Investigation of Autosomal Tetranucleotide STR loci and Male lineages among U.K. Leicestershire and Polynesian populations

By

Emma Watson MSc

Supervisor: Dr Mastana

Director of Research: Professor Cameron

Thesis submitted for the examination of the award: PhD

Loughborough University, Department Human Sciences

Abstract

This study reports the findings of an investigation of ten polymorphic autosomal tetranucleotide short tandem repeat loci and Y chromosome haplotype diversity among the genetically diverse populations of; U.K. Leicestershire caucasians, New Zealand Maori and other Polynesian Islanders.

The ten autosomal loci were initially isolated and sequenced by the Utah Marker Development group. This present study optimised the methodology for use with unlabelled primers and submarine gel electrophoresis technology. Little or no previous population or forensic genetic research had been carried out incorporating the ten loci presented in this study.

The autosomal loci investigated in the present study were highly polymorphic which when analysed together had the potential to offer a greater power of exclusion than the system currently employed by the Forensic Science Service in the U.K. The Polynesian populations of the New Zealand Maori and other Islanders provided an interesting and stark contrast both historically and genetically to the northern European, U.K. Leicestershire population. Polynesia is purported to be the 'last chapter' of mans colonisation of the World. In particular New Zealand, which was the very last to be occupied by seafaring Polynesian peoples, 1000 – 2000 years before present.

The Y chromosome study elaborated upon specifically male haplotypes and provided the first data of male lineages among the New Zealand Maori population. In agreement with the documented Polynesian prehistory, close genetic affinities were observed between the Maori and Polynesian Islanders with extensive European admixture. The European admixture was also observed among other Polynesian Islanders of this and other studies. The European admixture has been described as post-settlement contact during the last 400 years, with the advent of European voyagers 'discovering' the Pacific and its' inhabitants.

A common haplotype isolated among the New Zealand Maori of this study, was also present among the Polynesian Islanders and has previously been thought of as unique to Polynesia. The common haplotype, was dated within the New Zealand Maori to have a 'common ancestry' of around 1000 years before present, which is in good agreement with archaeological and anthropological estimations of the time/age of settlement of New Zealand.

Acknowledgement

It is my sincerest wish to thank all individuals who have guided me through the four years of research. In particular I am indebted to Dr Chambers at Wellington University who provided the Polynesian DNA samples and whose insight and enthusiasm kept me motivated through the harder times.

I also thank Dr Thomas at University College London who provided all the resources I needed (including his time and efforts) to carry out the Y chromosome research.

I am also appreciative of the kind and generous support of Dr Mastana and Professor Cameron who painstakingly reviewed the progress of this research.

Finally, and by no means least, a special thanks should be given to the laboratory staff, Alice and Pat, at Loughborough University. These unsung heroes provided moral support, countless laughter, tea, cake and general good humour.

Thank you to one and all.

Contents

Chapter 4

The prehistory of man in the Pacific Islands

Chapter 5

Section I: The Pilot Methodology

Section II: Amplification Methodology

Chapter 9

Discussion: How informative were the 10 autosomal tetranucleotide STR loci systems?

Chapter 10

Y Chromosome Results

Chapter 11

Discussion: Male lineages among the U.K. Leicestershire caucasian, New Zealand Maori and Polynesian Islander

Chapter 12

Comparing and contrasting issues:

Appendix

Chapter 1

The Short Tandem Repeat

1.a. Introduction

Deoxyribonucleic acid (DNA) is located within the nucleus of all living and replicating cells. The DNA molecule is complex, comprised of singular units or 'nucleotides' joined to form a long chain. An estimated 6 X 10^9 nucleotides are present within a diploid or non-sex cell (Craig et al. 1988). The nucleotide in turn contains a 'base' of one of the following; adenine, cytosine, thymine or guanine. Two long chains of nucleotides coil round each other, forming the characteristic double helix.

The DNA sequence can be crudely divided into 'coding' and 'non-coding' regions. The coding regions are the specific areas of DNA termed 'genes' that code for the production of a specific protein. They also determine the degree of expression of the gene in any tissue at any time (Craig et al. 1988). The coding regions therefore rarely alter their base composition, as such 'mutations' may prevent the successful production of a vital protein necessary for the healthy functioning of the organism.

The non-coding regions are, as the name suggests, DNA sequences that do not code for any particular proteins. These sequences can be found as a single copy between coding regions or as multiple copies termed 'repetitive' DNA (Craig et al. 1988, Hancock 1999). Specific sites of repetitive DNA have been shown to bind proteins and enhance gene expression. This 'function' is possibly affected by the number of tandem repeats of the microsatellite (Kashi and Soller 1999). The non-coding DNA has no sequence constraints in comparison to the coding DNA; thus the repetitive sequences can spread and diversify throughout the genome. Furthermore, their invaluable source of interindividual genetic variation has proved an asset in Forensic identification (Craig et al. 1988, Evett et al. 1996, Edwards et al. 1992).

1.b. Repetitive DNA

An estimated 20-40% of mammalian genomes contain moderately or highly repeated sequences (Craig et al. 1988). The identification of a repeated sequence was established following one of the three observations:

i) Single stranded tandemly repeated sequences of DNA re-anneal at a faster rate than the coding or single copy sequences.

ii) Centrifugation in the presence of caesium chloride separates highly repetitive DNA from the remaining DNA due to the differences in 'buoyancy'.

iii) Enzymatic digestion of the genome at specific restriction sites produces bands of different lengths, when separated through an ethidium bromide stained agarose gel. This indicates the dispersion of similar sequences throughout the genome (Craig et al. 1988).

1.c Repetitive DNA and their polymorphisms

The tandem repeats dependent on their size can be classified as minisatellites or microsatellites. The minisatellites may be present at hundreds or thousands of loci per genome, in which a repeat unit sequence (motif) of a sufficient length to be locus specific is repeated to an overall size of 0.5-30kb. The microsatellites or short tandem repeats (STR) are equally abundant, with repeat motifs 2-8 base pairs in length tandemly repeated giving an overall length in the region 20-100 base pairs (Armour et al. 1999, Gill and Evett 1995). The term microsatellite was historically used for dinucleotide repeats (Hancock 1999) although is now commonly used to describe repeat motifs of 2-8 base pairs. The term short tandem repeat abbreviated to STR is also commonly used (Edwards et al. 1992) and is the preferred terminology to describe the repeat sequences in this thesis.

The repetitive DNA is subdivided to tandem repeats or interspersed elements (Craig et al. 1988). The interspersed elements, for example the Alu insertion, are commonly scattered throughout the genome as single units of specific length (Craig et al. 1988). Similarly, the tandem repeats are also dispersed throughout the genome, although are multiple tandemly repeated units (Rostedt et al. 1996, Craig et al. 1988).

Tandem repeat polymorphisms are the product of insertion/deletion events, which produces a lengthening or shortening of the overall fragment length. The process of such events may be through:

i) integral numbers of unit slippage at replication or

ii) unequal recombination between the tandemly repeated sequences, or both (Craig et al. 1988).

1.d. Mutational Mechanisms – length alteration

The most common short tandem repeats in all genomes are the poly(A)/poly(T) repeats (Hancock 1999). However, their instability during the polymerase chain reaction makes this repeat undesirable for use in mapping, population or forensic analyses (Hancock 1999).

In the year 1989, the dinucleotide STR was first demonstrated to have a length polymorphism. Since that time thousands of dinucleotide STR polymorphisms have been characterised throughout the human genome (Deka et al. 1995). It was purported that CA/TG dinucleotide repeats were most common, occurring twice as frequently as AT repeats and three times as often as AG/TC repeats (Hancock 1999). Deka et al. (1995) observed the apparent lack of population genetic analyses using the dinucleotide STR's. This could possibly be due to the PCR 'stuttering' during *in vitro* amplification causing spurious artefacts to be visualised after electrophoresis of the PCR product (Wall et al. 1993). Stuttering is caused by DNA slippage in the presence of a polymerase enzyme during amplification (Perez-Lezaun et al. 1997a). Often the amplified fragments are one or more repeats shorter than the true length (Goldstein and Schlotterer 1999).

Trinucleotide repeats have associations with a variety of genetic disorders. The trinucleotide repeat diseases are caused by the expansion of a 'normal' trinucleotide STR within a gene (Rubinsztein 1999). In the year 1991, the first trinucleotide STR locus was isolated in association with the fragile X syndrome, known to be the most common form of familial mental retardation (Kashi and Soller 1999).

The longer the repeat unit, less slippage is observed and greater accuracy analysing the PCR product is achieved (Wall et al. 1993). This has been observed at tetranucleotide STR loci. These loci are abundant throughout the genome, and because of their four base pair repeat units unambiguous allele size designation is possible (Perez-Lezaun et al. 1997a).

1.e. Tetranucleotide Short Tandem Repeats - Inheritance and mutations

The transmission of genetic material in the form of ova and sperm at fertilisation fuse the male and female haploid gametes forming diploid genetic material. In general, the diploid DNA will remain unaltered, although replicated a vast number of times throughout the 'life' of the developing individual. Therefore, the STR loci contain inherited genetic information from both parents. It is this fact that is exploited in both population and forensic genetics. Population genetics compares and contrasts genetic structures within and between populations using allele frequencies and calculated mutation rates (Hancock 1999). Forensic genetics considers the population genetic dynamics to assess the rarity of specific genetic information and also investigates specific polymorphic loci which may be of use in aiding the identification of individuals (Hammond et al. 1994).

At a specific locus, there are varying lengths of repeat motifs or 'alleles'. Individuals have two alleles per locus, in accordance with the inheritance pattern as described above. Where both alleles are of the same length, the term 'homozygous' is used. Similarly, if the alleles are of different sizes, the term 'heterozygous' is used. Within a population a number of alleles are observed, the greater the number of different alleles the greater the 'diversity'. Varying numbers of different alleles and size differences between populations exist, creating 'population trends' noticeable through 'allelic frequencies' (Watson et al. 1998; Lum et al. 1998; Jorde et al. 1997).

An allele length mutation at a specific locus can occur resulting in either an increase or decrease in repeat number. It is generally accepted that the mutational mechanism for microsatellites is that of 'replication slippage' (Levinson and Gutman 1987; Hancock 1999). Replication slippage involves the dissociation of the replicating strand from the template, and the incorrect reannealing of the strands to continue replication. If a tandemly repeated sequence misanneals, the replicate will be longer or shorter than the template. If a loop occurs in the template the replicate will be shorter, or a loop in the replicate strand will increase the replicate length (see figure 1.1).

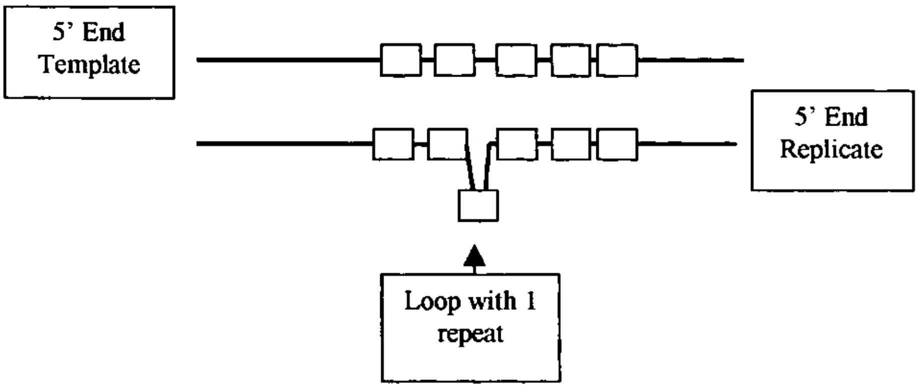


FIGURE 1.1: REPLICATION IN THE DIRECTION 5’ TO 3’.

THE REPLICATE RE-ANNEALED TO THE TEMPLATE FORMING A LOOP CONTAINING A SINGLE REPEAT. THUS, THE REPLICATE INCREASED THE ALLELE LENGTH BY ONE REPEAT ‘MOTIF’.

Recombination has also been suggested as another possible mutational event altering the allele length (Hancock 1999). There have been two proposed mechanisms of recombination; crossing-over between mis-aligned chromosome strands resulting in an insertion/deletion and ‘gene conversion’ allowing unidirectional transfer of DNA from one strand to the other (Hancock 1999).

Brinkmann et al. (1998) observed the influence of the structure and length of tetranucleotide and pentanucleotide short tandem repeats on the mutation rate in humans. In general, the mutation rate increased with increasing repeat length and there was a strong indication that replication slippage was the cause of the mutations and not recombination (Brinkmann et al. 1998). Unimodal and bimodal distributions of allele frequencies versus allele lengths have also been investigated. The bimodal distributions of alleles were described to be an indication that two subgroups of alleles were present with differences in the sequence and arrangement of repeats (Brinkmann et al. 1998; Freimer and Slatkin 1996). This was believed to be an effect of genetic drift and founder effects (Brinkmann et al. 1998). Interrupted alleles (i.e., alleles with interspersed irregular motifs) were longer and less susceptible to mutation as the chances of replication slippage were reduced (Brinkmann et al. 1998). Mutation rate differences also exist between the male and female germ line, in the ratio of 17:3 males to females as reported by Brinkmann et al. (1998), and a similar ratio of 15:4 males to females as reported by Weber and Wong (1993). This was purported to be associated with the number of cell divisions before mitosis, whereby females undergo approximately 22 divisions before meiosis in comparison to males who continuously go through mitosis with some cells to meiosis. Therefore, the increased number of divisions in males increases the susceptibility to mutation (Brinkmann et al. 1998).

Weber and Wong (1993) concluded that the tetranucleotide short tandem repeats mutated at a higher rate than dinucleotide repeats. However, Chakraborty et al. (1997) disagreed with Weber and Wong's (1993) findings, and observed the exact opposite, whereby dinucleotides mutated at a faster rate than the tetranucleotides. However, the disease causing trinucleotides mutated 2-3 times faster than the dinucleotides. The mutation rates of these short tandem repeats have been observed in the region of 10^{-4} – 10^{-5} per gamete (Bruford and Wayne 1993). In addition to microsatellite mutation rates, the direction in which the mutation occurs (i.e., increasing or decreasing allele lengths) has also been investigated (Rubinsztein et al. 1995; Goldstein and Pollock 1997; Estoup and Cornuet 1999).

Researchers have concluded that mutations at microsatellite loci favour, on average, a gain in allele-length (Amos 1999; Rubinsztein et al. 1995; Goldstein and Pollock 1997; Amos et al. 1996; Primmer et al. 1996). Evidence supporting this finding, stems from mutation research across numerous diverse human loci (Amos et al. 1996) and the incorporation of human germlines (Weber and Wong 1993). In addition to studies examining allele length alteration through mutation, sequence variation also alters the susceptibility of a locus to mutation, hence its mutation rate (Brinkmann et al. 1998). Jin et al. (1996) observed alleles with low mutation rates had shorter uninterrupted repeat sequences, than alleles with longer repeat arrays. Thus when examining the allelic distribution an asymmetry was observed with the outlying alleles tending to be shorter than the main group, consistent with a bias favouring expansion. In contrast, the more complex an allelic sequence, the less susceptible it is to mutation, as long uninterrupted tandem arrays have a greater potential to mutate (Brinkmann et al. 1998).

Another theory proposed by Amos et al. (1996) purports that heterozygous genotypes mutate more frequently than the equivalent homozygous genotypes, brought about by mismatch repair between the homologous chromosomes during meiosis termed 'heterozygote instability'. This was further corroborated by

Rubinsztein et al. (1995), who further added the possibility of a population size effect, whereby the allele lengths were influenced by interchromosomal events. This theory explains if mutations were more likely in heterozygous individuals, then large populations (i.e., humans) support greater genetic diversity, therefore high levels of heterozygosity, thus a higher mutation rate than smaller populations.

1.f. Classification systems – allele sizing, sequence variation and allelic ladders

Ideally all STR analyses should be carried out following a standard procedure to ensure allele sizing/motif repeat number continuity between independent laboratories, so that direct comparisons and contrasts can be made between data sets. This is of-course, at best highly problematic, and at worst close to impossible, mainly because of the differing analytical methodologies that can be used. Goldstein and Schlotterer (1999) believe a 'repeat count' (i.e., the number of repeat motifs of a specific allele) is the best policy and where this is unknown, the allele size in base pairs can be used. Repeat counts offer an easier system to compare result data from independent laboratories, as an exact number of repeat motifs are recorded. In comparison allele sizes in base pairs not only includes the repeat motifs but also the DNA sequences flanking these regions. Furthermore, PCR artefacts such as adenylation (the addition of adenosine to the end of a PCR product during the thermocycle programe) cause allele sizes to differ by one or two repeats. Thus, this choice of data recording makes it less amenable for cross comparison with other independent laboratories. Equally as difficult is the use of an arbitrary scale, whereby neither allele sizes nor repeat counts are used. In these instances, the results cannot be compared to data other than that generated using the same arbitrary scale, with the same methodology at the same locus. An example of this was the comprehensive work of Jorde et al. (1997). Sixty microsatellite systems were analysed using genetically diverse world-wide populations. A battery of statistical analyses were carried-out, but all incorporating the arbitrary scale, hence difficult for comparison to other studies.

In order to faithfully record the repeat count, one may sequence the amplified product. Sequencing not only provides data on the number of repeat motifs, but also the structure of the microsatellite locus (i.e., 'perfect' (or simple), 'imperfect', 'interrupted' or 'compound') (Goldstein and Schlotterer 1999).

