banjarmasin	26	(3)3.(1)13.(3)34.(1)6.(3)49.(4)6
BJM016	Banjarmasin	26	(1)18.(3)46.(4)4
BJM020	Banjarmasin	10	(1)17.(3)40.(4)15
BJM025	Banjarmasin	26	(1)16.(3)43.(4)4
BJM038	Banjarmasin	26	(3)3.(1)16.(3)35.(4)20
BJM042	Banjarmasin	26	(3)2.(1)16.(3)35.(4)17
BJM047	Banjarmasin	10	(1)17.(3)41.(4)16
BJM053	Banjarmasin	26	(3)2.(1)17.(3)35.(4)17
BJM054	Banjarmasin	26	(3)3.(1)21.(3)31.(4)1.(3)2(4)16
BJM057	Banjarmasin	26	(3)1.(1)16.(3)3(1)2.(3)24.(4)26
BJM058	Banjarmasin	3	(1)20.(3)57.(4)11
BJM060	Banjarmasin	26	(1)18.(3)45.(4)6
BJM063	Banjarmasin	10	(1)16.(3)23.(4)16
BJM067	Banjarmasin	26	(3)1.(1)17.(3)40.(4)10
BJM069	Banjarmasin	10	(1)15.(3)37.(4)16
BJM076	Banjarmasin	26	(3)1.(1)17.(3)32.(4)16
BJM077	Banjarmasin	26	(3)3.(1)17.(3)3(1)1.(3)32.(4)19
BJM078	Banjarmasin	2	(3)4.(1)10.(3)40.(4)16
AMI1	Ami	26	(1)15.(3)47.(4)6
AMI3	Ami	26	(3)4.(1)13.(3)45.(4)7
AMI4	Ami	26	(1)16.(3)37.(4)14
AMI6	Ami	26	(3)3.(1)12.(3)62.(4)6
AMI8	Ami	26	(1)15.(3)48.(4)6
AMI9	Ami	26	(3)1.(1)18.(3)38.(4)9
AMI14	Ami	26	(3)1.(1)16.(3)39.(4)8
AMI15	Ami	26	.(1)12.(3)32.(1)5.(3)1.(1)3.(3)25.(4)2.(3)1.(4)4
AMI17	Ami	26	(3)1.(1)18.(3)38.(4)9
AMI18	Ami	26	(3)1.(1)17.(3)43.(4)9
AMI19	Ami	26	(3)1.(1)16.(3)38.(4)9
AMI20	Ami	26	(3)3.(1)12.(3)63.(4)7
AMI21	Ami	26	(3)2.(1)15.(3)64.(4)8
AMI23	Ami	26	(1)15.(3)47.(4)6
ATA3	Atayal	26	(3)1.(1)17.(3)42.(4)6
ATA5	Atayal	26	(3)1.(1)17.(3)42.(4)6
ATA6	Atayal	26	(3)1.(1)17.(3)43.(4)8
ATA8	Atayal	26	(3)1.(1)17.(3)42.(4)8
ATA9	Atayal	26	(3)1.(1)17.(3)41.(4)6
ATA17	Atayal	26	(3)1.(1)15.(3)43.(4)8
ATA19	Atayal	26	(3)1.(1)17.(3)42.(4)8
BUN1	Bunumi	26	(3)1.(1)18.(3)38.(4)9
BUN3	Bunumi	26	(1)16.(3)42.(4)6
BUN4	Bunumi	26	(1)16.(3)42.(4)6
BUN11	Bunumi	26	(1)16.(3)42.(4)6
BUN12	Bunumi	26	(1)16.(3)42.(4)6
BUN14	Bunumi	26	(3)1.(1)17.(3)38.(4)12
BUN15	Bunumi	26	(3)2.(1)16.(3)41.(4)11
BUN16	Bunumi	26	(1)17.(3)41.(4)6
PAI2	Paiwan	26	(3)1.(1)17.(3)36.(4)3.(0)2.(4)6
PAI4	Paiwan	26	(1)18.(3)41.(4)5
PAI17	Paiwan	26	(3)1.(1)17.(3)1.(1)2.(3)42.(4)6
PAI6	Paiwan	26	(3)1.(1)16.(3)2.(1)2.(3)42.(4)6
PAI19	Paiwan	26	(3)1.(1)18.(3)37.(4)10
PAI20	Paiwan	26	(1)17.(3)43.(4)5
PAI9	Paiwan	26	(1)15.(3)40.(4)6
PAI21	Paiwan	26	(1)17.(3)44.(4)9
PAI12	Paiwan	26	(1)15.(3)44.(4)6
PAI15	Paiwan	26	(1)18.(3)41.(4)5
F7	Filipino	26	(3)3.(1)12.(3)63.(4)5
F16	Filipino	26	(1)16.(3)42.(4)6
F28	Filipino	26	(3)3.(1)13.(3)63.(4)4
F42	Filipino	26	(3)1.(1)17.(3)39.(4)8

F8	Filipino	26	(3)3.(1)12.(3)63.(4)5
F17	Filipino	26	(3)2.(1)13.(3)1.(1)1.(3)25.(4)2.(3)2.(4)15
F30	Filipino	26	(3)3.(1)13.(3)63.(4)4
F43	Filipino	26	(3)1.(1)25.(3)45.(4)8