Lareu et al. (1998), listed the sequence composition of each allele within five tetranucleotide STR systems. A simple or perfect (Goldstein and Schlotterer 1999) STR was observed, composed of GATA motifs tandemly repeated with no intermediate alleles. Compound repeats were also observed consisting of variable numbers of AGAT and AGAC with intermediate GAT motifs (Lareu et al. 1998). The imperfect microsatellite was described to be where one or more repeat motifs include a base pair that is not congruent with the repeat structure, e.g., AGAT AGAT AGCT AGAT. Lastly, the interrupted microsatellite is characterised by the insertion of a number of base pairs that are not consistent with the repeat structure, e.g., AGATAGATCCCAGAT (Goldstein and Schlotterer 1999).

The sequencing of alleles is clearly advantageous not only to characterise the microsatellite but also to identify the regions within the microsatellite that tend to mutate most frequently. This point is best clarified by reiterating an observation made by Brinkmann et al. (1998), whereby interrupted repeats were less susceptible to 'slippage' during replication. Hence, this reduced the chances of mutation in comparison to simple repeats with no anomalous insertion of additional nucleotides.

Knowledge of allele size and repeat number is invaluable when constructing allelic ladders for a specific microsatellite locus. Allelic ladders form a set of different sized alleles identical in sequence and structure to a specific microsatellite. Amplified samples of unknown size can be measured against the corresponding allelic ladder. The allelic ladders are used with both silver staining and fluorescent scanning technologies. However, the fluorescent scanners also use an 'internal size standard' whereby fragments of known length are electrophoresed as a mixture with the amplified DNA sample. Although the internal ladder does not have exactly the same sized fragments as the unknown alleles, by means of computer software the unknown allele sizes are 'estimated'. The term 'estimated' has been used as differences in migration occur because of the differences in electrophoretic ability of the sequences of the alleles and the fragments of the internal size standard (Promega Technical Manual 1997)[1].

[1] N.B. For a more detailed explanation on migration rates, refer to Chapter 6 of this study.

1.g. Genetic Diversity of the Short Tandem Repeat – The influence of population structure

Perez-Lezaun et al. (1997a), isolated 178 different alleles across 20 tetranucleotide STR loci, among major continental populations. On average, 8-9 different alleles per locus were observed, giving an insight to the vast diversity present within the human population. Even although substantial differences in allele frequencies existed between major populations, in general populations within a major racial group exhibited more similarities than differences (Deka et al. 1995).

The gene diversity (heterozygosity) is 'a measure of the proportion of individuals whose homologous chromosomes carry distinguishable alleles' (Goldstein and Schlotterer 1999). Genetic diversity is calculated from the allele frequencies of the population being studied and is one of the important statistical measures used to understand basic population structures.

Similarly to Perez-Lezaun et al.'s (1997a) study, varied gene diversities across World populations have been observed, calculated from data incorporating 60 tetranucleotide microsatellite polymorphisms (Jorde et al. 1997). Greatest diversity was recorded in continental Africa (76%) and least in Asian populations (60%).

Although most comparisons of genetic variation between human populations express excess African diversity using up to 30 microsatellite markers (Deka et al. 1995), it was not until Jorde et al. (1997) included 60 microsatellite markers that a significant difference in variation was observed. Even then, it was purported that a further one hundred microsatellite systems would be required to confidently resolve inter-population relationships (Jorde et al. 1997).

1.h. Factors affecting genetic diversity

Genetic diversity has been described as having two components; additive genetic variance, and allelic diversity. Additive genetic variance is 'the proportion of genetic differences among individuals/populations' and allelic diversity described 'the number of different alleles present at any locus' (Holsinger 1999). A number of different factors have been purported to affect genetic diversity including genetic drift (Bowcock et al. 1991) and selective sweeps (Holsinger 1999).

1.h.i. Genetic Drift

When variation is considered selectively neutral, the only factor altering the status quo is drift (Bowcock et al. 1991). Drift is expected to be equal for all DNA as drift depends only on the demographic properties of populations (Bowcock et al. 1991).

Two types of genetic drift occur, the founder effect and the genetic bottleneck;

The founder effect occurs when allele frequencies in a group of migratory individuals are not representative of their population of origin. If these individuals were then isolated, after a few generations the progeny may develop different characteristics to the original population (http://library.thinkquest.org/19926/java/tour/09.html).

The other type of genetic drift is the bottleneck. This occurs when a population suffers a dramatic decrease in size, possibly caused by any number of climatic, environmental or ecologial conditions. The populations may increase in number with time, but the allelic frequencies may have been considerably altered. Genetic differences between human populations today have been purported to be the consequence of bottlenecks in ancestral populations (http://library.thinkquest.org/19926/java/tour/09.html). Holsinger (1999) purported that population bottlenecks had little effect on variance but a dramatic effect on diversity.

1.h.ii. Selective Sweep

At some point in time a mutation occurs which is advantageous to the human more so than the 'ancestral' allele. Natural selection perpetuates the spread of the advantageous mutation throughout the population until it becomes fixed (i.e., present in the whole population). During the time required to 'fix' the mutation polymorphisms in the region flanking the mutation may be lost. Hence the 'selective sweep' reduces the variation and hence diversity of the polymorphic flanking region (Schlotterer and Wiehe 1999). However, if the flanking region is in linkage to the advantageous mutation, then not only is the mutation fixed but also the linked flanking region. Therefore, immediately after a selective sweep, variation will be reduced in the flanking region linked to the mutated site. As new mutations in the flanking region occur variation increases until the mutation-drift equilibrium is restored and during this time an excess of rare variants are observed (Schlotterer and Weihe 1999).

1.i. The tetranucleotide markers used in this study

A total of 17 tetranucleotide short tandem repeat markers were initially chosen for optimisation using the Perkin Elmer 480 thermocycler and submarine gel electrophoresis. Elimination of loci that did not produce consistently clear results reduced the number of working systems to ten. The ten useable short tandem repeat loci were; D1S407, D2S262, D3S1514, D4S2285, D5S592, D7S618, D7S1485, D9S252, D10S520 and D12S297. These loci were originally all isolated by the Utah Marker Development group and their relationship to other markers with reference to odds of inversion were statistically evaluated (Utah Marker Development group 1995).

The ten loci have been discussed further;

D1S407

This locus is located on the short arm of chromosome 1. It has an AGAT repeat motif which has been reported to produce distinct allelic bands (Ballard personal communication). This marker has been described as an 'anchor' marker, with 1 in 1000 chance of inversion with adjacent loci (Utah Marker Development group 1995). The maximum heterozygosity has been reported by the Genome Database as 75% of a caucasian population.

This locus has recently received attention because of its possible linkage to a rare prostate cancer (Gibbs et al. 1999). An explanation for the linkage, was that within the D1S407 region a gene exists which functions as a type of tumour-suppressor. Therefore, mutations within this region may cause a susceptibility to cancer (Gibbs et al. 1999).

D2S262

This locus has an AAAG motif, which is repeated simply. Similarly to D1S407, D2S262 has been described as an anchor marker (Ballard personal communication). The genome database has reported a maximum heterozygosity (gene diversity) of 80% in a caucasian population. There has been no reported linkage to any disease loci.

D3S1514

This locus has an AAAG motif with an AGAG repeat motif insertion (Ballard personal communication) making this locus a compound repeat (see Lareu et al. 1998). This marker has also been described as an anchor marker as there is less than a 1 in 1000 chance of inversion with adjacent loci (Utah Marker Development Group 1995).

A maximum heterozygosity of 83% has been reported by the Genome Database.

D4S2285

This locus has a complex compound repeat structure with a predominant AAGG repeat motif (Ballard personal communication). A maximum heterozygosity was observed at 75% in a caucasian population, as reported by the Genome Database. The odds of an inversion event occurring at this locus have been estimated at 1 in 1000, hence D4S2285 has been termed an 'anchor' marker (Utah Marker Development Group 1995).

There have been no reported observations of linkage to genes or disease loci, at this locus.

D5S592

This locus is located on chromosome five and is characterised by an AGAT repeat motif (Ballard personal communication). This repeat also has an imperfect structure, with an AAAT repeat possibly formed from an insertion deletion event of G in the AGAT motif (Ballard personal communication). The maximum heterozygosity was 88% as observed by the Genome Database.

A thorough literature search of this locus found no information assuming linkage to genes or diseased loci.

D7S1485 and D7S618

These loci are both located on chromosome seven and are situated 49 centimorgans apart and exhibit no linkage to each other (Ballard personal communication).

D7S1485 has been characterised as having primarily an AAGG and AAAG motif repeats, making this a compound locus (see Lareu et al. 1996). The genome database has observed a maximum heterozygosity value of 88% at this locus. The tetranucleotide marker D7S1485, was reported as an anchor marker with a low chance of inversion with adjacent loci (Utah Marker Development group 1995). D7S618 has an observed maximum heterozygosity of 86% (Genome Database). Unlike the aforementioned loci, D7S618 has not been distinguish as an anchor marker and has an estimated chance of inversion of one in one hundred (Utah Marker Development Group 1995). Both loci have not been linked to genes or diseased loci.

D9S252

This marker is located on the long arm of chromosome 9 (Xin et al. 1999) and is characterised by a simple AGAT repeat motif (Ballard personal comunication). The genome database estimated a maximum heterozygosity of 88% at this locus. This locus has exhibited linkage to a tumor suppressor gene causing a form of skin cancer (Xin et al. 1999). In 40% of skin cancer patients, a loss of heterozygosity was observed at this locus (Xin et al. 1999).

D10S520

This marker was located on chromosome 10 and was characterised by a compound AAAG and CAAA repeat structure (Ballard personal communication). The genome database estimated the maximum heterozygosity to be 86%. The chance of inversion with adjacent loci was reported to be very low and so was termed an ‘anchor marker’ (Utah Marker Development Group 1995).

Literature searches have not provided any evidence of this markers’ linkage to either genes or disease loci.

D12S297

This marker was located on chromosome 12 (Utah Marker Development Group 1995) and was characterised by an imperfect AGAT repeat (Ballard personal communication, Goldstein and Schlotterer 1999). The Genome database has estimated a maximum heterozygosity of 80% in a caucasian database. The odds of inversion with adjacent loci were estimated to be one in one thousand, therefore this too was described as an anchor marker (Utah Marker Development Group 1995). Presently, there is no published evidence of this markers' linkage to either genes or disease loci.

1.j. Previous studies

The STR loci D12S297, D7S618 and D1S407 investigated in the present thesis, have also been investigated previously on separate occasions (Jorde et al. 1995, Jorde et al. 1997, Lum et al. 1998). Other studies have found possible linkage between the STR loci; D1S407 (Gibbs et al. 1999) and D9S252 (Xin et al. 1999) to rare forms of cancer.

Jorde et al. (1997) examined differences in diversity at 60 microsatellite loci among human population samples. Included were the microsatellite loci; D7S618 and D12S297. An earlier study by Jorde et al. (1995) incorporated 30 microsatellite markers and various mitochondrial DNA sequences. Within Jorde et al.'s (1995) study, the D1S407 tetranucleotide microsatellite marker was characterised using genetically diverse human populations. Information concerning the sizing method was not included in Jorde et al.'s (1995) paper, although an arbitrary allele sizing of 1-9 was used at the D1S407 locus. There was no mention of the use of allelic ladders and an internal size standard would not have been included (e.g. as used with fluorescent technology) as the electrophoresed bands were visualised by autoradiography (Jorde et al. 1995). Loci, D7S618 and D12S297 were analysed using the same methodology (Jorde et al. 1997). The heterozygosity values at these loci were 93.7% (D1S407), 89.4% (D7S618) and 84.2% (D12S297) in a northern European sample population (Jorde personal communication). Similar heterozygosities were observed in a Finnish population with values of 93.5% (D1S407), 83.9% (D7S618) and 81.5% (D12S297) (Jorde personal communication). In general, the D12S297 tetranucleotide microsatellite held the lowest heterozygosity value and D1S407 the highest (Jorde personal communication).

In the year 1998, Lum et al. investigated the genetic relationships among Pacific Island and Asian populations. Autosomal tetranucleotide short tandem repeat were used including D1S407 and D12S297 similarly to previous studies (Jorde et al. 1995, Jorde et al. 1997). In general, Lum et al. (1998) observed a loss of STR genetic diversity in geographically isolated Remote Oceanic populations, and higher diversities were observed in regions geographically intermediate between the Remote Pacific and Near Oceania (e.g. Samoa, Yap and Vanatu). The heterozygosity values at the D1S407 locus were; 68.5%, 57.9%, 66.8% and

67.8% in Chinese, Papua New Guinea, Samoa and Yap populations respectively. Similarly, at the D12S297 locus, the heterozygosities were 82.8%, 77.7%, 79.4% and 82.7% in Chinese, Papua New Guinea, Samoa and Yap populations respectively (Lum personal communication). Interestingly, Jorde et al. (1997) observed locus D1S407 to have on average a higher heterozygosity than D12S297, in comparison, Lum et al. (1998) observed the exact opposite.

1.k. Autosomal marker studies among Pacific and U.K. Leicestershire populations

The Pacific

In 1999, Hagelberg et al. investigated the genetic structure of the Pacific using mitochondrial, Y-chromosome and autosomal analyses. Nucleotide sequence polymorphisms at four loci within the HLA system expressed reduced genetic diversity in Island Melanesia. The Samoans, Taiwanese and Roro were closest to each other and most genetically distant to the Papua New Guinea Highlanders. Hagelberg et al. (1999), supported the idea of close genetic affinities between the Austronesians in Taiwan to the Samoan population.

Minisatellite analyses of New Zealanders revealed that the closest relatives to the Maori were the Polynesians (Clark et al. 1995). Interestingly, there was no evidence to indicate that a bottleneck event occurring when the Maoris colonised New Zealand, had reduced genetic variation (Clark et al. 1995). However, the Maoris themselves were purported to be a racially admixed population, hence complicating the analysis of their origin (Clark et al. 1995).

The U.K.

In the year 1995, Gill and Evett compared frequency distributions between 16 different databases, including seven different Caucasian databases, using four tetranucleotide short tandem repeat loci. The Caucasian populations included two general UK populations and a regional population (Derbyshire). The variance of allele distribution (incorporating four loci), was greatest in the general UK population and least in the Derbyshire region (Gill and Evett 1995). The genetic diversities exhibited only marginal differences between populations, hence the numbers of different alleles were also similar (Gill and Evett 1995).

Concentrating on just one locus, Drozd et al. (1994) observed heterozygosity values at the vWA locus to be similar to those found in other Northern European Caucasian populations. However, a significant departure from Hardy-Weinberg equilibrium was observed at the 5% level. The departure was explained as a mis-typing of the results.

Other departures from Hardy-Weinberg equilibrium have been observed at microsatellites among U.K. caucasian populations (Kimpton et al. 1993). Kimpton et al. (1993) observed that one of 14 different microsatellites expressed a departure from Hardy-Weinberg equilibrium. This was purported to be the result of sampling error.

Chapter 2

The Y Chromosome:

Haplotypes and their use in Forensic and Population Genetics

2.a. Introduction

The Y chromosome is the second smallest of the human chromosomes, containing approximately 60 million base pairs of DNA (Hammer and Zegura 1996). The Y chromosome or male sex chromosome is inherited paternally, and passed from father to son at fertilisation with the fusion of the sperm and ova. The probability of a woman giving birth to a male or female is statistically the same. Therefore, some families may have one or more offspring of the same sex. In families where only females are born, the male lineage is lost. Conversely, families where only males are born the male lineage remains.

One of the oldest recorded male lineages is that of Adam's descendants to Noah in the Old testament, Genesis Chapter 5 and records: When Adam had lived one hundred and thirty years, he became the father of a son in his likeness, according to his image, and named him Seth. Seth begat Enos, Enos begat Cainan, Cainan begat Mahalaleel, Mahalaleel begat Jared, Jared begat Enoch, Enoch begat Methuselah, Methuselah begat Lamech and Lamech begat Noah – he of the Arc and great flood.

2.b. Mutational processes

The genetic information contained within the Y chromosome passes from generation to generation usually as an exact copy (Jobling and Tyler-Smith 1995). Mutational events, although rare, alter the 'content' of the genetic information. The types of mutations that can occur, include;

♦ *Indels*

These are insertions or deletions of nucleotide bases or sequences into or out of specific regions of the DNA.

♦ *Single nucleotide polymorphisms (SNPs)*

Single nucleotide polymorphisms (SNPs) or diallelic polymorphisms occur when a specific nucleotide is replaced for another (Jobling and Tyler-Smith 1995).

Indels and SNPs occur infrequently and may have only occurred once during human evolution and therefore have been termed 'unique event polymorphisms' (UEPs) (Thomas et al. 1998).

- Microsatellites

These are 2-5 base pair DNA sequences (motifs) repeated tandemly a variable number of times. Microsatellite mutations alter the number of motifs either by replication slippage increasing the number of repeats or a deletion decreasing the number.

- Minisatellites

Similarly to the microsatellite, minisatellites also have a varying number of tandem motifs, however these are normally 10-60 base pairs long and the number of repeats can be several dozen.