F56	Filipino	26	(3)3.(1)15.(3)46.(4)13
F10	Filipino	26	(3)1.(1)16.(3)41.(4)7
F20	Filipino	26	(3)3.(1)14.(3)57.(4)8
F31	Filipino	26	(3)3.(1)13.(3)63.(4)4
F47	Filipino	26	(3)1.(1)25.(3)45.(4)8
F22	Filipino	26	(3)3.(1)14.(3)57.(4)8
F33	Filipino	26	(3)3.(1)13.(3)63.(4)4
F48	Filipino	26	(3)4.(1)11.(3)37.(4)13
F60	Filipino	26	(3)3.(1)15.(3)46.(4)13
F23	Filipino	26	(3)3.(1)14.(3)57.(4)8
F34	Filipino	26	(3)1.(1)17.(3)37.(4)9
F61	Filipino	26	(3)3.(1)15.(3)46.(4)13
F2	Filipino	26	(1)18.(3)24.(0)5.(3)11.(4)9
F14	Filipino	1	(1)16.(3)41.(4)17
F38	Filipino	26	(3)1.(1)17.(3)37.(4)9
F50	Filipino	2	(3)3.(1)13.(3)38.(4)15
F3	Filipino	26	(3)1.(1)16.(3)41.(4)7
F26	Filipino	26	(3)1.(3)14.(3)54.(4)4
F40	Filipino	26	(3)1.(1)17.(3)39.(4)8
F4	Filipino	26	(3)3.(1)13.(3)65.(4)7
F41	Filipino	26	(3)1.(1)17.(3)39.(4)8
F53	Filipino	26	(3)1.(1)16.(3)42.(4)7
CI120	Cookl	1	(1)15(3)39(4)21
CI139	Cookl	1	(1)17(3)38(4)18
CI149	Cookl	1	(1)15(3)39(4)18
CI123	Cookl	10	(0)1(1)13(3)33(4)6(0)3(4)16
CI138	Cookl	10	(0)1(1)14(3)32(4)6(0)3(4)16
CI145	Cookl	10	(0)1(1)12(3)38(4)11
CI153	Cookl	10	(0)1(1)13(3)33(4)5(0)4(4)16
CI155	Cookl	10	(0)1(1)12(3)34(4)5(0)3(4)16
CI167	Cookl	10	(0)1(1)13(3)>11(4)5(0)3(4)16
CI186	Cookl	10	(0)1(1)13(3)33(4)6(0)3(4)16
CI151	Cookl	3	(1)20(3)51(4)18
CI128	Cookl	26	(3)2(1)13(3)63(4)6
CI135	Cookl	26	(3)3(1)14(3)59(4)8
CI166	Cookl	26	(3)2(1)13(3)63(4)6
CI75A	Cookl	1	(1)17(3)36(4)21
CI115	Cookl	1	(1)15(3)37(4)20
CI140	Cookl	10	(0)1(1)13(3)33(4)6(0)3(4)16
CI142	Cookl	2	(3)3.(1)13.(3)37.(4)24
CI147	Cookl	10	(0)1(1)13(3)33(4)5(0)4(4)15
CI156	Cookl	10	(0)1(1)12(3)35(4)5(0)3(4)17
CI175	Cookl	10	(0)1(1)13(3)33(4)6(0)3(4)16
CI180	Cookl	10	(0)1(1)13(3)32(4)5(0)4(4)14
CI181	Cookl	10	(0)1(1)13(3)33(4)6(0)3(4)16
CI183	Cookl	1	(1)16(3)38(4)20
CI185	Cookl	10	(0)1.(1)13.(3)33.(4)25
CI188	Cookl	1	(1)17(3)38(4)18
CI190	Cookl	26	(3)1(1)11(0)1(1)1(3)51(4)15
CI191	Cookl	2	(3)3(1)14(3)37(4)23
CI192	Cookl	2	3)3(1)1(1/37)1(1)5(1/37)2(1)4(3)28(4)24
CI194	Cookl	1	(1)15(3)36(4)21
CI196	Cookl	1	(1)16(3)38(4)20
CI198	Cookl	10	(0)1(1)13(3)34(4)5(0)3(4)14
CI206	Cookl	10	(0)1(1)13(3)33(4)5(0)4(4)15
2592	PNG	24	(3)1(1)13(3)36(4)12
2792	PNG	24	(3)1(1)12(3)34(4)15
2892	PNG	24	(3)1(1)13(3)34(4)17
2992	PNG	24	(3)1(1)13(3)40(4)12
3091	PNG	24	(3)1(1)12(3)31(4)16
3092	PNG	24	(3)1(1)14(3)36(4)8
3392	PNG	24	(3)2(1)13(3)33(4)15
3792	PNG	24	(3)2(1)11(3)39(4)15
4292	PNG	24	(3)1(1)13(3)32(4)12
4992	PNG	24	(3)1(1)13(3)35(4)17
5092	PNG	24	(3)1(1)13(3)36(4)19
5492	PNG	24	(3)1(1)12(3)34(4)13
6092	PNG	24	(3)1(1)12(3)37(4)16
8092	PNG	24	(3)1(1)13(3)37(4)12
8392	PNG	24	(3)1(1)13(3)38(4)13
9391	PNG	24	(3)1(1)11(3)36(4)13
O9921	PNG	24	(3)1(1)12(3)34(4)14
12191	PNG	24	(3)2(1)11(3)37(4)15
12591	PNG	24	(3)1(1)12(3)36(4)16
12991	PNG	24	(3)2(1)11(3)35(4)16
13191	PNG	24	(3)1(1)12(3)33(4)13
6192	PNG	24	(3)1(1)13(3)35(4)1(3)1(4)15
8692	PNG	24	(3)1(1)11(0)1(3)32(4)14
14591	PNG	24	(3)1(1)12(3/1)2(3)34(4)16
14991	PNG	24	(1)14(3)37(4)14
6991	PNG	24	(1)12(3)38(4)14
7292	PNG	24	(1)11(3)40(4)14
9192	PNG	24	(1)12(3)38(4)14
1992	PNG	26	(1)12(3)25(4)24
5192	PNG	26	(1)15(3)47(4)12
9791	PNG	26	(1)14(3)48(4)14
13591	PNG	26	(1)14(3)46(4)15
2692	PNG	26	(3)2(3/1)1(1)13(3)61(4)8
9191	PNG	26	(3)3(1)13(3)61(4)8
5692	PNG	26	(3)3(1)12(3)63(4)6
13991	PNG	26	(3)3(1)14(3)60(4)8
4591	PNG	26	(3)1(1)13(3)39(4)14
5791	PNG	26	(3)1(1)12(3)39(4)14
7792	PNG	26	(3)1(1)10(3)37(4)16
8792	PNG	26	(3)1(1)14(3)44(4)13
14391	PNG	26	(3)1(1)13(3)39(4)15
6591	PNG	10	(3)2(1)12(3)29(4)4(0)5(4)24
7092	PNG	10	(0)1(3)1(1)11(3)29(4)7(0)3(4)9(0)3(4)11
14791	PNG	2	(1)14(3/4)21(0)27

All Oceanic Hg26 chromosomes

Sample	Population	Haplogroup	MSY1	MSY1 length	Modular structure	network codes
MJ12	Majoro	26	(1)15.(3)7.(0)1(3)37.(4)4	64	1,3,0,3,4	
F2	Filipino	26	(1)18.(3)24.(0)5.(3)11.(4)9	67	1,3,0,3,4	
MJ23	Majoro	26	(1)15.(3)7.(0)1(3)37.(4)4	64	1,3,0,3,4,	
PO390	Port Olry	26	(1)10.(3)19.(1)1.(3)8.(4)23	61	1,3,1,3,4	
PO397	Port Olry	26	(1)10.(3)3.(1)3.(3)41.(4)8	65	1,3,1,3,4	
PO360	Port Olry	26	(1)12.(3)3.(1)3.(3)39.(4)11	68	1,3,1,3,4	
MF060	Maewo fathers	26	(1)14.(3)1.(1)6.(3)38.(4)8	67	1,3,1,3,4	
MF040	Maewo fathers	26	(1)14.(3)2.(1)3.(3)35.(4)10	64	1,3,1,3,4	
MF019	Maewo fathers	26	(1)14.(3)3.(1)2.(3)34.(4)10	62	1,3,1,3,4	
TL67	Tongan lysates	26	(1)16.(3)3.(1)2.(3)34.(4)10	65	1,3,1,3,4	
MF066	Maewo fathers	26	(1)10.(3)30.(4)19	59	1,3,4	1
MF022	Maewo fathers	26	(1)11.(3)28.(4)22	61	1,3,4	2
MF067	Maewo fathers	26	(1)11.(3)28.(4)22	61	1,3,4	2
MF106	Maewo fathers	26	(1)11.(3)28.(4)22	61	1,3,4	2
V42	Vav'ua	26	(1)11.(3)29.(4)22	62	1,3,4	3
WS97	Western Samoa	26	(1)11.(3)30.(4)22	63	1,3,4	4
2052	Rapa	26	(1)11.(3)30.(4)22	63	1,3,4	4
2085	Rapa	26	(1)11.(3)30.(4)22	63	1,3,4	4
2102	Rapa	26	(1)11.(3)30.(4)22	63	1,3,4	4
2103	Rapa	26	(1)11.(3)30.(4)22	63	1,3,4	4
2104	Rapa	26	(1)11.(3)30.(4)22	63	1,3,4	4
2231	Rapa	26	(1)11.(3)30.(4)22	63	1,3,4	4
KG10	Kapamarangi	26	(1)11.(3)31.(4)20	62	1,3,4	5
KG51	Kapamarangi	26	(1)11.(3)31.(4)20	62	1,3,4	5
KG61	Kapamarangi	26	(1)11.(3)32.(4)19	62	1,3,4	6
1992	PNG	26	(1)12(3)25(4)24	61	1,3,4	7
KG30	Kapamarangi	26	(1)12.(3)23.(4)27	62	1,3,4	8