The rarity of the aforementioned Y chromosomal mutations in comparison to mitochondrial or autosomal mutations have been attributed in part to the physical number of copies of the genetic material. The recombination and mutation of diploid autosomal DNA can occur quite freely and at varying frequency between the paired chromosomes (Jobling et al. 1997). Mitochondrial genes are polyploid and tend to mutate at a high rate (Bianchi et al. 1998). Y-specific genes are haploid and so recombination of genetic material is somewhat restricted to the tips of the Y chromosome and the X chromosome known as the *pseudoautosomal* region (Jobling et al. 1997). The remaining non-recombining portion of the Y chromosome has all its genes in linkage disequilibrium, hence commonly transmitted as an unaltered unit generation to generation (Bianchi et al. 1998).

2.c. Y chromosome diversity

The reduction in mutation frequency observed within the Y-chromosome also coincides with a lack of polymorphic diversity which Jobling and Tyler-Smith (1995) explain as;

- A reduced Y chromosome population compared to the X chromosome as there is a 1:3 ratio of Y:X
- The effective population size may be reduced if a small percentage of men produce a large number of offspring thus Y chromosome diversity is reduced.
- Finally, due to a lack of recombination with other chromosomes it is possible that if an advantageous mutation arose which spread throughout a population

accompanied by a 'hitchhiker' low in variation, a reduction in diversity would occur.

The difference in diversity between Y chromosomal, mitochondrial and autosomal DNA is highlighted within African populations. A low level of Y chromosome genetic diversity and a contrasting high level of mtDNA and autosomal diversity was observed (Jobling and Tyler-Smith 1995). Diversity not only changes with different marker systems but can also differ between populations analysed with the same system. For example, Y chromosome diversity is generally lower than autosomal or mtDNA by virtue of mutational frequencies (Jobling and Tyler-Smith 1995), but the diversity between populations is also affected by population stresses including genetic bottlenecks, gene flow and selection.

An extensive Y chromosomal study of male lineage's in Polynesia found a difference in opinion to mtDNA studies with not only diversity issues, but also the movements of peoples throughout the Pacific (Hurles et al. 1998).

Mitochondrial studies of New Zealand Maoris have recently been used to test the migration patterns and to estimate the founding population size (Murray-McIntosh et al. 1998). It was postulated that between 50-100 women founded New Zealand from eastern Polynesia, which was consistent with Maori oral history, forming more of a planned settlement, rather than a chance finding (Murray-McIntosh et al. 1998). Mitochondrial diversities of Maori, East Polynesian, West Polynesian and Melanesian populations revealed a decreasing diversity from Melanesia to New Zealand was observed, consistent with bottleneck and founding population theories (Murray-McIntosh et al. 1998). Although Murray-McIntosh et al. (1998) did not explicity study European ancestry in the female lineages in Polynesia, there was no indication of indeed any European female lineages in Polynesia. In fact, the same 'native' mitochondrial sequence was observed in 87% of the Maori, 64% of eastern Polynesian, in 56% of the western Polynesian and 23% of the Melanesian mtDNA samples.

The observation that almost all maternal lineages were derived from native Polynesian ancestors and at least one third of the Y-chromosome lineages' were of recent European origin was further qualified by Hurles et al. (1998).

2.d. Towards a unified classification system

In the year 1996, the first Y-user workshop was held. At this international gathering it was agreed that as many population samples as possible should be typed at seven of the 'established' Y STRs: DYS19, DYS389 I+II, DYS390, DYS391, DYS392, DYS393 (de Knijff et al. 1997). The Y chromosome is useful when comparing closely related populations that otherwise would not be distinguished using autosomal STRs or other classical marker systems. However, one should also use more stable Y polymorphisms, for example base substitutions, when examining an evolutionary context as Y STRs have higher mutation rates than diallelic polymorphisms (de Knijff et al. 1997).

Ruiz-Linares et al. (1996) examined the geographic clustering of human Y-chromosome haplotypes using five polymorphic markers (including DYS19) in 13 populations. A total of 78 haplotypes were distinguished and the relationship between them assessed using a neighbour joining tree. These workers found there to be a particularly close relationship between Asian and Oceanic populations and in accordance with the 'out of Africa' theory the greatest haplotype difference was between African and non-African populations.

In the same year, Santos et al. (1996) studied the allele frequency distribution of the DYS19 locus in several genetically diverse populations. Predominant alleles were observed in an Amerindian, Asian and African populations. Unfortunately, due to the nomenclature system employed by Santos et al. (1996) comparisons between this study and others could not be made.

2.e. Background information on the Y-Chromosomal loci used in this study.

Previously, Y-chromosome studies of populations in the Pacific have examined prehistoric migrations (Hurles et al. 1998, Hagelberg et al. 1999). However, no previous work has been published specifically examining male lineages in New Zealand Maoris. The markers used to carry out this work and methods of analysis have been developed during the past five years, beginning with the typing of the YAP marker (Hammer 1995).

A total of ten polymorphic short tandem repeat loci were multiplexed following the protocol of Thomas et al. (2000). Four tetranucleotide polymorphic markers (DYS19, DYS390, DYS391 and DYS393) and the two trinucleotide polymorphic repeat markers (DYS388 and DYS392) were used as part of a multiplex system.

A total of ten diallelic (or UEP) markers and the YAP (Hammer 1995) insertion were also used. The biallelic markers were; 92R7 (Mathias et al. 1994), sY81 (Seielstad et al. 1994), SRY465 (Shinka and Nakahori not published), TAT (Zerjal et al. 1997), M9, M13, M17, M20 (Underhill et al. 1997), SRY+4064 and SRY +10,831 (Whitfield et al. 1995).

The single nucleotide polymorphisms, M17, M20 (Underhill et al. 1997) and SRY465 (Shinka and Nakahori unpublished) have not been extensively used for male lineage studies (Thomas et al. 1998). Hence, this study will provide an interesting insight as to the genetic affinities of these loci in the Maori and U.K. Leicestershire populations.

This male lineage study of New Zealand Maoris complements the novel autosomal study carried out on both male and female Maori DNA samples at ten tetranucleotide short tandem repeat loci.

The U.K. Leicestershire DNA samples provided an interesting population comparison, especially in the light of evidence indicating a strong European ancestry in Polynesia (Hurles et al. 1998).

2.f. Background information on Y chromosomal STR loci used in this study

DYS19 (or DYS394)

This is an extensively studied tetranucleotide GATA repeat located on the short arm of the Y chromosome (Jobling and Tyler-Smith 1995; Roewer et al. 1996). The repeat sequence $(GATA)_3GGTA(GATA)_{12}$, has an allele size range of 186-194bp with 194bp corresponding to 12 repeats (Carvalho-Silva et al. 1999). Heterozygosity level at this locus was observed to be 66% in a Caucasian population (Genome DataBase).

At this locus the most frequent allele varied across regions. A unified classification system has not been implemented, thus comparisons between populations can be problematic. However, allele 13 (Carvalho-Silva et al.'s 1999 nomenclature) was most frequent among the Inuit (77.5%), whereas allele 16 was most frequent among the Samoans (80%). Allele 14 was regarded as a predominantly a caucasian allele observed in 49.9% Europeans and Indians (35.4%), and in agreement with the findings of the Genome DataBase.

Alternatively, Ruiz Linares et al. (1996) classified alleles by base pairs. These workers observed the 192bp allele to be most frequent in a European population (56%), the 196bp allele most frequent in a Chinese population (69%) and the Melanesian population (100%). Kittles et al. (1998) also classified alleles by base pairs. A total of five alleles were recorded (243, 247, 251, 255 and 259 base pairs) with the 247bp allele as most frequent in a Finnish population.

Hammer and Horai (1995) observed a total of six different alleles at this locus. Arbitrarily naming these alleles A to F ranging in size from 186bp to 202 bp. This corresponded to 10-14 copies of the GATA motif. The B allele was most frequent in Western European populations (50%) and regarded as characteristic of western European populations. In contrast, the C allele was most frequent in Japanese populations (47%) with the B allele least frequent (<5%).

De Knijff et al. (1997) observed a total of five different alleles across eight genetically diverse populations. Following the recommendations of the International Society of Forensic Haemogenetics (ISFH), de Knijff et al. (1997) designated alleles by the number of repeat motifs. At this locus five different alleles; 13, 14, 15, 16 and 17 were observed. Above 40% of the sample population in New Guinea/Australia, North East Asia, India and Africa had the 15 allele (de Knijff et al. 1997). Allele 14 was observed to be most frequent, in the European population (>40%) (de Knijff et al. 1997; Forster et al. 1998) and allele 13 in the Arctic (80%) and American (<40%) populations (de Knijff et al. 1997).

DYS388

This is a trinucleotide short tandem repeat marker with an ATA motif (Genome DataBase). Heterozygosity values were observed to be 60%, 51%, 38% and 55% in Caucasian, German, Japanese and Chinese populations respectively (Genome Data Base).

The Finnish populations' most frequent allele was 129bp in length, with a range of allele sizes between 126-144 base pairs (Kittles et al. 1998). The 129bp allele was also most frequent in German, Japanese, and Chinese populations, however the Samoan's most frequent allele (50%) was the 141bp length allele (Deka et al. 1995).

Similarly, the most frequent allele in both Basque and Catalan populations was 129bp in length occurring at frequencies 87% and 86% respectively (Ruiz-Linares et al. 1999).

Ruiz-Linares et al. (1999) and de Knijff et al. (1997) followed Kayser et al.'s (1997) PCR protocol. Kittles et al. (1998) followed procedures as listed in the Genome Data Base and Deka et al. (1995) followed primer listings and locus information as given by Jobling and Tyler-Smith (1995). In turn Jobling and Tyler-Smith gained locus information from the Genome Data Base. It is reasonable to assume therefore that a 129bp allele observed by Kittles et al. (1998) was equivalent to a 129bp allele observed by Deka et al. (1995). Similarly, an equivalence of allele sizing existed between Ruiz-Linares et al. (1999) and de Knijff et al. (1997).

Interestingly, Thomas et al. (1998) found this marker to deviate from the stepwise mutation model, due to a bimodal distribution of allele frequencies and a suspected deletion event resulting in a clear division of longer:shorter allele lengths. This observation was pertinent to statistical models of common ancestry which assume a stepwise mutational process, thus the DYS388 locus may need to be excluded from such calculations.

DYS390

This is a tetranucleotide short tandem repeat locus with a GATA motif. The repeat sequence is $(GATA)_4(GACA)(GATA)_{10}(GACA)_8(GATA)_2$, with alleles in the size range 212-224bp and 212bp corresponding to 10 repeats (Carvalho-Silva et al. 1999). Heterozygosity values of 71%, 76%, 74% and 46% were recorded in Caucasian, German, Japanese and Chinese populations respectively (Genome Data Base).

Deka et al. (1995), observed a total of ten different alleles at this locus with alleles varying in length from 187bp to 227bp. The 215bp allele was most frequent in German (32%), Japanese (32%) and Samoan (35%) populations. However, the 211bp allele was most frequent in the Chinese population (Deka et al. 1995). A total of four different alleles were observed in a Finnish population (206, 210, 214 and 218 base pairs). The 214bp length allele was most frequent in the Finnish sample population (Kittles et al. 1998).

De Knijff et al. (1997) observed seven different alleles across eight genetically diverse populations. Using nomenclature as recommended by the ISFH the seven alleles were 7, 8, 9, 10, 11, 12 and 13. Allele 11 was most frequent in the New Guinea/Australian (>40%), Arctic (80%), American (<30%) and European (<40%) populations. Allele 12 was most frequent in the Indian (<40%) and North

East Asian (<40%) populations. Finally, allele 10 was most frequent in a South East Asian population and allele 8 in an African population (de Knijff et al. 1997).

Ruiz-Linares et al. (1999) observed three different alleles in a Basque population and four alleles in a Catalan population. The alleles were classified according to the amplified PCR product length and not the number of repeat motifs. The most frequent allele was 215bp long and was observed at a frequency of 77% in the Basque population and 69% in the Catalan population (Ruiz-Linares et al. 1999).

Forster et al. (1998), observed particularly short alleles (18-20 repeats) in aboriginal Australians, Papuans and Island southeast Asians in comparison to European alleles typically 3-4 repeat motifs longer. Sequencing of the short repeats expressed a deletion of four repeats within the $(CTGT)_8$ block. This deletion was observed in primarily the Australian samples. Similarly, the Papuan alleles 19-22 repeats in length had the CTGT and the CTAT motif (of varying repeat sizes) deleted, thus simplifying the structure of the locus (Forster et al. 1998).

DYS391

This is a tetranucleotide short tandem repeat marker with a GATA motif. The repeat sequence is $(GATA)_{10}(GACA)_3(GATA)$ with alleles in the size range 279-287bp and allele of length 283bp corresponds to 10 repeats (Carvalho-Silva et al. 1999).

Heterozygosity values were observed to be 49%, 29%, 41% and 33% in Caucasian, German, Japanese and Chinese populations respectively (Genome Data Base).

Deka et al. (1995) observed a total of six different alleles at this tetranucleotide repeat locus, with allele sizes in the range 275-295bp. The 283bp allele was most frequent in the German (82%), Chinese (80%), Japanese (75%) and Samoan (85%) populations. A total of three different alleles were observed with lengths 283bp, 287bp and 291bp in a Finnish population. The most frequent allele in this Finnish sample population was 287bp in length (Kittles et al. 1998).

De Knijff et al. (1997) observed a total of five different alleles across eight genetically diverse populations. The most frequent allele in every population was allele 10. All five alleles (8, 9, 10, 11 and 12) were only observed in an African sample population, with allele 10 at a frequency greater than 40%. Indian, New Guinea/Australian South East Asian and North East Asian populations had similar allele frequencies. These populations were observed to have three alleles (8, 9, and 10) with allele 8 below a frequency of 20% (de Knijff et al. 1997). The European and American sample populations were similar to the Indian, New Guinea/Australian and South East Asian populations but also present at a low frequency (<20%) was allele 12. Finally, the Arctic sample population expressed two alleles; 10 and 11 with respective frequencies 80% and 20%.

At this tetranucleotide repeat locus, Ruiz-Linares et al. (1999) observed four different alleles (279bp, 283bp, 287bp and 291bp) in the Basque population and three different alleles (279bp, 283bp, and 287bp) in the Catalan population. Both populations' most frequent allele was 287bp long and was present in over 50% of the samples, similarly to a Finnish population studied by Kittles et al. (1998).

DYS392

This is a trinucleotide repeat marker with an ATA motif.

A total of five different alleles were observed with lengths 248bp, 251bp, 254bp, 257bp and 260bp, with the 257bp length most common in a northern European (Finnish) population (Kittles et al. 1998).

At this locus de Knijff et al. (1997) observed a total of six different alleles (10, 11, 12, 13, 14 and 15) across eight genetically diverse populations. The alleles were named following the recommendations of the ISFH. Of the eight populations allele 13 was most frequent (>40%) in New Guinea/Australian, South East Asian, Arctic and Indian populations. The remaining four populations studied (North East Asian, American, European and African) all had allele 11 at frequency above 40% (deKnijff et al. 1997).

De Knijff et al. (1997) further commented that the allele frequencies formed a bimodal distribution. A bimodal distribution of allele frequencies at this locus is also observed in a Finnish sample population (Kittles et al. 1998).

A total of four different alleles were observed (248bp, 251bp, 254bp and 257bp) in a sample Catalan population and just two alleles in a sample Basque population (248bp and 254bp) (Ruiz-Linares et al. 1999). The most frequent allele in the Basque population was 254bp long and found in 74% of the sample population. In comparison, the most frequent allele in the Catalan population was 248bp long and observed in 55% of the sample population (Ruiz-Linares et al. 1999).

DYS393

This is a tetranucleotide repeat marker with a GATA motif. The repeat sequence is $(GATA)_{13}$ with alleles in the size range 119-131bp and the 123bp allele corresponds to 13 repeats (Carvalho-Silva et al. 1999).

A total of four different alleles were observed with lengths 120bp, 124bp, 128bp and 132bp, with the 128bp length most common in a Finnish sample population (Kittles et al. 1998). Ruiz-Linares et al. (1999) also classified alleles by their PCR product length at this locus. Three alleles sized as 120bp, 124bp and 128bp were observed in Basque and Catalan sample populations. The most frequent allele in both populations was 124bp long and was observed in over 80% of the samples studied (Ruiz-Linares et al. 1999).

At this locus a maximum of five alleles (classified according to the recommendations of the ISFH as 11, 12, 13, 14 and 15) across eight genetically diverse populations were observed (de Knijff et al. 1997). Allele 13 was most frequent (>40%) in New Guinea/Australian, North East Asian, American, European and African sample populations. The Arctic and Indian populations most frequent allele was 14, however only two alleles were observed in the Arctic population compared to four in the Indian population. Finally allele 12 most frequent in the South East Asian population (de Knijff et al. 1997).

2.g. Single Nucleotide Polymorphisms (SNPs)

YAP (DYS287)

The Y Alu polymorphic (YAP) element is a member of the Alu family of repeated DNA elements present on the long arm of the human Y chromosome, Yq11 (DYS287). The insertion event possibly originated during the past 29,000-334,000 years and is now present (YAP+) in some humans but not in others (YAP-) (Hammer 1995). Hammer (1995) was first to locate the YAP element and reported a clear pattern in the frequencies of the insertion. The sub-Saharan Africans had the highest YAP insertion frequency, followed by the northern Africans, Europeans, Oceanic populations and Asians. However, an exception was found, whereby a high frequency of the insertion was observed in a Japanese sample population (Hammer 1995).