KG78	Kapamarangi	26	(1)12.(3)24.(4)26	62	1,3,4	9
MJ15	Majoro	26	(1)12.(3)25.(4)24	61	1,3,4	7
KG14	Kapamarangi	26	(1)12.(3)26.(4)24	62	1,3,4	11
KG62	Kapamarangi	26	(1)12.(3)26.(4)24	62	1,3,4	11
KG75	Kapamarangi	26	(1)12.(3)26.(4)24	62	1,3,4	11
PO350	Port Olry	26	(1)12.(3)49.(4)14	75	1,3,4	12
PO351	Port Olry	26	(1)12.(3)49.(4)14	75	1,3,4	12
13591	PNG	26	(1)14(3)46(4)15	75	1,3,4	13
9791	PNG	26	(1)14(3)48(4)14	76	1,3,4	14
MJ1	Majoro	26	(1)14.(3)39.(4)12	65	1,3,4	15
KK051	Kota Kinabalu	26	(1)14.(3)44.(4)6	64	1,3,4	16
KK132	Kota Kinabalu	26	(1)14.(3)46.(4)5	65	1,3,4	17
5192	PNG	26	(1)15(3)47(4)12	74	1,3,4	18
PAI9	Paiwan	26	(1)15.(3)40.(4)6	61	1,3,4	19
KK151	Kota Kinabalu	26	(1)15.(3)42.(4)10	67	1,3,4	20
KK046	Kota Kinabalu	26	(1)15.(3)44.(4)4	63	1,3,4	21
PAI12	Paiwan	26	(1)15.(3)44.(4)6	65	1,3,4	22
KK043	Kota Kinabalu	26	(1)15.(3)45.(4)6	66	1,3,4	23
KK044	Kota Kinabalu	26	(1)15.(3)45.(4)6	66	1,3,4	23
KK162	Kota Kinabalu	26	(1)15.(3)47.(4)4	66	1,3,4	24
AMI1	Ami	26	(1)15.(3)47.(4)6	68	1,3,4	25
AMI23	Ami	26	(1)15.(3)47.(4)6	68	1,3,4	25
AMI8	Ami	26	(1)15.(3)48.(4)6	69	1,3,4	26
AMI4	Ami	26	(1)16.(3)37.(4)14	67	1,3,4	27
PO331	Port Olry	26	(1)16.(3)40.(4)6	62	1,3,4	28
TL63	Tongan lysates	26	(1)16.(3)42.(4)5	63	1,3,4	29
BUN3	Bunumi	26	(1)16.(3)42.(4)6	64	1,3,4	30
BUN4	Bunumi	26	(1)16.(3)42.(4)6	64	1,3,4	30
BUN11	Bunumi	26	(1)16.(3)42.(4)6	64	1,3,4	30
BUN12	Bunumi	26	(1)16.(3)42.(4)6	64	1,3,4	30
F16	Filipino	26	(1)16.(3)42.(4)6	64	1,3,4	30
BJM025	Banjarmasin	26	(1)16.(3)43.(4)4	63	1,3,4	31
WS13	Western Samoa	26	(1)17.(3)39.(4)9	65	1,3,4	32
BUN16	Bunumi	26	(1)17.(3)41.(4)6	64	1,3,4	33
PAI20	Paiwan	26	(1)17.(3)43.(4)5	65	1,3,4	34
KK004	Kota Kinabalu	26	(1)17.(3)43.(4)6	66	1,3,4	35
KK020	Kota Kinabalu	26	(1)17.(3)43.(4)6	66	1,3,4	35
KK094	Kota Kinabalu	26	(1)17.(3)44.(4)9	70	1,3,4	36
PAI21	Paiwan	26	(1)17.(3)44.(4)9	70	1,3,4	36
KK070	Kota Kinabalu	26	(1)18.(3)40.(4)9	67	1,3,4	37
PAI4	Paiwan	26	(1)18.(3)41.(4)5	64	1,3,4	38
PAI15	Paiwan	26	(1)18.(3)41.(4)5	64	1,3,4	38
BJM060	Banjarmasin	26	(1)18.(3)45.(4)6	69	1,3,4	39
BJM016	Banjarmasin	26	(1)18.(3)46.(4)4	68	1,3,4	40
BJM001	Banjarmasin	26	(1)19.(3)33.(4)18	70	1,3,4	41
KK055	Kota Kinabalu	26	(1)19.(3)39.(4)9	67	1,3,4	42
PO371	Port Olry	26	(1)6.(3)44.(4)17	67	1,3,4	43
CI190	Cook	26	(3)1(1)11(0)1(1)1(3)51(4)15		3,1,0,1,3,4	
AMI15	Ami	26	(3)3.(1)12.(3)32.(1)5.(3)1.(1)3.(3)25.(4)2.(3)1.(4)4	88	3,1,3,1,3,1,3,4,3,4	
2240	Rapa	26	(3)1(1)16(3)1(1)3(3)40(4)9	70	3,1,3,1,3,4	
KK023	Kota Kinabalu	26	(3)1.(1)1.(3)1.(1)16.(3)18.(4)17	54	3,1,3,1,3,4	