Similarly to Hammer (1995) Ruiz-Linares et al. (1996) found African populations to have the highest frequency of the insertion and interestingly 45% of the

Japanese sample population. The Alu insertion was observed in only one from fifteen European samples (7% frequency), and absent in Melanesian, New Guinea and Chinese populations (Ruiz-Linares et al. 1996). Hammer and Horai (1995) in agreement with Ruiz-Linares et al. (1996) observed 42% of Japanese males to have the YAP element, and absent in males from Asian populations (Taiwan and Korea). Santos et al. (1999) recorded a lower European Alu insertion frequency of 4% in comparison to Ruiz-Linares et al.'s (1996) European frequency of 7%. However, in agreement with previous studies (Hammer 1994, Hammer and Horai 1995 and Ruiz-Linares et al. 1996), Alu insertions were present at a high frequency in African populations and absent in Asian and Indian populations (Santos et al. 1999).

In addition to the insertion / deletion event at the Yq11 site, polymorphic nucleotide sites encompassing the YAP insertion site plus a variable length poly(A) tail associated with the YAP element were also observed (Hammer et al. 1997). A total of five YAP haplotypes could be distinguished; haplotypes 1 and 2 were YAP- and haplotypes 3-5 were YAP+. A global survey of these haplotypes revealed that all the haplotypes could only be found in sub-Saharan African populations, and various combinations of the haplotypes found in non-African populations (Hammer et al. 1997).

sY81 (DYS271)

First identified by Seielstad et al. (1994) as an A to G transition. PCR and digestion with *NlaIII* (Jobling and Tyler-Smith 1995) can monitor the base transition.

The derived allele G at this locus has only been observed in African populations and solely in connection with Alu insertion (YAP+) (Ruiz-Linares et al. 1999, Jobling et al. 1997, Underhill et al. 1997 and Seielstad et al. 1994). Scozzari et al. (1999), examined African genetic diversity and observed the derived allele G at frequencies of 85.3%, 53.8% and 43.3% in western, central and southern Africa respectively and virtually absent in the rest of the continent.

92R7

92R7 is a probe used with restriction fragment length polymorphisms (RFLPs). It detects seven bands in a *HindIII* digest in males but none in females (Mathias et al. 1994). One of the detected bands is polymorphic, either 4.6kb or 6.7kb in length. Mathis et al. (1994) observed no variation with other digests and purported the lieklihood that the polymorphism detected with 92R7 was caused by a point mutation at a *HindIII* site. Developments are being made to make detection of the polymorphism PCR compatible (Hurles, Santos, Pandya and Tyler-Smith unpublished data). The ancestral state of this polymorphism was difficult to define as examination of ape DNA produced a different pattern of digestion (Mathias et al. 1994). Jobling and Tyler-Smith (1995), followed Mathias et al.'s (1994) protocol and also observed that the ancestral state for this polymorphism was unknown. However, Jobling and Tyler–Smith (1995) used the binary nomenclature to express the state of the polymorphism (0=4.6kb, 1=6.7kb) in agreement with Mathias et al. (1994). Hurles et al. (1998) also used the same binary nomenclature and termed the '0' as ancestral, although this has not yet been proven using higher primate DNA.

The 4.6kb allele was observed in all Asian samples, and with varying frequency across Italy (52% in the North and 60% in the South) (Mitchell et al. 1997). Similarly, 64% of European caucasians were observed to have the 4.6kb allele (Mathias et al. 1994).

SRY465

Thomas et al. (1998) included this single nucleotide polymorphism as part of a multiplex system, to examine the male lineage of the Cohanim and Levites (see Thomas et al. 1998; Thomas et al. 1999).

SRY10,831

This single nucleotide polymorphisms was identified using sequencing strategies (Whitfield et al. 1995). The standardised protocol involves PCR followed by digestion with the restriction enzyme *Mae*III. The mutation from the ancestral to the derived (A to G) form of the nucleotide codes for the *Mae*III site. When digested the ancestral form produces two fragments (180bp and 100bp) and the derived form, three fragments (142bp, 100bp and 38bp) (Scozzarri et al. 1999). SRY10,831 (or SRY1532) is not a unique event polymorphism. In fact, although a derived form of the G allele from A is present, there is also evidence of a reversion to the ancestral 'A' allele (Hammer et al. 1998; Santos et al. 1999, Karafet et al. 1999). This reversion to the ancestral allele occurred after the 92R7 mutation to the derived form (Santos et al. 1999). Whitfield et al. (1995), analysed samples from chimpanzee and human subjects. The chimpanzee sequence contained the A allele, hence this was considered the ancestral form of the polymorphism.

A world wide study expressed little variation between diverse populations. The G allele or derived form was present in African, Indian, Central East Asian, Mongolian, European and Native American (Santos et al. 1999). The A allele or ancestral form was present at lower frequencies than the derived form in African (1 in 18 samples), Indian (12 in 55), Mongolian (4 in 45 samples) and European (1 in 53 samples) populations (Santos et al. 1999). Whitfield et al. (1995) used just five subjects and one chimpanzee for the sequence variation. The five subjects origins were European, Melanesian, Rondonian suri, Tsumkwe san and Mbuti pygmy with respective forms polymorphism being, A, G, G, A and G.

It was interesting to observe in Whitfield et al.s' (1995) study that the only European sample tested had the ancestral form of the allele especially when Santos et al. (1999) found just 1 ancestral form in a sample population of 53 Europeans.

SRY4064

Sequence variation at this locus was first recognised by Whitfield et al. (1995) during a survey of five ethnically diverse humans and a chimpanzee. The chimpanzees' form of the polymorphism (G allele) was regarded as ancestral with the derived form observed as the A allele.

The PCR protocol was used to detect this single nucleotide polymorphism, whereby two sets of primers were designed. One set were specific to the A allele and the other set specific to the G allele. The A allele fragment measures 164bp and the G allele fragment measures 189bp (Scozzari et al. 1999).

The SRY4064 alleles are closely associated with the YAP element. The ancestral form of SRY4064 (G allele) was exclusive to YAP- (no Alu insertion). YAP+

have the SRY4064 A allele (derived form), although a subgroup of the YAP+ also contained the SRY4064 G allele (ancestral form) (Altheide and Hammer 1997).

TAT

First identified using single stranded conformation polymorphism (SSCP) methodologies (Zerjal et al. 1997). TAT was the name given to the primers designed to amplify a 112bp fragment that included the polymorphism. The 112bp fragment was analysed by *Hsp92*II digestion followed by electrophoresis. TAT-C alleles were confirmed by further digestion with the *Mae*II enzyme (Zerjal et al. 1997).

The single nucleotide T to C transition polymorphism was purported to be restricted to a subset of Asian and European populations indicating genetic evidence for a substantial Asian paternal contribution to northern European populations (Zerjal et al. 1997). A number of genetically diverse populations were analysed for the TAT polymorphism with the derived form observed in European and Mongolian populations and absent in Central East Asian, Indian and African populations (Santos et al. 1999).

Interestingly, it has been observed that within the TAT-C lineages, the polymorphic locus DYS19 has low diversity and variance values associated with short alleles (low average repeat number) (Carvalho-Silva et al. 1999).

M9

This single nucleotide polymorphism is characterised by an ancestral C allele and a derived G allele. The ancestral element was determined from the chimpanzee polymorphism (Underhill et al. 1997). Underhill et al. (1997) observed that the M9 defined a major lineage found in all geographic regions except Africa. This was suggestive of the mutation occurring outside of Africa, alternatively the M9 mutant may still be undetected in few descendants who survived drift and selection.

The derived form of the polymorphism was found at high frequencies in Asia and Australia, and less common in Europe and Africa (Karafet et al. 1999).

M13

This single nucleotide polymorphism is characterised by a C to G transition and can be detected using PCR and digestion methodologies (Underhill et al. 1997). The derived form of the polymorphism has been observed in 42% of the Sudan population and nowhere else (Underhill et al. 1997).

M17

This Y marker is characterised by a deletion of one base pair rather than a base substitution (Underhill et al. 1997). It has been observed primarily with Central Asia, India and Pakistan (Underhill et al. 1997).

M20

This single nucleotide biallelic marker is characterised by an A to G substitution and has been observed in Pakistan (Underhill et al. 1997).

2.h. Haplotype analyses

The term haplotype as defined by Goldstein and Schlotterer (1999), is a term describing a haploid genotype with respect to a specific set of alleles across two or more loci. Haplotypes are in linkage disequilibrium, hence inherited as unaltered units from generation to the next.

The loci selected in this study for haplotype analyses incorporate slowly evolving polymorphisms, (for example single nucleotide polymorphisms), to draw a basic Y chromosome tree, and also more rapidly evolving microsatellite polymorphisms allowing one to discriminate between independent Y chromosomes (Jobling and Tyler-Smith 1995). However, with respect to rapidly evolving polymorphisms, two identical haplotypes may not necessarily be derived from a recent ancestral haplotype as haplotypes may mutate to another via multiple pathways (Deka et al. 1996). Therefore, to represent the relationships between the different haplotypes a network rather than a phylogenetic tree may be constructed. The hypothetical 'ancestral' (or most common) haplotype is central in the network and haplotypes with the least variation to the ancestral haplotype are observed closest to it (Hurles et al. 1998). Similarly haplotypes with the greatest difference are most distant to the ancestral haplotype. The network is not arranged by population and there is often a lack of clustering of haplotypes from the same population or geographic region. However, distinctive ethnic clusters can be apparent indicative of a specific Y chromosome haplotype Deka et al. (1995).

2.i. The population genetics of Y-chromosome haplotypes

In the year 1992, Roewer and workers examined simple repeat sequences consisting of a GATA motif isolated to the Y-chromosome. The polymorphisms were deemed suitable for application in forensic casework, linkage studies and studying ethnological questions (Roewer et al. 1996). During the latter half of the 1990's the polymorphisms detected on the Y-chromosome have greatly increased (Mathias et al. 1994, Hammer 1995, Underhill et al. 1997). Also, the knowledge of male lineages across most populations (Ruiz-Linares et al. 1996), including those of Oceania (Hurles et al. 1998, Hagelberg et al. 1999). Furthermore, specific questions have been answered as to the origins of peoples, for example the Lemba 'black jews' of southern Africa (Spurdle and Jenkins 1992) and the ancestry of the Cohanim and Levite priests (Thomas et al. 1998) which previously would have remained a mystery with mtDNA and autosomal analyses.

There are at present an estimated 18 single nucleotide polymorphisms or rare event markers and a further 23 microsatellite markers and a minisatellite marker to type the Y-chromosome (Jobling et al. 1997). Therefore the possible combinations of haplotype are many thousands.

For the purposes of this study, discussion of haplotypes will be confined to those included in this study and also with reference to known Polynesian and European markers.

This study carried out University College London, Department of Anthropology incorporated the same multiplex loci and UEPs with additional UEPs as was previously used to determine the ancestry of the Cohamin and Levite Jewish priests (Thomas et al. 1998). The UEP haplotypes observed were; i) YAP-, SRY4064-G, sY81-A, SRY465-C, 92R7-C and TAT-T; ii) YAP-, SRY4064-G, sY81-A, SRY465-C, 92R7-T and TAT-T; and iii) YAP+, SRY4064-A, sY81-A, SRY465-C, 92R7-C and TAT-T. The key issue to observe here is the occurrence of YAP+ and SRY4064 A, this association is consistent (Scozzari et al. 1999).

A further association, that is possibly population specific, is that of the YAP+ and allele G at locus sY81 (DYS271). This has only been observed in African populations (Jobling and Tyler-Smith 1995, Ruiz-Linares et al. 1999).

2.j. Haplotypes in Oceania

Underhill et al. (1997) studied haplogroups in Oceanic populations and found 54% (28/52 samples) were of type: YAP-ve, M9 (derived form), M4 (derived form), 25% (13/52 samples) had the YAP-ve M9 (derived form) M4 (ancestral form) and 19% (10/52 samples) had the ancestral haplotype found in all populations including higher primates.

The most common European haplotype (YAP-ve, M9 (derived form)) was also isolated in Oceanic populations and Central and East Asia, Pakistan and Tibet (Underhill et al. 1997).

Hurles et al. (1998) examined European Y-chromosomal lineages in Polynesians. Similarly to this present study, Hurles et al. (1998) used the biallelic markers; YAP, SRY10,831, 92R7, M9 and SRY1532. Additionally, Hurles et al. (1998) also used the markers; DYS199, M4 and SRY3225. The microsatellite markers analysed were; DYS19, DYS389I and II, DYS390, DYS391, DYS392 and DYS393 and the minisatellite MSY1 was also typed (Hurles et al. 1998). Following the allele sizing strategy of de Knijff et al. (1997) a Polynesian specific haplogroup was isolated. The Polynesian specific haplogroup consisted of 27% of the Polynesian sample (9/33 chromosomes) sharing the same MSY1 subtype. Of these nine Polynesian chromosomes there were eight different microsatellite haplotypes, the allele repeat ranges for each locus were; DYS19 12-13 (2-3), DYS389I 10 (2), DYS389II 15-17 (1-3), DYS390 23-25 (4-6), DYS391 10-11 (2-3), DYS392 13-14 (3-4) and DYS393 12,13 and 15 (1,2 and 4). The numbers in brackets are the allele numbers designated by Hurles et al. (1998). The biallelic markers were: DYS199 (ancestral), YAP-ve, SRY3225 (ancestral), SRY1532 (derived), 92R7 (derived), M4 (ancestral), SRY2627 (ancestral) and M9 (derived).

A second haplogroup studied by Hurles et al. (1998) contained chromosomes found in both Melanesians (3/58 samples) and Polynesians (19/33 samples). These belonged to two different MSY1 subtypes. Of these 22 samples, eleven had different microsatellite haplotypes, the allele repeat ranges for each locus were; DYS19 12-14 (2-4), DYS389I 9-11 (1-3), DYS389II 16,17 and 19 (2,3 and 5), DYS390 20-21 (1-2), DYS391 10-11 (2-3), DYS392 11-12 (1-2) and DYS393 13-14 (2-3). The biallelic markers were; DYS199 (ancestral), YAP-ve, SRY3225 (ancestral), SRY1532 (derived), 92R7 (ancestral), M4 (ancestral), SRY2627 (ancestral) and M9 (ancestral).

It is noteworthy to observe differences between Hurles et al.'s (1998) and Underhill et al.'s (1997) findings. The derived form (G) of the M4 allele was characterised as 'Polynesian' and found in over 50% of the chromosomes studied (Underhill et al. 1997). In comparison, Hurles et al. (1998) did not observe the derived form of this allele in either the exclusive Polynesian or the Polynesian/Melanesian haplogroups. In fact the derived form (G) of the M4 allele was only observed in 37 (64%) of 58 Papua New Guinean chromosomes.

2.k. Comparison of Y-chromosome microsatellites in the Pacific

The allele repeat numbers at specific loci vary between populations (Hurles et al. 1998, Forster et al. 1998) as do the sequences of alleles (Forster et al. 1998). The most common Western Samoan microsatellite haplotype (5/7 chromosomes) was characterised as; DYS19 –16, DXYS156-Y –11, DYS390 –20, DYS391 –10, DYS392 –12 and DYS393 –14 (Forster et al. 1998). Other Western Samoan and North Coastal Papuan chromosomes differed from this haplotype by single-step mutations at one or more loci. Highland Papuans had on average larger alleles at three of the six loci (DXYS-156-Y, DYS390 and DYS392), similar sized alleles at two loci (DYS391 and DYS393) and smaller alleles at one locus (DYS19) (Forster et al. 1998). The Western Samoan haplotypes were similar to Hurles et al. (1998) haplogroup 2 containing Polynesian and Melanesian chromosomes.

Deka et al. (1995) used five microsatellite haplotypes to form a 'network' of 15 genetically diverse populations. Included in these populations were Samoan chromosomes that formed a distinct cluster within the network analysis. The most common haplotype observed in the Samoan cluster was DYS391-283bp (or 10 repeats Carvalho-Silva et al. 1999), DYS388-138bp, DYS395-127bp, DYS394-255bp and DYS390-199bp. The amplified products were sized differently to Carvalho-Silva et al. (1999). Deka et al. (1995) followed Jobling and Tyler-Smith's (1995) allele sizes as well as locus and primer descriptions. The primer sequences for loci, DYS390/1/3 and DYS388 were the same as listed in the genome database. Therefore, the amplified lengths of the alleles should theoretically be the same between Carvalho-Silva et al.'s (1999) and Deka et al.'s (1996) studies. The exception to this is the observation of a one base addition at locus DYS390 in Carvalho-Silva et al.'s (1999) study. This may have been caused by an adenylation reaction during the PCR process.

2.1. Forensic use of Y chromosome haplotypic data

Despite the observed lack of diversity between individuals in comparison to autosomal marker systems, highlighted further by apparent mutation frequencies, Y-chromosome data has a useful place in forensic studies.

Forensic scientists require a highly discriminatory system to identify individuals. There is a clear advantage of using Y chromosome haplotype analyses when examining mixed DNA samples in the case of rape, where both the female and male fractions are mixed and need to be separated. However, caution should be noted here, as instances where father and brother are possible suspects if all three are from the same lineage the Y chromosome alone will not be informative enough. It could therefore be argued that the use of the Y chromosome is restricted to rape cases alone where two unrelated suspects are to be tested (Jobling and Tyler-Smith 1995).