PAI6	Paiwan	26	(3)1.(1)16.(3)2.(1)2.(3)42.(4)6	69	3,1,3,1,3,4	



BJM057	Banjarmasin	26	(3)1.(1)16.(3)3(1)2.(3)24.(4)26	72	3,1,3,1,3,4	
PAI17	Paiwan	26	(3)1.(1)17.(3)1.(1)2.(3)42.(4)6	69	3,1,3,1,3,4	
KK075	Kota Kinabalu	26	(3)3.(1)1.(3)1.(1)11.(3)42.(4)10	68	3,1,3,1,3,4	
BJM014	Banjarmasin	26	(3)3.(1)13.(3)34.(1)6.(3)49.(4)6	111	3,1,3,1,3,4	
BJM077	Banjarmasin	26	(3)3.(1)17.(3)3(1)1.(3)32.(4)19	75	3,1,3,1,3,4	
KK093	Kota Kinabalu	26	(3)2.(1)2.(3)1.(1)12.(3)43.(4)8	68	3,1,3,1,3,4,	
F17	Filipino	26	(3)2.(1)13.(3)1.(1)1.(3)25.(4)2.(3)2.(4)15	63	3,1,3,1,3,4,3,4	
7792	PNG	26	(3)1(1)10(3)37(4)16	64	3,1,3,4	1
5791	PNG	26	(3)1(1)12(3)39(4)14	66	3,1,3,4	2
4591	PNG	26	(3)1(1)13(3)39(4)14	67	3,1,3,4	3
14391	PNG	26	(3)1(1)13(3)39(4)15	68	3,1,3,4	4
8792	PNG	26	(3)1(1)14(3)44(4)13	72	3,1,3,4	5
MF093	Maewo fathers	26	(3)1.(1)12.(3)27.(4)21	61	3,1,3,4	6
PO318	Port Olry	26	(3)1.(1)12.(3)28.(4)20	61	3,1,3,4	7
MF023	Maewo fathers	26	(3)1.(1)12.(3)29.(4)21	63	3,1,3,4	8
MF069	Maewo fathers	26	(3)1.(1)12.(3)33.(4)20	66	3,1,3,4	9
PO349	Port Olry	26	(3)1.(1)13.(3)31.(4)26	71	3,1,3,4	10
PO370	Port Olry	26	(3)1.(1)13.(3)46.(4)13	73	3,1,3,4	11
PO384	Port Olry	26	(3)1.(1)13.(3)46.(4)13	73	3,1,3,4	11
PO388	Port Olry	26	(3)1.(1)13.(3)46.(4)13	73	3,1,3,4	11
KK115	Kota Kinabalu	26	(3)1.(1)14.(3)37.(4)15	67	3,1,3,4	12
MF061	Maewo fathers	26	(3)1.(1)14.(3)39.(4)16	70	3,1,3,4	13
MF126	Maewo fathers	26	(3)1.(1)14.(3)42.(4)13	70	3,1,3,4	14
MF051	Maewo fathers	26	(3)1.(1)14.(3)47.(4)13	65	3,1,3,4	15
TL115	Tongan lysates	26	(3)1.(1)14.(3)63.(4)6	84	3,1,3,4	16
ATA17	Atayal	26	(3)1.(1)15.(3)43.(4)8	67	3,1,3,4	17
KK001	Kota Kinabalu	26	(3)1.(1)15.(3)43.(4)9	68	3,1,3,4	18
AMI19	Ami	26	(3)1.(1)16.(3)38.(4)9	64	3,1,3,4	19
AMI14	Ami	26	(3)1.(1)16.(3)39.(4)8	64	3,1,3,4	20
F10	Filipino	26	(3)1.(1)16.(3)41.(4)7	66	3,1,3,4	21
F3	Filipino	26	(3)1.(1)16.(3)41.(4)7	65	3,1,3,4	21
F53	Filipino	26	(3)1.(1)16.(3)42.(4)7	66	3,1,3,4	23
KK021	Kota Kinabalu	26	(3)1.(1)16.(3)43.(4)10	70	3,1,3,4	24
BJM076	Banjarmasin	26	(3)1.(1)17.(3)32.(4)16	66	3,1,3,4	25
F34	Filipino	26	(3)1.(1)17.(3)37.(4)9	64	3,1,3,4	26
F38	Filipino	26	(3)1.(1)17.(3)37.(4)9	64	3,1,3,4	26
BUN14	Bunumi	26	(3)1.(1)17.(3)38.(4)12	68	3,1,3,4	27
F42	Filipino	26	(3)1.(1)17.(3)39.(4)8	65	3,1,3,4	28
F40	Filipino	26	(3)1.(1)17.(3)39.(4)8	65	3,1,3,4	28
F41	Filipino	26	(3)1.(1)17.(3)39.(4)8	65	3,1,3,4	28
BJM067	Banjarmasin	26	(3)1.(1)17.(3)40.(4)10	68	3,1,3,4	30
ATA9	Atayal	26	(3)1.(1)17.(3)41.(4)6	65	3,1,3,4	31
KK131	Kota Kinabalu	26	(3)1.(1)17.(3)41.(4)8	67	3,1,3,4	32