Jobling et al. (1997) wrote of the merits and pitfalls of the use of the Y chromosome in forensic analyses. If a suspects haplotype were compared to a reference population, care would be need as the ethnic divisions used in US databases are not ideal and Y chromosome analyses may highlight a population subdivision not necessarily apparent through autosomal analyses. However, the microsatellite diversity of the Y chromosome makes them the markers of choice as opposed to UEPs and minisatellites. As with all evidential material the integrity of DNA for analysis is important. Practically, the Y microsatellites are suited to analysing degraded DNA samples as they use small PCR amplicons. However, if degraded DNA is to be used the MSY1 locus may be uninformative as minisatellite alleles of length 1.7-2.7kb are used. Although if good quality DNA is provided this locus would be of preferential use due to its very high diversity (Jobling et al. 1997).

Instances where Y-chromosomal analyses are of particular forensic use:

i) Deficiency cases in paternity testing, where an alleged father is unavailable. In these instances, male blood relatives can be tested, and if different Y chromosome profiles are observed then an 'exclusion' is reported.

ii) In instances such as rape cases where the assailants' semen and victims vaginal cells are mixed, autosomal typing would be inconclusive without a suspect profile for comparison. Y-chromosome analyses will type the male component easily, without the need for differential extraction of DNA. A clear advantage of this system is that blood-blood or blood saliva mixtures can be analysed where differential lysis could not be applied (Jobling et al. 1997).

Overall, Jobling et al. (1997) note that in practice the Y chromosome analyses will be only part of a battery of other autosomal analyses providing a more detailed picture.

Jobling and Tyler-Smith (1995) write of the use of the Y chromosome in population genetic studies and note that one of the long-term aims would be to produce a tree depicting the evolutionary relationships of modern Y chromosomes. This could also include the origin or root of all lineage's and the ability to date the branch points.

Chapter 3

The prehistory of Man in the United Kingdom

3.a. The Leicestershire Caucasian DNA Samples

The U.K. DNA samples used in this present study were all from the county of Leicestershire in the East Midlands region of England. All the samples were randomly chosen, unrelated individuals, whose families had resided in Leicestershire for at least three generations. This reduced the possibility of selecting DNA samples not 'native' to the county or region. Archaeological and placename evidence of the East Midland region in England, evinces the ancestry to be strongly influenced by Danish, Norwegian and Scandinavian invasions and settlements. Although the DNA samples used in this study ensure the ancestry for only three generations, ancient northern European gene flow to Leicestershire may persist through genetic transmission. The Leicestershire DNA samples used in this study are a subset of those used by Mastana and Sokol (1998) in their examination of genetic variation in the East Midlands. Eighteen conventional genetic systems were used in Mastana and Sokol's (1998) study. In general, the results indicated no statistically significant differences in comparison to other regions of the U.K and some European countries, and similar results to those reported for several populations of Europe (Mastana and Sokol 1998). Although the Leicestershire samples were observed to be similar to the European 'gene pool' a bias may occur if one were to classify the Leicestershire samples as typically European. Therefore, the Leicestershire samples could be described as a 'subgroup' of the European gene pool although not representative of the total Northern Europeans.

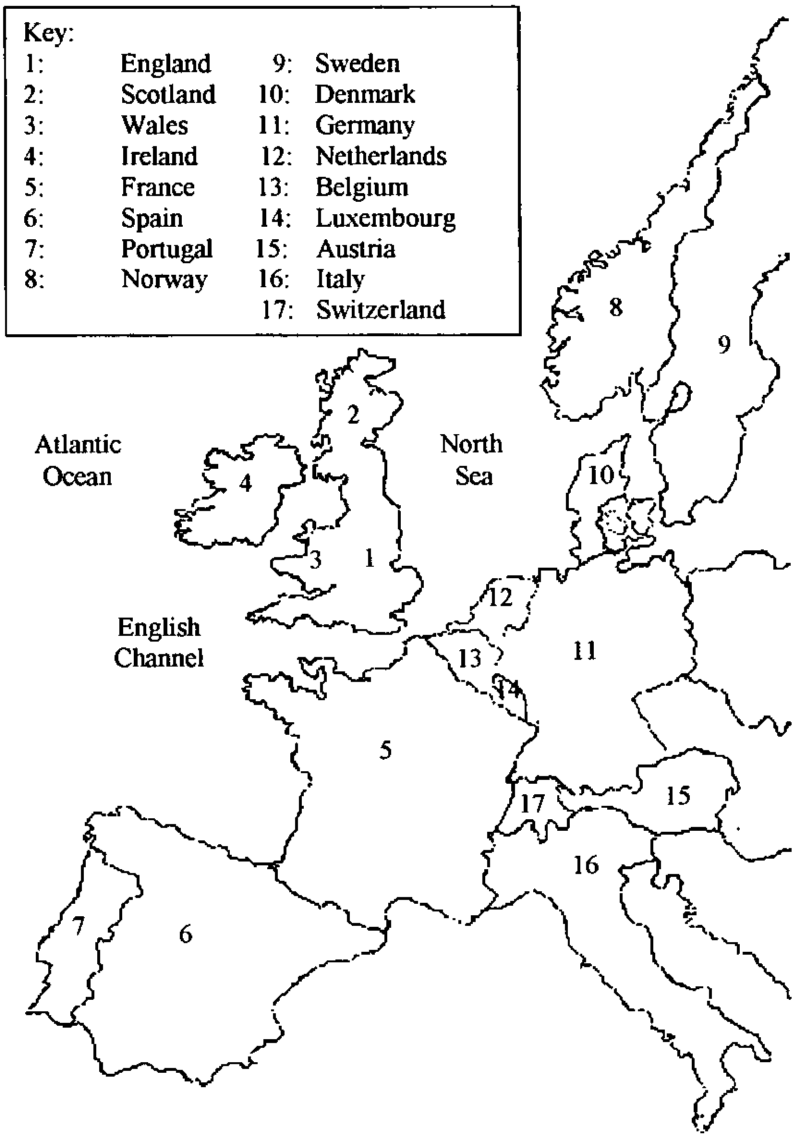
The purpose of this study was to examine 10 novel autosomal tetranucleotide short tandem repeat loci and male lineages analysing Y chromosomal short tandem repeat loci and single nucleotide polymorphisms. Therefore, the UK Leicestershire DNA samples were an interesting population to compare to the Polynesian sample population.

3.b. Geographic and Climatic considerations

'It cannot be stressed too strongly that Britain was on the edge of the inhabitable world, an inhospitable land, much of it difficult to tame, all of it subject to inclement weather. The rapid dissemination of new ideas was made difficult by such geographical obstacles as the great mountain spine separating east from west, and surrounded by dangerous and treacherous seas which made the dissemination of ideas from the Continent difficult in the extreme' (Laing and Laing 1980).

Although this refers to the movement and exchange of ideas between peoples and lands, it can as easily describe the lack of gene flow. The inhospitable environment to man made the U.K an undesirable land in which to live, until the ambient climate became warmer and the land more arable (refer to figure 3.1 for a European map depicting the position of the British Isles in relation to other land masses).

FIGURE 3.1: EUROPEAN MAP.
COUNTRY BORDERLINES ON THE CONTINENT REPRESENT PRESENT DAY BOUNDARIES AND ARE NOT A REPRESENTATION OF PREHISTORICAL BOUNDARIES



3.c. Palaeolithic Britain

Ancestral Danes crossed were able to cross land from Denmark to the British Isles before it was submerged by what is now called the North Sea (Laing and Laing 1980). However, they were not the only inhabitants of the United Kingdom. The Augrignacians' and their culture spread westward from central Europe and south-west Asia to their furthermost habitation, 'remote' Britain. The ancient 'Gravettians' also roamed the land. These peoples were hunters of the prehistoric mammouth. The Gravettian origin has been traced to southern Russia (Hawkes and Hawkes 1958). The hunters and gatherers feeding from the plentiful offerings Britain offered were not replenishing their harvests. Towards the end of the Palaeolithic period, the climate was considerably warmer. The ice sheets receded rapidly and the natural habitat was changing. The hunter-gatherers bereft of crop sowing and harvesting skills adapted to their new environment (Hawkes and Hawkes 1958).

3.d. Mesolithic

By 8000 B.C the population increased from hundreds to thousands, with the beginnings of a distinctive culture. Evidence of Danish contact at such an early stage was apparent by their distinctive stone axes and array of small flints. The Danes traversed the North Sea and were of a Forest culture known as the Maglemosian, adept in the art of fishing and fowling (Hawkes and Hawkes 1958). A further culture with a strong French descendancy was the 'Magdalenians'. These peoples preferred the sandy coastlines to the Forests and marshlands the Maglemosians inhabited.

Most importantly, there was archaeological evidence of the cultures mixing and living harmoniously side by side (Hawkes and Hawkes 1958). Thus, it was also then probable that the gene pools of the different cultures did not remain isolated and admixture of the cultures occurred.

3.e. Neolithic Britain: the age of control

Simple farming communities travelling northwards from southern France embraced the 'primitive' British. The skeletal remains of the southern European people revealed they were of slight build with long narrow heads and delicate features, not unlike the Mediterranean race (Hawkes and Hawkes 1958). The demic diffusion of these people was observed from southern England to northern England and parts of Ireland (Hawkes and Hawkes 1958).

3.f. Bronze Age Britain - Archaeological evidence

Bronze Age Britain was a time of technological advancement. The use of metals for tool and weapon making was greatly advanced by the invasion of peoples (about 1900 B.C) already accustomed to manipulating metal, in particular bronze (Hawkes and Hawkes 1958). The invaders were warriors from the Continent labelled as the 'Beaker people'. This name was given as the Beaker people commonly used a distinctive type of pot (Laing and Laing 1980, Hawkes and Hawkes 1958). Skeletal remains of the Beaker people were different to the Neolithic peoples they would have encountered. Beaker people had, in general, rounder skulls with pronounced brows in comparison to the native British Neolithics (Hawkes and Hawkes 1958). The origin of the Beaker people is an archaeological 'bone of contention'. Hawkes and Hawkes (1958) reported inland Spain to be their origin, thereafter spreading across much of Europe mixing with other violent peoples. Burgess (1974) however, believed ancient Germany to be the source of the British Beakers, although acknowledges their occupation across Europe.

3.g. Late Bronze Ages

The late Bronze Age was beginning to mark broad racial differences. The Mediterraneans were purported to have dark skins and were smaller in stature in comparison to the tall, fair Nordics (Hawkes and Hawkes 1958).

However, towards the end of the Bronze Age, the 'Celts' dominated the British Isles. The Celts originated in France and West Germany, although their expansion to southern and eastern Britain around 800 B.C, was a necessity brought about by an increasing demand for land and resources as population numbers increased (Hawkes and Hawkes 1958). Waves of Celtic invasions continued together with fresh ideas of metal work from the Mediterranean. It was postulated that around 500 B.C the Celts were arriving on southern British shores with iron blades. These peoples were forced from their homelands by flooding of their villages and farmlands and by increasing warfare in Continental Europe (Hawkes and Hawkes 1958).

3.h. Iron Age

The transition from Bronze Age to Iron age was clearly defined in the south, however further north and to the east it appeared that the diffusion of the Celtic influence was taking longer (Hawkes and Hawkes 1958). Therefore, the Southerners were more technologically advanced than the Northerners, as the contact with Continental Europeans aided the transfer of ideas (Cunliffe 1974).

During the first century before Christ, the Roman Empire increased in strength. The roman emperor Julius Caesar had conquered the south of France and much of Germany. The people of Belgium in the north-east, were difficult to overthrow not least because of aid from generations of Belgium settlers in southeast Britain (Hawkes and Hawkes 1958). The Belgian settlers were barbaric people and expanded their population, within twenty years, from Kent to Hertfordshire (Hawkes and Hawkes 1958). Decades passed with the growth of Belgian dwellings to kingdoms. However, the strong rule was not to last. The death of their Ruler left the Belgians in turmoil and Britain vulnerable to Roman invasion (Hawkes and Hawkes 1958).

In 43 A.D a successful Roman army invasion occurred and again Britain was about to change. Forty years of warfare and the movement of Romans north from the British south coast ended with the invaders advancing no further than the Scottish Highlands. The terrain too unrewarding, Roman government relinquished its battle to occupy all of the British Isles and just the cultural identity of prehistoric northern Britain survived (Hawkes and Hawkes 1958). The Roman occupation in Britain was primarily military in nature, marked by the construction of roads forming a network across the country. However, it was only the physical structure that 'stood the test of time', very little of the language and customs remained beyond their occupation in Britain (Leeds 1913). By 400 A.D the Romans had left (Leeds 1913). They were recalled to Rome to defend it from constant attack, leaving Britain to fend for itself. Three main nations then attacked Britain to claim fertile land and overthrow the 'worthless Britons': the Saxons, Angles and Jutes (Taken from: The Anglo-Saxon invasion of Britain http://www.fordham.edu/halsall/basis/gildas-full.html).

The numbers of invaders were large enough to affect the British culture at that time, even replacing the primary language. However, the native population was not wiped out (Taken from: The Anglo-Saxon invasion of Britain http://www.fordham.edu/halsall/basis/gildas-full.html).

The genetic implications of the repeated invasions can be summarised.

The Roman occupation of Britain lasted roughly half a century during this time one could understand the gene pools of the native Britons and Romans mixing. At the time the Roman army left Britain, the Roman gene pool obviously suffered a severe bottleneck. The Saxons, Angles and Jutes entering Britain displacing native Britons from their villages, would have at the very least added to the European gene pool in Britain and reduced the native Britons gene pool through warfare.

3.i. The Saxons

The Saxons invaded English shores and by a gradual process of absorption, made England their own. The absorptive process included the conquering of villages, followed by their own settlement and alteration of agricultural methods to suit their lifestyle. The Saxon invaders originated along the length of the southern coast of the North Sea in particular northern Germany. There was little communication between the Saxon settlements and intertribal violence was common (Leeds 1913). No archaeological evidence has been excavated to indicate the presence of Anglo-Saxons in the Celtic inhabited Wales, West Coast of England, Scotland or Ireland. Furthermore, skulls recovered from burial grounds among the purported Saxon settlements, had characteristics not found in England in Roman or pre-Roman times. Instead the physical features of these skulls were similar to features observed in North Germany (Leeds 1913).

3.j. The Angles

The largest numbers of invaders came from the Danish peninsula and were called the Angles. Archaeological records indicated their settlements were within central and East England in regions known then as East Angles, Middle Angles and Mercia (Leeds 1913). The early kingdom of Mercia included all the territory 'watered' by the river Trent. Leicestershire was greatly favoured by the first settlers in particular through the river Soar valley, recognised by the number of 'Danish' burial sites in the area (Leeds 1913).

By the time of the eighth century, seven Anglo-Saxon kingdoms existed. Mercia formed part of the West Midlands. Mercia was bordered by all the Old English kingdoms except Kent and Sussex (Cameron 1988). The division of the country to *shires (counties)* took place in the middle 9-10th centuries. The East Midland shires came into existence during the 10th Century. Derbyshire, Nottinghamshire, and Leicestershire represent a few of the districts ruled by Danish law and occupied by divisions of the Danish army. Leicester was a supposed hybrid Celtic-Old English name. Its origins are not truly understood, although believed to be a folk-name derived from a Celtic-river name. The river would not have been the Soar, as this name was recorded at an early stage and has always kept the form *Sore*. However, a river was named *Leire (Lei)*, thus Leicester means 'dwellers on the banks of the *Leire'*. However, the meaning of the river name Leire was unknown (Cameron 1988). Place names within Leicestershire also have strong foreign connections. The hybrid names like 'Leicester' cluster in areas where Scandinavian names are uncommon and where English settlement had taken place before the introduction of the Danes. The explanation for this occurrence was that established English settlements were overthrown by the Danes and the area renamed (Cameron 1988).

3.k. The Vikings

During the period between the 8th to 10th centuries Scandinavians traversed the northern waters terrorising the people they met (Barker and Cook 1977).

Some forty years of battles raged between the Scandinavians and English, to gain complete control of England. The Midlands were the first region chosen by the Scandinavian Vikings as this was previously 'Scandinavian' in character, hence the people were more likely to accept a Viking king. This belief was not misplaced and shortly the rest of England was under the rule of a Viking King (Web address 1996 http:// viking.no/england/danelaw/e-heritage danelaw.html).

3.l. The Population Genetic Implications since the time of the Saxons

The Romans moved from England by 400 A.D after ruling Britons for an estimated period of 400 years. The Romans fashioned lifestyles and architecture also inevitably contributed to the British gene pool forming admixtures with the natives.

An influx of tribal peoples from Germany known as the Saxons invaded England. This would have had a three-fold effect on the gene pool at that time. Firstly, the Saxons killed many native Britons, thus decreasing the native gene pool.

Secondly, the Saxons tribes would have undergone a bottleneck effect when abandoning their homeland and settling in England. Lastly, admixture with the natives may have occurred altering gene pools of both native tribal Saxons and native Britons, through admixed offspring.

The greatest alteration in the British gene pool during the 6th-8th Centuries occurred with the invasion of the Angles from the Danish peninsula. However, their colonisation in Britain was concentrated to within central and east England. Therefore, the gene pool within the midlands would have been the most affected by the large influx of 'Danish' genes with little affect in areas outside of this region. In addition Scandinavian Vikings between the 8th-10th centuries attacked England slaying many Saxons. Thus, localised settlements of Saxons were destroyed and replaced by the Scandinavian Viking army. However, in retaliation the Saxons massacred as many 'Danish' people as possible. This would clearly have reduced the Danish gene pool in both localised areas and regions. The Vikings replied to the insult and did not cease their warfare until the Vikings had control of all of England. Though much bloodshed occurred in the conquering of these lands, under one rule peace prevailed and instead of the further rivalry between nations.