KK032	Kota Kinabalu	26	(3)1.(1)17.(3)41.(4)9	68	3,1,3,4	33
ATA3	Atayal	26	(3)1.(1)17.(3)42.(4)6	67	3,1,3,4	34
ATA5	Atayal	26	(3)1.(1)17.(3)42.(4)6	67	3,1,3,4	34
ATA8	Atayal	26	(3)1.(1)17.(3)42.(4)8	68	3,1,3,4	35
ATA19	Atayal	26	(3)1.(1)17.(3)42.(4)8	68	3,1,3,4	35
ATA6	Atayal	26	(3)1.(1)17.(3)43.(4)8	69	3,1,3,4	36
AMI18	Ami	26	(3)1.(1)17.(3)43.(4)9	70	3,1,3,4	37
BJM012	Banjarmasin	26	(3)1.(1)17.(3)44.(4)6	68	3,1,3,4	38
KK007	Kota Kinabalu	26	(3)1.(1)18.(3)36.(4)10	65	3,1,3,4	39
PAI19	Paiwan	26	(3)1.(1)18.(3)37.(4)10	66	3,1,3,4	40
KK145	Kota Kinabalu	26	(3)1.(1)18.(3)38.(4)11	68	3,1,3,4	41
AMI9	Ami	26	(3)1.(1)18.(3)38.(4)9	66	3,1,3,4	42
AMI17	Ami	26	(3)1.(1)18.(3)38.(4)9	66	3,1,3,4	42
BUN1	Bunumi	26	(3)1.(1)18.(3)38.(4)9	66	3,1,3,4	42
KK022	Kota Kinabalu	26	(3)1.(1)19.(3)40.(4)8	68	3,1,3,4	43
KK144	Kota Kinabalu	26	(3)1.(1)19.(3)40.(4)9	69	3,1,3,4	44
KK173	Kota Kinabalu	26	(3)1.(1)19.(3)40.(4)9	69	3,1,3,4	44
F43	Filipino	26	(3)1.(1)25.(3)45.(4)8	79	3,1,3,4	45
F47	Filipino	26	(3)1.(1)25.(3)45.(4)8	79	3,1,3,4	45
PO315	Port Olry	26	(3)1.(1)4.(3)41.(4)22	68	3,1,3,4	46
PO344	Port Olry	26	(3)1.(1)4.(3)41.(4)22	68	3,1,3,4	46
PO303	Port Olry	26	(3)1.(1)4.(3)42.(4)22	69	3,1,3,4	47
PO326	Port Olry	26	(3)1.(1)4.(3)42.(4)22	69	3,1,3,4	47
PO359	Port Olry	26	(3)1.(1)4.(3)42.(4)22	69	3,1,3,4	47
PO380	Port Olry	26	(3)1.(1)4.(3)42.(4)22	69	3,1,3,4	47
PO366	Port Olry	26	(3)1.(1)4.(3)42.(4)23	70	3,1,3,4	48
F26	Filipino	26	(3)1.(3)14.(3)54.(4)4	73	3,1,3,4	49
CI128	Cookl	26	(3)2(1)13(3)63(4)6		3,1,3,4	50
CI166	Cookl	26	(3)2(1)13(3)63(4)6		3,1,3,4	50
2692	PNG	26	(3)2(3/1)1(1)13(3)61(4)8	85	3,1,3,4	51
PO330	Port Olry	26	(3)2.(1)12.(3)43.(4)13	70	3,1,3,4	52
KK091	Kota Kinabalu	26	(3)2.(1)12.(3)45.(4)15	74	3,1,3,4	53
T18	Tongan	26	(3)2.(1)12.(3)62.(4)6	82	3,1,3,4	54
T20	Tongan	26	(3)2.(1)12.(3)62.(4)6	82	3,1,3,4	54
PO322	Port Olry	26	(3)2.(1)13.(3)40.(4)18	73	3,1,3,4	55
MJ7	Maloro	26	(3)2.(1)13.(3)56.(4)13	84	3,1,3,4	56
WS21	Western Samoa	26	(3)2.(1)13.(3)63.(4)6	84	3,1,3,4	50
WS29	Western Samoa	26	(3)2.(1)13.(3)64.(4)6	85	3,1,3,4	57
WS36	Western Samoa	26	(3)2.(1)13.(3)64.(4)6	85	3,1,3,4	57
WS40	Western Samoa	26	(3)2.(1)13.(3)64.(4)6	85	3,1,3,4	57
WS31	Western Samoa	26	(3)2.(1)13.(3)66.(4)6	87	3,1,3,4	58
KK009	Kota Kinabalu	26	(3)2.(1)14.(3)54.(4)8	78	3,1,3,4	59
KG31	Kapamarangi	26	(3)2.(1)14.(3)57.(4)14	87	3,1,3,4	60
TL64	Tongan lysates	26	(3)2.(1)14.(3)58.(4)8	82	3,1,3,4	61
MF112	Maewo fathers	26	(3)2.(1)14.(3)62.(4)7	85	3,1,3,4	62
KK047	Kota Kinabalu	26	(3)2.(1)15.(3)44.(4)5	66	3,1,3,4	63

lineages