3.m. The Normans

The 1st Century A.D saw much unrest with invasion from the Normans including the famous Battle of Hastings in 1066A.D. England had regressed from civilisation to savagery (Size 1930).

No further foreign invasions occurred after 1154 A.D, however civil wars were common and 'peasant' revolts altered the course of Reign in England from the 15th century onwards.

3.n. Population Genetic implications since the time Normans

The middle of the eleventh century saw the savagery of war when at least three rival factions were brought together within the confines of the relatively small British Isles.

The native Britons at the time of the Norman invasion were admixed with Scandinavians, Danes and Norwegians, thus an admixture of the respective gene pools occurred. However, the extent of the admixture may have been confined to specific regions. The Viking invasion in 1066 although short-lived caused many English and Viking deaths. Furthermore, the following battle between the Normans and English reduced the English male gene pool further. The influx of Normans from France lasted roughly 200 years and brought close links with the continent.

The Vikings did not attack again, therefore, for the most part, the Nordic gene pool in Britain remained the same or was incorporated into an admixed gene pool. At an early stage in history, regional cultural variations existed revealed by the archaeological remains. The most obvious example of this was the high concentration of Viking peoples in the Midlands region of Derbyshire, Nottinghamshire and Leicestershire. Therefore, a regional genetic variation would have also been present. The 'Class' systems also ensured a level of genetic variation, whereby, the upper classes of nobility seldom were associated with the lower class of the 'peasant'.

3.o. The Biological Evidence

Relatively little biological and genetic evidence of British history and prehistory of individuals with an UK ancestry exists (Roberts 1973, Kopec 1973, Mastana and Sokol 1998, Lanchbury et al. 1990).

Distinct genetic regional variations throughout the British Isles have been obscured by geneflow between them (Roberts 1973). However, subtle genetic differences between regions have been observed (Mastana and Sokol 1998). The movement of armies and their women, merchants and refugees over large distances, brought about appreciable gene flow, and a presumed reduction in genetic difference between the North and South, East and West of Britain (Roberts 1973). Random variation in gene frequencies due to geographical isolation was also considered (Roberts 1973).

The source of genetic variation in the U.K. has rarely been attributed to historical invasions and subsequent settlements (Piazza 1993; Kopec 1973; Lanchbury et al. 1990 and Mastana and Sokol 1998). Therefore, few studies have compared regional variation to that found in nearby Continental European populations (Lanchbury et al. 1990 and Mastana and Sokol 1998).

In 1940, Fisher and Taylor examined blood group frequencies throughout England and Scotland. The country was divided into three distinct regions, Scotland, Northern England and Southern England (see Kopec 1973). The frequency of blood group O decreased and A increased from north to south, whilst blood group B was higher in Scotland and found at low levels in both northern and southern England. Interestingly, Kopec's study (1973) expanded upon Fisher and Taylor's study and found not only an increase in the frequency of blood group A north to south but also from east to west, reaching a maximum in northern east Anglia. Kopec (1973) did not wish to comment on the implications of historical settlements on the influence of the cline of blood group 'A' frequencies, because of the world wide variation in frequency of this blood group. However, Imaizumi (1974) did observe that the regions with the highest frequency of blood group A corresponded to some extent to the locations of historic Norwegian or Viking settlement.

A genetic study of the East Midlands using conventional blood group and serological systems of analysis revealed little difference between the A and O blood group frequencies (Mastana and Sokol 1998), with no major difference to previous regional studies (Kopec 1973 and Imaizumi 1974). Comparisons of East Midland allele frequencies were made to Danish, German and Norwegian samples. Similar frequency patterns were observed (Mastana and Sokol 1998). It was interesting to observe the genetic distances between Leicestershire and North-East Derbyshire to the European populations (Germany, Norway and Denmark) was less than the other East Midland populations (Mastana and Sokol 1998).

3.p. U.K. Leicestershire Caucasians: Regional variations and the short tandem repeat.

In the year 1995, Gill and Evett analysed a number of Caucasian, Asian, African and Chinese populations at four tetranucleotide repeat loci. Included in the U.K Caucasian populations was a Derbyshire sample population and two general U.K populations. Also included were two European populations from Germany and Sweden. Interestingly the Derbyshire population had a smaller genetic distance to Sweden than to either of the general U.K populations (Gill and Evett 1995). However, this finding may be observed with caution, as although a very large Derbyshire sample population was used, far fewer Swedish samples were analysed which may slightly bias the results. In the year 1998, Watson et al. examined the World wide allele frequency distribution at the HUMTHO1 short tandem repeat locus. A total of twenty one Caucasian populations were included in the study. The allele frequencies of the U.K Derbyshire population at the HUMTHO1 locus were taken from Gill and Evett's (1995) study. Comparing population data, similar allele frequencies were observed between the Derbyshire population (Gill and Evett 1995) and a Danish Caucasian population (Nellemann et al. 1994), French Caucasian population (Pftzinger et al 1994) and Dutch Caucasian population (Sjerps et al. 1995). Although no firm conclusions can be drawn as to the genetic affiliations between the aforementioned populations, it would be of interest to conduct more conclusive genetic investigations (see Watson et al. 1998).

3.q. Origins of Genetic Disorders

Clinical diagnoses and prevalence of specific genetic disorders and diseases have led scientists to trace the origins and movements of historic and prehistoric man (Hummel et al. 1999). For example, a mutation on the male sex chromosome causing a haemochromatosis has been reported in various European populations. However, the highest recorded incidence of this mutation was observed in residual Celtic populations in the U.K and France (Hummel et al. 1999). Similarly, certain types of skin cancer were claimed to be more prevalent in persons with a Celtic ancestry than those of non-Celtic ancestry (Long et al. 1998). However, as with any inheritable disease, not all people of Celtic ancestry were affected and with genetic admixture increasing with every generation, successfully tracing all inheritable disorders to their origins would be more problematic.

3.r. In Summary

The ancestry of the native British population of today has been well documented through archaeological and historical findings (Leeds 1913, Hawkes and Hawkes 1958). However, conclusive genetic evidence is not available and only tenuous links between British Regions and Northern European countries can be made (Mastana and Sokol 1998 and Lanchbury et al. 1990). The English county of Leicestershire in the East Midlands has a varied and interesting history of foreign invasions and settlements (Leeds 1913, Hawkes and Hawkes 1958, Renfrew 1984 and Laing and Laing 1980), with close historical 'ties' to the Scandinavians and Danes (Mastana and Sokol 1998).

Although the U.K. Leicestershire population in this study may not be 'representative' of a European population, it presented an interesting comparison to the New Zealand Maori and Polynesian Islander populations.

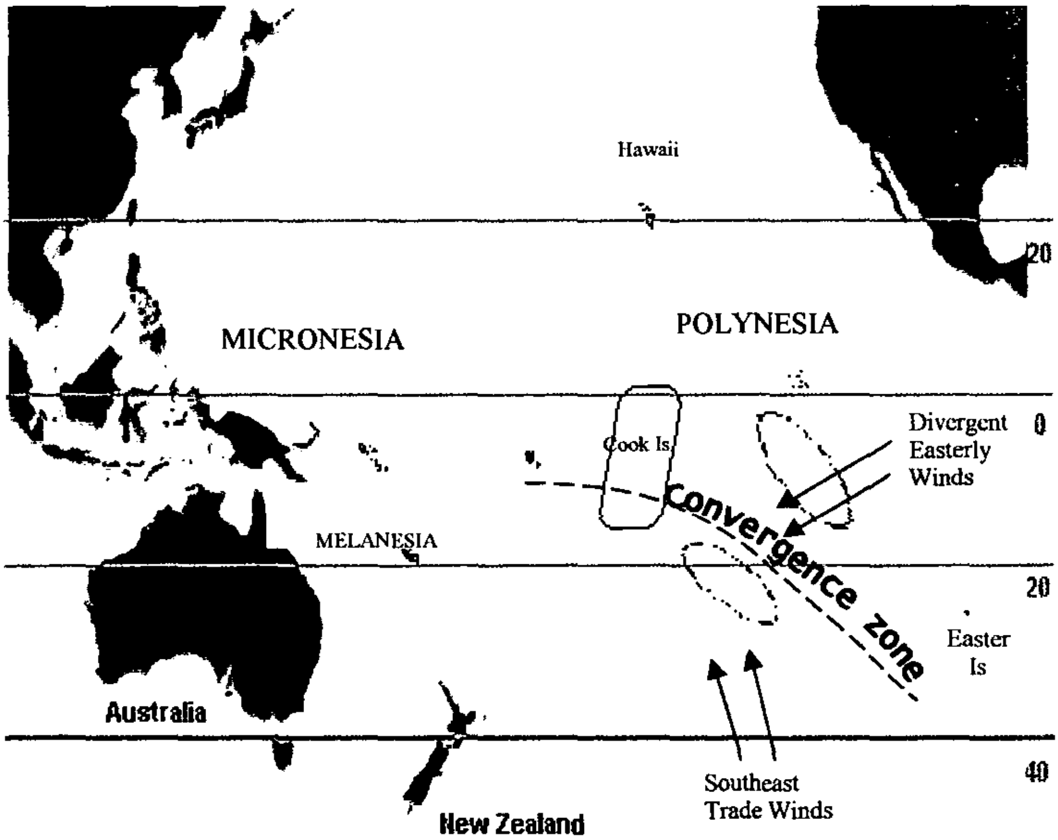
Chapter 4

The Prehistory of Man in the Pacific Islands

The Pacific Ocean covers one third of the surface of the Earth, scattered in it are atolls, basaltic Islands, and volcanic islands. These islands can fall into two distinct groups, those of the continental regions in the Western Pacific and those on the Pacific plate that are east of the 'Andesite Line' (Irwin 1992, Thomas 1965).

The larger and higher islands in the west (Australia and the large Melanesian Islands which run from New Guinea eastwards to Fiji) are geologically diverse and rich in resources, hence an ideal environment to support settlements (Bellwood 1989). Conversely, the Basaltic islands on the Pacific plate offer less resources, with deep valleys alluvial soils and treacherous reefs.

The most extreme habitats are atolls, these are volcanic islands sunk or eroded below the surface of the ocean. The atolls are characterized by a lagoon encompassed with a reef, with sand 'bars' only a few meters above sea level,



notably difficult to support many diverse colonies (Irwin 1992; Terrell 1986).

The sea level some 50,000 to 30,000 years ago was 40-70m below current levels due to the effects of the Ice Age and sea crossings were shorter than at present. Though the sea level was lower, island-bridging would have only occurred to the Bismark-Solomon region. In order to understand how the Pacific was populated one should consider all the environmental conditions influencing and affecting colonization (Bellwood 1989).

Much of the explanation of the prehistory of the Pacific begins with human settlements in southeast Asia and Australia roughly 30,000 – 40,000 years before present and termed the 'Pleistocene' period. 'Neolithic' man in the Pacific has been dated from about 4000-6000 years before present and was characterized by primitive farming and the use of polished stone, flint tools and weapons.

4.a Weather Patterns.

The warm air at the equator rises and flows toward the north and south poles where it cools descending to the surface regions (30°N and 30°S). The cold dense air from the poles flows towards the equator and upon mixing with the warm air at 30° latitude causes the mixture to ascend at 60° latitude.

At the time of Polynesian settlement, westerly winds would have affected New Zealand and the Chatham Islands especially in winter, with varying wind direction and strength. Inspite prevailing easterly winds in the tropical pacific, variations occurred allowing progress to the east. Monsoonal reversal of the southern-hemisphere in the western pacific, blew westerlies and northerlies into Melanesia (Irwin 1992). The trade winds and currents in many tropical regions where the Pacific Islands occur tend from East to West during most of the year. Furthermore, an El Nino phenomenon draws winds and currents eastwards across the Pacific towards South America (Bellwood 1989). John Williams a missionary in the Society Islands from 1817-1839 noted that, '*the direction of the wind is not so uniform as to prevent the Malays from reaching the various islands and groups, in which their descendants, I believe now found.*' Further to his observations, he knew of the westerly winds, their Tahitian names, the months they blew and their pattern of change (Irwin 1992).

4.b Human movement through the Pacific – the prehistory

' The settlement of the Pacific Islands is the final chapter in man's occupancy of the terrestrial quarters of the Globe' (Ward 1972). No one person has disputed this fact, however the 'homeland', timing and direction of the settlement of the Polynesians in the Pacific Islands had differed with scholarly opinion since the eighteenth century. These differences of opinion can be grouped into three 'theories' (refer to figure 1 for a flow chart of the three theories).

The first theory was the so-called 'express train to Polynesia' (Diamond 1988). This theory assumed a precursor 'Polynesian' culture termed Lapita originated in Island southeast Asia and spread rapidly west to the Bismarck Archipelago (Melanesia) (Diamond 1988). Bellwood (1989) in agreement with Diamond (1988) further detailed the ancient expansional events via two major stages. Firstly, between 50,000 to 30,000 years before present (ybp), hunting and gathering populations crossed to Australia, New Guinea the Bismarck Archipelago and Northern Solomons. Secondly, around 4,000 ybp, during an agricultural expansion, horticultural populations (Lapita) with voyaging technology entered southeast Asia either of Indonesian or southern Chinese origin. Thereafter, the Lapita traversed the 'unexplored' ocean separating the Bismarcks from Samoa within a few hundred years (Diamond 1988).

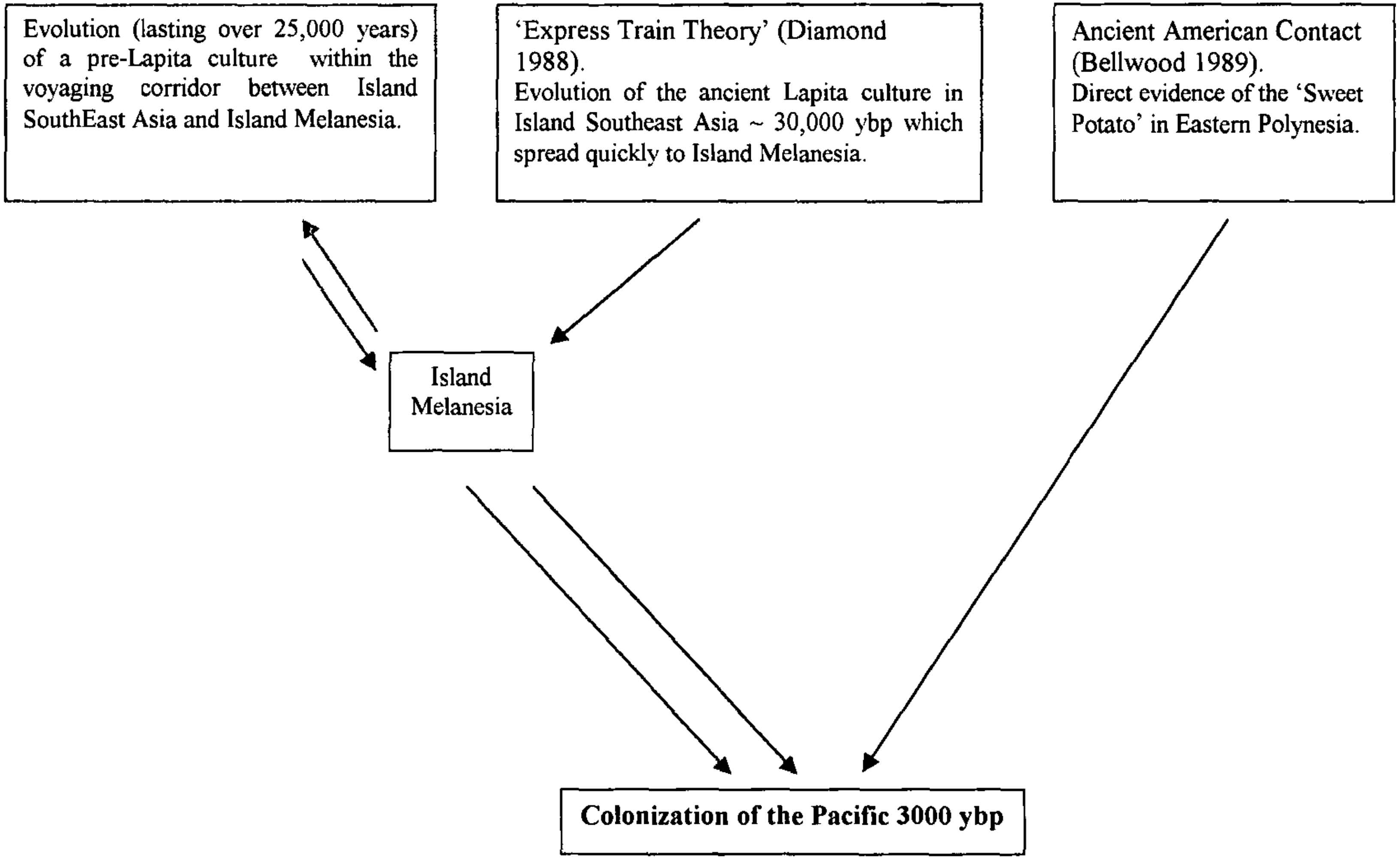
The second theory, was presented by the anthropologist Terrell (1986), backed by Irwin (1992) and later Richards et al. (1998). The first pre-Polynesian colonists in this theory were those of the Pleistocene era (pre-Lapita). The origins and development of the 'Polynesian' culture occurred between Island southeast Asia, coastal New Guinea (Inland Melanesia) and the Bismarck Archipelago (Island Melanesia) (Terrell 1986, Irwin 1992). This formed a safe sailing 'corridor', whereby the culture evolved for 25,000 years (Irwin 1992). Settlement expansion was fastest in southeast Asia and slowest in inland Melanesia, with Lapita pottery found in the Bismarck Archipelago dated 1,500BC (Irwin 1992). This theory was consistent with the hypothesis that Austronesian languages originated within island southeast Asia during the Pleistocene period migrating to Melanesia and then the remote Pacific during the past 6,000 years (Richards et al. 1998).

An interesting theory was proposed in 1930, which not only predates Terrells' (1986) proposal of origin but also traces the ancestry of the Polynesians and Maoris further back from Island southeast Asia. Without the hindsight of linguistic, archaeological and biological evidence of today, Cowan in 1930 wrote that no date, even approximate, could be given when the Polynesian ancestors began their migration eastwards from south-west Asia to India and Indonesia and then finally the Pacific. Furthermore, the Polynesians formed part of an ancient Gangetic race present in India from remote antiquity, but were modified by the intrusion of Semitic (Hebrews), Tibetan and other races. The ancestors of the Maori, were living in a land known as Atia-te-varinga-nui, believed to be India, about 450 BC and were ruled over by a Supreme Chief. Not long before the beginning of the Christian era Maori ancestors began to migrate to the East Indie Islands, where they gradually settled Eastwards into the Pacific. However, ancient humans were known to be in New Zealand at this time and these were of Polynesian-Melanesian admixture (Cowan 1930). The coastal people of the south-western Asia were among the earliest sailors, with a knowledge of navigation by the stars. Travel to the African continent would have taken place and partial colonization of Madagascar, explaining the Maori-Polynesian and Malagasy language mixture. Monsoons thereafter would have taken founding groups to New Guinea and further to the Western Pacific. This passage of travel would have taken hundreds of years, and the mixing of tribes would have taken place (Cowan 1930).

The third theory, is that of ancient American contact in the pacific prior to the arrival of the first 'Polynesians'. However, archaeological findings in Eastern Polynesia appear to be of Asian rather than South American derivation (Bellwood 1989). Although, one cannot refute the evidence presented in the form of the 'typically American' Andean sweet potato introduced into ancient eastern Oceania (Bellwood 1989). Interestingly, further possible connections with America, stem from the observations of Charles Nelson a keen linguist and ethnologist who travelled extensively across the globe and held a strong interest in Maori-Polynesian history (Cowan 1930). In the 1850's Nelson spent some time sailing in the Pacific and then the Arctic Ocean. Returning along the coast of Alaska, Nelson observed carving on Totem poles, mats, garments, fishhooks and weapons alike to those of the Maori. Furthermore, Nelson met Indians in Nevada who lashed reeds and rushes to walls of their houses and made knots and bends in exactly the same way as the Maoris (Cowan 1930). However, Nelson does not make any assumption as to whether the Indians influenced the Maoris or *vice versa*.

These theories differ on a number of levels not least the scientific disciplines, as Bellwood is principally an archaeologist, Terrell (1986) an anthropologist, Cowan (1930) an ethnographer and linguist and Richards et al. (1998) a geneticist. It has been observed that the use of linguistics in analysing and reconstructing ancient history will not always and need not always corroborate DNA analyses or even archaeological records. Terrell (1986) writes that the Pacific islanders linguistics may vary among themselves although still share a common ancestry. Hence, linguistic differences may have developed during the course of human settlement.

FIGURE 4.1: THREE THEORIES OF THE COLONIZATION OF THE PACIFIC



Evolution (lasting over 25,000 years) of a pre-Lapita culture within the voyaging corridor between Island SouthEast Asia and Island Melanesia.

'Express Train Theory' (Diamond 1988).
Evolution of the ancient Lapita culture in Island Southeast Asia ~ 30,000 ybp which spread quickly to Island Melanesia.

Ancient American Contact (Bellwood 1989).
Direct evidence of the 'Sweet Potato' in Eastern Polynesia.

4.b.i. Movement and settlement since 6,000 years before present.

Irwin (1992) disagrees with Wards' (1972) date of the settlement of New-Guinea-Australia (of 20,000 years), instead Irwin notes that archaeological evidence of settlement approaches a much earlier date (40,000 years). Furthermore, a New England cave site (33,000 years ago) was indicative that coastal marine and lowland tropical forests were used as resources for both food and shelter.

At 6,000 years before present a widespread and diffuse human ancestry may have existed within Western Melanesia and possibly southern China (Bellwood 1989). At 3,000 years before present, Polynesian ancestry can be placed in the Tonga-Samoa region (Bellwood 1989, Irwin 1992, Terrell 1986), with the final settlement in the South Pacific Islands, including New Zealand, during 1,000 AD (Ward 1972, Hill and Serjeantson 1989).

In the early 1970's the radiocarbon dating technology provided a crude outline of the course of Island settlement. Western Micronesia was purported to be inhabited by 2000 BC and the south east Melanesian Islands of New Caledonia and Fiji by 1000BC and the western Polynesian Islands of Tonga and Samoa before the birth of Christ (Ward 1972).

The settlement of more isolated Islands in the Pacific was believed not to be from deliberate voyages of exploration and colonization, rather accidental discovery by canoes, either blown off course or those exiled. Ward (1972) believed that humans inhabited south-east Melanesian islands of New Calendonia and Fiji by 1000 B.C and in the western Polynesian islands of Tonga and Samoa much before that. Ward (1972) also reported that the earliest date for colonization in the periphery of eastern Polynesia was around 500 AD with the central Islands dated almost 500 years later.

It was predicted that the settlement of New Zealand occurred during the last 1000-2000 years before present and was not a single colonization, rather a wave of settlements throughout a few hundred year span (Irwin 1992). Although this has recently been disputed (Murray-McIntosh et al. 1998).

4.b.ii.The Canoes

However, little information about the types of craft used to cross the sea, have been uncovered by archaeologists. Bark canoes and bamboo rafts have been suggested and Irwin (1972) observed that in a period of 25,000 years during the Pleistocene (600,000 years before present), water-crossing boats would not only have varied but more importantly improved.

Founder groups attempting to cross the sea to settle on an Island could scarcely contain less than four or five people, hence boat sizes would have needed to be substantial even if basic (Irwin 1992, Murray-McIntosh et al. 1998).

Undisputedly, the voyaging canoe was central to Polynesian culture. Seafarers ancestral to 'Polynesians' traversed the waters in craft capable of sailing hundreds or thousands of kilometers (Finney 1977). The crafts would have been of sufficient size, most probably double canoes, to carriage migrants, food, water and domesticated plants and animals (Finney 1977). This in turn indicated intentional voyages and knowledge of at least basic sailing and navigational skills, to safely travel against the direction of prevailing winds and currents (Finney 1977). However, evidence suggests that westerly winds blew several times per summer for a few days to one week facilitating travel Eastwards (Irwin 1992). Tongans, Cook Islanders and Tahitians used these winds to stage travels Eastwards (Finney 1977).

4.c. The scientific evidence of human settlement and movement

The theories of migration have been examined with little reference to the actual scientific evidence. This section concentrates on the academic disciplines of archaeology, anthropology, linguistics and biological marker systems.

In the years preceding the use of genetic analyses in the examination of population affinities, a wealth of archaeological anthropological and linguistic evidence was studied to date and trace how Polynesia was populated.

4.c.i. Archaeological and Anthropological

The discovery of the Lapita culture and its significance in piecing together Polynesian pre-history.

In the 1920's archaeologists excavating sites on the Tongan Islands recovered potsherds of a plain ancient pottery. Twenty years on similar pottery was excavated from Islands in New Caledonia and Melanesia. These findings dissolved the once thought 'boundary line' between Melanesia and Polynesia which had been based on obvious racial distinctions (Terrell 1986). These racial distinctions are founded on the assumption that the Melanesians drifted from the north west to settle in the 'isolated' Fiji islands for a lengthy period of time before Polynesian groups traversed the waters westward from Tonga. These Polynesian groups were considered 'more developed' than the Melanesians, although they adopted some of the Melanesian culture, thus integrating both groups forming an admixed population (Thurn 1914).

The distinctive type of pottery, 'Lapita' was named after an archaeological site in New Caledonia excavated in the 1950's (Terrell 1986). Lapita, together with a range of stone and shell tools provided evidence of communities of specific cultures in the southwest Pacific spanning supposed boundaries between Melanesia and Polynesia (Ward 1972). Lapita pottery has been further located on Islands from Papua New Guinea in the west to Tonga and Samoa in the East. Carbon dating estimated the pottery to be 3,600 years old. It would be reasonable to assume that people who knew how to make Lapita pottery moved from place to place rather than the technical skills were passed from one ancient community to another.

Lapita pottery not only linked Polynesian with Melanesian but also early south east Asian pottery, thus inferring an Asian 'homeland' (Bellwood 1989). Peter

Bellwood and John Terrell have long been in disagreement as to the origins of the Polynesians. In the year 1997, Terrell et al., wrote the 'Human diversity and the myth of the primitive isolate'. In this, the simple assumption that people who lived on islands led isolated lives was no longer accurate. Rather, societies did not exist in isolation with cultural and biological traits combining and recombining (Terrell et al. 1997).

Bellwood commented on the paper published by Terrell et al. (1997) and argued that a debate about isolation and interaction based on the 'myth of the primitive isolate' created a 'smokescreen' without attacking the 'real issue'. The real issue Bellwood insisted, was the debate of the concept of a radiative dispersal of Austronesian speaking peoples from Island South East Asia to the Pacific roughly 3,500 years before present. However, as Bellwood pointed out, Island south east Asia rich in archaeological prehistory with cultures and traditions closely resembling those of Pacific Austronesians, was omitted from Terrell et al.'s (1997) arguments. Patrick Kirch also commented on Terrell et al.'s (1997) paper with similar criticisms to Bellwood. Firstly, Kirch observed the 'myth of the primitive isolate' to be of Terrell et al.'s (1997) own making and that inter-island voyaging was an important part of the pacific culture. The term 'isolation' varies in meaning between islands. Hawaii is isolated by some 3,862km from the nearest occupied archipelago, in comparison Samoa is isolated although one to three days travelling ensures contact with several other islands.

Decorated lapita pottery was not found in any of the excavated sites in Micronesia and so its prehistory cannot be linked closely to the southern Melanesian Islands (Bellwood 1989).

4.d. Linguistics

It was estimated that more than 1,600 languages have been spoken in the Pacific Islands, South East Asia and nearby mainland (Terrell 1986).

These are arranged into three groups, referred to as Australian, Papuan and Austronesian (Bellwood 1989). The Australian languages are confined to Australia with no clear links with nearby regions. However, one should note that gene flow from the North could occur without altering language.

Austronesian languages were purported to have spread from a 'homeland' in Southern China and Taiwan within the last 6000 years. There was a great geographical spread of this language, hence subgroups or variations in this language were vast (Bellwood 1989). However, the sub-grouping in the Pacific Islands was easy to define if one ignored western Melanesia. Bellwood (1989) observed that the location of the proto- Austronesian start-point could be placed in Taiwan. Subsequent proto-Austronesian expansions indicated the Phillipines, northern Borneo, Sulvesi and the Moluccas may have been settled before movement to western Indonesia and Malaya. Thereafter, the Austronesian language formed a sub-group in Oceania termed Oceanic. The expansion of this language followed the route; Admiralty Islands to Solomon Islands to Vanuatu. From that point forward, the settlement of Polynesia occurred from Tonga and Samoa spreading to the outlier islands and atolls (Bellwood 1989). The outlier atolls were believed to have involved population assimilation or replacement, although this was believed to be a rare occurrence (Bellwood 1989).

According to comparative linguistics, Polynesian settlement was straightforward. Combining linguistic information with archaeological artefacts, the initial settlement was in Tonga and Samoa and possibly east Uvea and east Futuna and occurred during the late second millennium BC. Furthermore, the languages of these island groups may have begun to diversify from the beginning with the largest separation occurring with the settlement in eastern Polynesia around 300 B.C, however this was purely speculative. At that time, the Tongan and Samoan languages would have diverged, with Tongan developing with little diversification until European contact.

Terrell (1986) agreed, regarding the Lapita settlers of Tonga as forming the basic elements of Polynesian language and customs. By the middle of the first millenium BC, the Polynesian way of life had been carried through to Samoa by Tongan colonists. Thereafter, Samoa became a hopping stone for the Polynesians to colonize the East Pacific. However, it would appear that the assumption that Polynesians first became 'Polynesian' in Tonga was incorrect, as archeaologists had found Lapita pottery in Samoa. Therefore, the first colonization of Samoa could not have taken place after the Polynesian way of life had been formed. It seems more appropriate to assume from archaeological evidence that the early settlers of Fiji, Tonga and Samoa who made Lapita pottery formed the beginnings of a Polynesian culture (Terrell 1986).

Anatomists have speculated that peoples of Tonga, Fiji and Samoa and resulting Polynesia may owe their origin to an extremely small founding group of people, who may have been a biased sample of closely related people not at all typical of the population they left behind. Therefore, the belief was that Polynesians were too different from Melanesians to have originated in Melanesia. Perhaps through a lack of diversity in a small founding population which caused the divergence (Terrell 1986).

In the 18050's Charles Nelson an accomplished linguist sailed along the East African coast. During his time aboard ship he learnt a little of the Suahili language. He observed similar words with similar meanings to both the east Africans and the Maoris, which suggested that the Maori were akin to peoples inhabiting Arabia, Egypt and other parts of Africa. Further corroborating Nelsons theory, some distance from the mouth of the Lower Euphrates (Africa) were bitumen springs. The Hebrew word for spring is 'mimis'. A maori would pronounce this as *mimiha* which is the exact word used to designate the balck bitumen often chewed by the Maoris (Cowan 1930).

4.e. Biological systems of discrimination

Since the beginning of the twentieth century, archaeology, linguistics and anthropological studies have tried to uncover scientifically, the truth behind the colonization of the Pacific and the order of events leading to present day populations.

Here, the biological systems describe the biological affiliations in the Pacific and how these findings differ to, or agree with, other scientific disciplines.

4.e.i. Classical Blood Systems

The twentieth century has seen many scientific advancements and biological basis of identification is no exception. The 'classical' marker systems including ABO, MNS, and Rh, since their discovery have been noted for their use in anthropological studies (Kirk 1989). Cavalli-Sforza and Edwards in 1964 first demonstrated that blood genetic marker data could be used for genetic-distance estimates of the relationships between human populations on a worldwide scale.

In the year 1931, an early study examining the blood groups of the Maori revealed striking differences between the different tribal groups (Phillips 1931). The differences were thought to reflect the descendents of the 'original' Maoris named 'Maoriri' and the immigrants colonizing New Zealand at a later date

(Phillips 1931). Phillips (1931) purports that the Maori immigrated to New Zealand from Hawaii roughly 600 years before present, only to encounter an Aboriginal Maoriri race. The Maoriri were characterised by their fair skins and reddish tint in their hair. These natives were brutally slain by the Maori immigrants whilst others fled to the hills. Later generations formed an admixed group, of which individuals had characteristic red tints in their hair and fairer skins (Phillips 1931). Within the Maori tribes the most frequent blood group was A and the least frequent blood group observed was B (Phillips 1931). Similarly 56 years later, Woodfield et al. (1987) also observed over 50% of the population to have blood group A and an 'under representation' of groups B and AB. A further study of blood group systems reported the Maoris were easily distinguishable from the American Indians (North and South), the Siamese, Burmese and Malaysians. However, there were possible connections between Indonesians and other south Asian Mongoloids (Lehmann et al. 1958). Kirk (1989) observed varied ABO blood grouping distributions in the Pacific, however a generalisation was made. The subgene A_2 was rare or absent in the populations tested, also blood group B was largely absent (Kirk 1989). Early studies of the MN blood groups revealed a general pattern in the Pacific of a decline in the frequency of the M allele gene North to South (Kirk 1989).

Nei and Roychoudhury selected populations 'representative' of Polynesians, Micronesians, New Guineans and Australian Aboriginals. This study concluded the populations fell into six groups:

i) Caucasians, including English, Iranian, and North Indian,

ii) Mongoloids, including Malay, Chinese, Japanese, Polynesian and Micronesian

iii) South American Indians,

iv) North American Indians and Eskimos,

v) Australoids, - the combination of Australian Aborigines and New Guineans, and

vi) Negroids, including nigerians, Bantu and Bushmen.

These workers accepted that some of these groups would evolve at different rates due to passing through 'bottle-necks' or random genetic drift (Kirk 1989).

4.e.ii. Mitochondrial DNA

The ability to trace lineages and the movement of individuals rather than populations has a clear advantage for predicting the colonization movements throughout the Pacific. Mitochondrial DNA (mtDNA) is present in large numbers within somatic cells. It is inherited independently of nuclear DNA and is generally transmitted almost exclusively maternally. MtDNA encodes for only 13 polypeptides of which none are directly involved with its own replication, transcription or translation. Therefore, these factors are believed to contribute towards mtDNA faster rate of evolution that genomic DNA (Lum et al. 1994).

Hagelberg and Clegg (1993) reported that an occupation of Polynesia by Melanesians was probable. Ancient human skeletal remains recovered throughout Oceania were used for mtDNA sequence analysis. Implications have been made that the Lapita culture entered central Pacific via Melanesian indigenous inhabitants as opposed to Austronesian-speaking migrants of Southeast Asia. Furthermore, the effects of population bottlenecks and genetic drift removed nearly all mtDNA lineages during the colonization of the eastern Pacific (Hagelberg and Clegg 1993).