26.1
26.2
26.3
26.4
26.5
26.6
26.7

V26	Vav'ua	26	(3)2.(1)15.(3)59.(4)8	84	3,1,3,4	64
AMI21	Ami	26	(3)2.(1)15.(3)64.(4)8	89	3,1,3,4	65
KG1	Kapamarangi	26	(3)2.(1)16.(3)29.(4)16	63	3,1,3,4	66
BJM042	Banjarmasin	26	(3)2.(1)16.(3)35.(4)17	70	3,1,3,4	67
KK130	Kota Kinabalu	26	(3)2.(1)16.(3)37.(4)16	71	3,1,3,4	68
BUN15	Bunumi	26	(3)2.(1)16.(3)41.(4)11	70	3,1,3,4	69
WS47	Western Samoa	26	(3)2.(1)17.(3)28.(4)17	64	3,1,3,4	70
BJM010	Banjarmasin	26	(3)2.(1)17.(3)35.(4)17	71	3,1,3,4	71
BJM053	Banjarmasin	26	(3)2.(1)17.(3)35.(4)17	71	3,1,3,4	71
KK150	Kota Kinabalu	26	(3)2.(1)17.(3)37.(4)12	68	3,1,3,4	72
KK025	Kota Kinabalu	26	(3)2.(1)17.(3)41.(4)13	73	3,1,3,4	73
KK054	Kota Kinabalu	26	(3)2.(1)17.(3)52.(4)10	81	3,1,3,4	74
KK077	Kota Kinabalu	26	(3)2.(1)17.(3)53.(4)7	79	3,1,3,4	75
MF073	Maewo fathers	26	(3)2.(1)4.(3)43.(4)21	70	3,1,3,4	76
MJ30	Majoro	26	(3)3(1)12(3)57(4)14	86	3,1,3,4	77
5692	PNG	26	(3)3(1)12(3)63(4)6	84	3,1,3,4	78
9191	PNG	26	(3)3(1)13(3)61(4)8	85	3,1,3,4	79
CI135	Cookl	26	(3)3(1)14(3)59(4)8		3,1,3,4	80
13991	PNG	26	(3)3(1)14(3)60(4)8	85	3,1,3,4	81
KG41	Kapamarangi	26	(3)3.(1)12.(3)51.(4)13	79	3,1,3,4	82
KG56	Kapamarangi	26	(3)3.(1)12.(3)51.(4)13	79	3,1,3,4	82
KG73	Kapamarangi	26	(3)3.(1)12.(3)52.(4)12	79	3,1,3,4	83
KG39	Kapamarangi	26	(3)3.(1)12.(3)55.(4)13	83	3,1,3,4	84
MJ83	Majoro	26	(3)3.(1)12.(3)55.(4)14	84	3,1,3,4	85
MJ24	Majoro	26	(3)3.(1)12.(3)56.(4)14	85	3,1,3,4	86
MJ13	Majoro	26	(3)3.(1)12.(3)57.(4)14	86	3,1,3,4	77
KG47	Kapamarangi	26	(3)3.(1)12.(3)58.(4)12	85	3,1,3,4	87
MJ36	Majoro	26	(3)3.(1)12.(3)58.(4)12	85	3,1,3,4	87
AMI6	Ami	26	(3)3.(1)12.(3)62.(4)6	83	3,1,3,4	88
F7	Filipino	26	(3)3.(1)12.(3)63.(4)5	83	3,1,3,4	89
F8	Filipino	26	(3)3.(1)12.(3)63.(4)5	83	3,1,3,4	89
AMI20	Ami	26	(3)3.(1)12.(3)63.(4)7	85	3,1,3,4	90
PO307	Port Olry	26	(3)3.(1)13.(3)41.(4)15	72	3,1,3,4	91
PO346	Port Olry	26	(3)3.(1)13.(3)41.(4)16	73	3,1,3,4	92
F28	Filipino	26	(3)3.(1)13.(3)63.(4)4	83	3,1,3,4	93
F30	Filipino	26	(3)3.(1)13.(3)63.(4)4	83	3,1,3,4	93
F31	Filipino	26	(3)3.(1)13.(3)63.(4)4	83	3,1,3,4	93
F33	Filipino	26	(3)3.(1)13.(3)63.(4)4	83	3,1,3,4	93
F4	Filipino	26	(3)3.(1)13.(3)65.(4)7	88	3,1,3,4	94
KK084	Kota Kinabalu	26	(3)3.(1)14.(3)45.(4)5	67	3,1,3,4	95
KK028	Kota Kinabalu	26	(3)3.(1)14.(3)49.(4)5	71	3,1,3,4	96
F20	Filipino	26	(3)3.(1)14.(3)57.(4)8	82	3,1,3,4	97
F22	Filipino	26	(3)3.(1)14.(3)57.(4)8	82	3,1,3,4	97
F23	Filipino	26	(3)3.(1)14.(3)57.(4)8	82	3,1,3,4	97
2143	Rapa	26	(3)3.(1)14.(3)58.(4)8	83	3,1,3,4	98
TL8	Tongan lysates	26	(3)3.(1)14.(3)58.(4)8	73	3,1,3,4	98
WS57	Western Samoa	26	(3)3.(1)14.(3)59.(4)3	79	3,1,3,4	99
WS72	Western Samoa	26	(3)3.(1)14.(3)59.(4)3	79	3,1,3,4	99
T4	Tongan	26	(3)3.(1)14.(3)59.(4)8	84	3,1,3,4	80
TL20	Tongan lysates	26	(3)3.(1)14.(3)59.(4)8	84	3,1,3,4	80
TL27	Tongan lysates	26	(3)3.(1)14.(3)59.(4)8	84	3,1,3,4	80
TL31	Tongan lysates	26	(3)3.(1)14.(3)59.(4)8	84	3,1,3,4	80
TL32	Tongan lysates	26	(3)3.(1)14.(3)59.(4)8	84	3,1,3,4	80
TL83	Tongan lysates	26	(3)3.(1)14.(3)59.(4)8	84	3,1,3,4	80
TL126	Tongan lysates	26	(3)3.(1)14.(3)59.(4)8	84	3,1,3,4	80
V15	Vav'ua	26	(3)3.(1)14.(3)59.(4)8	84	3,1,3,4	80
V36	Vav'ua	26	(3)3.(1)14.(3)59.(4)9	85	3,1,3,4	100
TL99	Tongan lysates	26	(3)3.(1)14.(3)60.(4)8	85	3,1,3,4	81
KK153	Kota Kinabalu	26	(3)3.(1)14.(3)61.(4)8	86	3,1,3,4	101
MF107	Maewo fathers	26	(3)3.(1)14.(3)62.(4)6	85	3,1,3,4	102
MF082	Maewo fathers	26	(3)3.(1)14.(3)63.(4)6	86	3,1,3,4	103
KK053	Kota Kinabalu	26	(3)3.(1)15.(3)38.(4)16	72	3,1,3,4	104
F56	Filipino	26	(3)3.(1)15.(3)46.(4)13	77	3,1,3,4	105
F60	Filipino	26	(3)3.(1)15.(3)46.(4)13	77	3,1,3,4	105
F61	Filipino	26	(3)3.(1)15.(3)46.(4)13	77	3,1,3,4	105
KK002	Kota Kinabalu	26	(3)3.(1)16.(3)33.(4)18	70	3,1,3,4	107
BJM038	Banjarmasin	26	(3)3.(1)16.(3)35.(4)20	74	3,1,3,4	108
KK167	Kota Kinabalu	26	(3)3.(1)16.(3)38.(4)11	68	3,1,3,4	109
KK116	Kota Kinabalu	26	(3)3.(1)16.(3)38.(4)15	72	3,1,3,4	110
KK012	Kota Kinabalu	26	(3)3.(1)16.(3)38.(4)16	73	3,1,3,4	111
KK034	Kota Kinabalu	26	(3)3.(1)16.(3)38.(4)16	73	3,1,3,4	111
KK049	Kota Kinabalu	26	(3)3.(1)16.(3)38.(4)16	73	3,1,3,4	111
KK099	Kota Kinabalu	26	(3)3.(1)16.(3)38.(4)16	73	3,1,3,4	111
KK103	Kota Kinabalu	26	(3)3.(1)16.(3)38.(4)16	73	3,1,3,4	111
KK124	Kota Kinabalu	26	(3)3.(1)16.(3)38.(4)16	73	3,1,3,4	111
KK179	Kota Kinabalu	26	(3)3.(1)16.(3)38.(4)16	73	3,1,3,4	111
KK014	Kota Kinabalu	26	(3)3.(1)16.(3)48.(4)8	75	3,1,3,4	112
KK058	Kota Kinabalu	26	(3)3.(1)17.(3)37.(4)19	76	3,1,3,4	113
F48	Filipino	26	(3)4.(1)11.(3)37.(4)13	65	3,1,3,4	114
WS1	Western Samoa	26	(3)4.(1)11.(3)63.(4)7	85	3,1,3,4	115
AMI3	Ami	26	(3)4.(1)13.(3)45.(4)7	69	3,1,3,4	116
V39	Vav'ua	26	(3)4.(1)13.(3)62.(4)8	87	3,1,3,4	117
KK092	Kota Kinabalu	26	(3)4.(1)14.(3)46.(4)12	76	3,1,3,4	118
KK136	Kota Kinabalu	26	(3)4.(1)14.(3)58.(4)7	83	3,1,3,4	22
KK066	Kota Kinabalu	26	(3)5.(1)14.(3)45.(4)5	69	3,1,3,4	106
KK090	Kota Kinabalu	26	(3)5.(1)14.(3)45.(4)5	69	3,1,3,4	106
PAI2	Paiwan	26	(3)1.(1)17.(3)36.(4)3.(0)2.(4)6	65	3,1,3,4,0,4	
BJM054	Banjarmasin	26	(3)3.(1)21.(3)31.(4)1.(3)2(4)16	74	3,1,3,4,3,4	

lineages

- 26.1
- 26.2
- 26.3
- 26.4
- 26.5
- 26.6
- 26.7