Lum et al. (1994), failed to support Hagelberg and Cleggs' (1993) findings and found three lineages in Polynesian populations. The mitochondrial region V deletion marker present in high frequency in Polynesian populations expressed sequence variations that allowed three lineages to be assumed. The first group found in Remote Oceania included roughly 95% of Native Hawaiian, 90% Samoan and 100% Tongan mtDNA samples. The second group represented the predominant maternal lineage group of Papuan Melanesia. The third group tentatively linked Samoa to Indonesia. Lum et al. (1994) believed that these lineages corresponded to an admixture of Asian and Melanesian peoples who inhabited near Oceania and populated the isolated island archipelagoes. Melton et al. (1995) further substantiated the aforementioned link between Polynesians and Southeast Asians.

Interestingly Sykes et al. (1995) also observed more than one lineage in Polynesia and believed 'the major prehistoric settlement of Polynesia was from the west and involved two or possibly three genetically distinct populations'. Mitochondrial DNA analyses of Lum et al.'s (1994) study, indicated 94% of the Polynesian lineages were clustered forming one group and originated from a diverse group present in the western Pacific, that migrated to the eastern margin of Polynesia. The most notable aspect of mtDNA analyses was the lack of diversity and severe bottlenecks which was predicted to have restricted gene flow eastwards.

Archaeological evidence dated the colonization of Fiji at 3,200 ybp and Samoa 3,000 ybp. Beyond Samoa genetic diversity fell steeply indicating genetic bottlenecks as settlement took place in more remote areas of the Pacific. Sykes et al. (1995) proposed that there was a long settlement in Samoa before moving eastwards.

A comprehensive study testing migration patterns across Polynesia using mtDNA analyses indicated that between 50 –100 women could have founded New Zealand (Murray-McIntosh et al. 1998). Furthermore, there was little evidence to suggest an earlier permanent settlement especially since copious vegetation and no disease would be conducive to a rapid population expansion (Murray-McIntosh et al. 1998).

Investigations using DNA samples from live subjects in Polynesia today, will be subject to various forms of genetic admixture. Easter Island is one such example whereby it was purported to be almost completely depopulated of its native inhabitants after European contact. At present it is populated mostly by Chileans (Hagelberg et al. 1994). Therefore, the use of ancient human bones to extract DNA for mitochondrial analyses one method for re-tracing 'lost' female lineages in Polynesia. Hagelberg et al. (1994) extracted DNA from human bones dated from the 2nd century AD. All the samples expressed the same DNA sequence at the tested loci and were consistent with samples from 'present-day' Polynesians located near New Zealand and Hawaii. Hagelberg et al. (1994) presumed that the presence of only one principal mtDNA lineage in Polynesia was a consequence of population bottlenecks during colonization. Furthermore, Hagelberg et al. (1994) did not find evidence of ancient American contact in Polynesia, and Hurles et al. (1998) observed little European contact with respect to mitochondrial DNA analyses.

4.e.iii. Autosomal genetic relationships

Autosomal chromosomes are inherited biparentally and so contain genetic information from both parents. Therefore, unlike the mtDNA which is inherited maternally or the Y chromosome which is inherited paternally (thus tracing female or male lineages) autosomal DNA 'blends' the lineages of the origins of the parents.

Autosomal genetic marker systems used in population and forensic genetic work have been discussed elsewhere, therefore this section concentrates on the markers employed to examine the genetic relationships in the Pacific.

The purported differences of colonization numbers of men to women and differences in male and female gene flow between 'isolates', favored the use of mtDNA (Hagelberg and Clegg 1993, Murray-McIntosh et al. 1998) or Y chromosome (Hurles et al. 1998) analyses over autosomal analyses. However, autosomal analyses have provided additional valuable information to complement mtDNA and Y chromosomal studies (Lum et al. 1998, Chu et al. 1998 and Deka et al. 1995).

Through the use of short tandem repeat marker systems, it has been observed that Polynesians and Micronesians share 70% of their nuclear alleles with Asians and 30% with Papuan-speaking Melanesians. Also, remote Oceanic Islanders express close genetic affinities to the near Oceanic populations from highland Papua New Guinea and Australia (Lum et al. 1998). These findings were consistent with Chu et al. (1998), whereby Polynesians were tentatively linked to Austronesian speaking southern Asian populations.

The observation that there was a reduction in genetic diversity from southeast Asia to the remote Oceanic populations has been documented using mtDNA analyses (Hagelberg et al. 1998, Murray-McIntosh et al. 1998) and classical marker studies (Woodfield et al. 1987). This observation has also been documented using autosomal STR loci, particularly in New Guinea highlander and Samoan populations (Deka et al. 1998).

4.f. Y-Chromosomal Evidence

A chapter has been devoted to the Y-chromosome including Polynesian studies. Therefore this section will be focused to the theories of male lineages in Polynesia.

In contrast to mitochondrial studies, Hurles et al. (1998) observed over 55% of Polynesians shared a recent common origin with Melanesians, and 33% of Polynesians had a European ancestry. The European admixture observed in Polynesians was attributed to the post-settlement contact of European sailors, traders, and missionaries (Hurles et al. 1998). Similarly, Spurdle et al. (1994) observed the Polynesians to cluster closely to Caucasians. Interestingly Spurdle et al.'s (1994) study did not show a close relationship between Maoris and Samoans. This was explained as a prehistoric European gene flow into the Maori population, which was in contrast to Hurles et al.'s (1998) post-contact theory.

In general, with respect to mtDNA analyses, an increase in genetically homogeneous populations was observed from western Polynesia to the east (Hagelberg et al. 1999). However, the same has not been observed with respect to male lineage studies, whereby more male than female founders have been anticipated or settlements by a few ancient male lineages which have had time to diversify (Hagelberg et al. 1999).

4.g. Foreign Contact in Polynesia

The Greeks were well aware that the Earth was round and as far back as 50AD believed the unknown southern hemisphere to be ocean broken by a continent. Ptolemy 200 AD, proved the earth round and as practical illustration merchants at that time sailed down the African coast five degrees south of the equator. Following the downfall of the Roman Empire many scientific discoveries were lost. Europe at that time harvested many Theologists and pilgrimages were the only recorded journeys for the next few hundred years. In fact it was not until the middle of the 16Th Century that man truly regained the 'voyaging for discovery' spirit. Once again speculations as to the southern hemispheres ocean and landmass arrangement came into question (Beaglehole 1932).

In the year 1513, Vasco Nunez de Balboa, discovered the Pacific Ocean from the west coast of Central America and called this the 'Great South Sea' claiming it for his master the King of Spain **(Web site** www.knight.org/advent/cathen/02216c.htm). However the 'Great South Sea' or Pacific Ocean was not sailed across until 1521, by the Portuguese voyager Ferdinand Magellan, who from 1518-1522 was the first to circumnavigate the globe (Web sit www.mariner.org/age/magellan.html). Although he only saw two small uninhabited islands until he reached the Marianas (West Melanesia), it prompted the beginning of European exploration in the late 16Th Century (Ward 1972).

The introduction of European intervention in the Pacific caused direct alteration of the habitation by introducing 'foreign' plants to the Islands, disease to the people and slaughter to at least two hundred indigenous people. It was estimated that from the time Magellan entered the Pacific to the middle twentieth century the Oceanic population decreased from 3.5 million to two million. This was believed to be a direct consequence of the introduction of smallpox, measles, typhus, typhoid, leprosy, syphilis and tuberculosis, which were previously unknown in the Pacific (Ward 1972).

Of the Islands Magellan used to replenish his stocks of fresh fruit and vegetables, the natives were not receptive of the sailors. In return Magellan burned houses and boats and killed at least ten men, labeling them savages (Beaglehole 1932).

In the 18th Century over 100 ships sailed in the Pacific, included in these voyages was the HMS Endeavour under the Captaincy of James Cook. Europeans were curious of the peoples of the Pacific, wanting to know how they settled such a vast expanse of Ocean and secondly what did their existence imply, with reference to the origin of mankind (Durrans in the 'Captain Cook in the South Pacific' 1979). Therefore, studies of these indigenous peoples and recordings of their behaviour, customs, language and beliefs were recorded by naturalists. One such naturalist, G Forster who traveled with Captain Cook, noted in 1773 that

> '.... the further the nations live from the equator, the less numerous they are and of those within the tropics they are most numerous who are more civilised than the rest. The inhabitants of the South Seas are remarkably different in colour, form, habit and natural turn of mind. People at Taheitee and the Society Isles, New Zealand and Easter Island seem to constitute a race of men entirely different from those at New- Calendonia, Tanna and Mallicollo and all the rest living in the New Hebrides............'

Forster continued his observations and believed the two 'races' of men in the south sea arrived there by different routes and descended from two different origins. However, Forster does imply that without any written record or historical fact in favour of his opinion, how could he be certain that his assumptions were correct. Notwithstanding this caveat, he continued,

> ' The five nations of the first enumerated race seem to come from the Northward and by the Caroline Islands, the Ladrones, and the Island of Borneo, to have descended from the Malays. Whereas, on the contrary, the black race of men seems to have sprung from the people that originally inhabited the Moluccas, and on the approach of the Malay tribes withdrew into the interior parts of their Isles and countries.'(Durrans in the 'Captain Cook and the South Pacific 1979 Pg 144)

A unifying thought at that time was how skilled at navigation and boat building were the Polynesians who had successfully populated many outlier islands in the Pacific. Two schools of thought prevailed, firstly that chance happenings of colonization were non-deliberate and secondly that of deliberate colonization, indicating a good knowledge of navigation. The latter was not highly regarded as many Europeans saw the Polynesians not as far along the evolutionary time scale, being a fairly 'recent' population (Terrell 1986).

Kawaharada wrote the 'Settlement of Polynesia' and in this he summarised how Polynesia was colonised with respect to archaeological and linguistic evidence as well as common folk lore;

Using a time scale of 50,000 years before present to 1,000 AD the following time scale of events has been estimated;

i) 50,000 years before present hunter gatherers inhabited Australia and New Zealand

ii) 1600-1200 BC Lapita (as previously described) spread from New Guinea in Melanesia to the east (Fiji, Samoa and Tonga)

iii) 300 BC seafarers from Samoa and Tonga settled islands further east, the Marquesas Islands (Cook Islands, Tahiti-nui, Tuamotus, and Hiva)

iv) 300 AD Settlement of Easter Island from Eastern Polynesia

v) 400 AD Settlement of Hawaii from the Marquesas Islands

vi) 1000 AD settlement of New Zealand (Aotearoa) from the Society or Cook Islands.

Overall though, it would appear that any settlement could not be due to a single migration and from the information gained from archaeological and linguistic evidence. It would seem reasonable to assume re-colonization and re-migrations to islands occurred at various intervals, reasons such as over populating of islands and the need to 'move on' or perhaps populated islands gaining voyagers at various intervals bringing with them new plants/ animals and new traditions. Foreign contact with native Polynesians caused a 'noisy' genetic background (Hurles et al. 1998). There appears to be some confusion as to whether European lineages were an ancient (Spurdle et al. 1994) or recent addition (Hurles et al. 1998) to the Polynesian gene pool.

4.h. Aims of this study

Genetically diverse populations were chosen to evaluate 10 novel autosomal tetranucleotide short tandem repeat loci for use in population and forensic genetics and to examine the relationships between the male lineages.

The diverse populations were required to be able to invoke questions of not only whether, the autosomal loci were capable of correctly differentiating between the populations, but also, whether the loci could correctly separate closely related populations.

In order to test diversity issues, the use of a more isolated population versus an outbred stable population, (of known ethnic origin and admixture) was considered of greatest use.

The Polynesian sample population – an example of a historically 'isolated' population

The populations in Polynesia have received attention from anthropologists (Irwin 1992) and geneticists (Hagelberg and Clegg 1993, Hurles et al. 1998 and Murray-McIntosh et al. 1998) alike. The attention has been focused on the pathway of settlement across Polynesia and the origins of these peoples. Numerous bottleneck events have been purported across the south Pacific with the last colonization event occurring in New Zealand (Murray-McIntosh et al. 1998). However, few studies have specifically examined the New Zealand Maori (Murray-McIntosh et al. 1998). It has previously been reported that extensive European admixture exists in Polynesia, although male lineages exist which are native to the region (Hurles et al. 1998). Thus, one can hypothesise that the admixed DNA samples will share a closer genetic similarity to a European population than their own native population.

The U.K. Leicestershire population – an example of a historically 'heterogenous' population

The U.K. Leicestershire DNA samples have been incorporated previously as part of a regional U.K. study (Mastana and Sokol 1998). Regional genetic diversities within the U.K. have previously exhibited only marginal differences between populations, hence the numbers of alleles observed were also similar (Gill and Evett 1995).

Aims of the 10 autosomal short tandem repeat loci study

- To optimise the polymerase chain reaction (PCR) methodology at each locus and to incorporate a suitable method of electrophoresing and sizing the PCR product.
- To statistically evaluate the potential for each novel polymorphic locus for use in population and forensic genetics. This evaluation procedure incorporated genetically diverse populations with known admixture.
- Hypothesis: The autosomal marker systems will be able to distinguish between the U.K. Leicestershire and the Polynesian populations and a greater polymorphic diversity will be observed among the U.K. Leicestershire population than the Polynesian Islander populations.

Aims of the male lineage study

- To present the first New Zealand Maori male lineage study.
- To evaluate the extent of European admixture within the Maori and Polynesian populations.
- To compare the findings of this study to previous research carried out within Polynesia.
- Hypothesis: The New Zealand Maori males will be closely associated with the Polynesian Islander males, although extensive European admixture among the Polynesian populations will be observed.

Finally, the two studies can be compared and contrasted, in particular specific issues of admixture and the effects it has on analyses and interpretation of results. One can hypothesize that one will observe greater similarities than differences between populations at the autosomal level in comparison to the male lineages whereby clearer population distinctions will be made.

Chapter 5

Section I: The Pilot Methodology

5.a. The gathering of locus information.

The Genome Data Base (GDB), was searched for possible microsatellite marker systems that had not previously been examined for population genetic or forensic purposes. The Utah marker development group, carried out analyses of markers as part of the Genome project (Ballard, personal communication) They were contacted to obtain marker information on tetranucleotide short tandem repeat loci. A battery of marker information was provided which detailed the repeat motif, the primer base sequence and rough guidelines of the size range (in base pairs) of the alleles.

Information on 17 STR loci was provided for which little or no population work had been carried out. These markers all contained tetranucleotide repeats, which are known for their better stability during PCR than dinucleotide repeats (Utah Marker Development group 1995). Also most systems can resolve 4 base pair (bp) differences unambiguously, thus tetranucleotide markers present clearer resolution and allele calling after electrophoresis (Utah Maker Development group 1995).

The primer sequences were designed so that no sequence homology existed between primers or to Alu sequences (Utah Marker Development group 1995).

5.b. Equipment used in this study

- The *Perkin Elmer 480 thermocycler* was used for the polymerase chain reaction technique. This thermocycler was a 48 well array accommodating 500μl microtubes. No heated lid was present, thus oil 'overlay' techniques were employed.
- The *Poker Face II Nucleic Acid Sequencer SE 1650* (Hoefer Scientific Instruments), was used for polyacrylamide gel electrophoresis of the amplified oligonucleotide products. Thereafter silver staining was used to visualize the electrophoresed PCR product.
- *Hybaid submarine gel electrophoresis tanks* provided the means to electrophorese agarose gels to observe the efficacy of the DNA extraction and amplified PCR product.

- *'Elchroms' submarine gel electrophoresis system* provided an alternative to the polyacrylamide gel system. Elchroms apparatus included a circulating submarine gel electrophoresis tank, which has dimensions 40cm X 18.4cm X 17.5cm (w, d, h). Catamarans 'pin-down' the pre-formed gels for use with the tank.
- A *Fisons HAAKE D1 thermostatically controlled thermo circulator* was connected to Elchroms system. This provided a constant temperature of 55°C during electrophoresis.
- An ultra violet transiluminator and UVP Image store 7500 visualized the agarose check gels and spreadex gels stained with ethidium bromide.

5.c. The Polynesian DNA Samples

Dr Geoffrey Chambers of Wellington University New Zealand was contacted regarding DNA samples. In total, 175 ethanol precipitated DNA samples of Maori and Polynesian origin was provided following the guidelines and recommendation of the Wellington Ethics Committee. The blood samples were collected with informed consent from volunteer blood donors and extracted DNA banked according to American Society of Human Genetics guidelines. The samples were stored in Victoria University DNA bank and managed under the supervision from the University and Wellington ethics committees. The ethnic affiliation and admixture were also established by voluntary informed self-declaration at the time of collection.

The DNA was prepared from white blood cells recovered from 10ml blood. The final ethanol precipitated DNA pellets were resuspended in 200 - 1000 μl of tris-EDTA (TE) buffer (see appendix for composition). Aliquots were later assayed by eye after ethidium bromide staining and concentrations were 1 - 2μg per 10μl of stock solution (Chambers. G. personal communication).

5.d. The U.K. Leicestershire DNA Samples

The Leicestershire sample population provided an example of a population within Europe that supports an outbred and diverse genetic history (Mastana and Sokol 1998; see also Chapter 3).

U.K. Leicestershire Caucasian samples were collected by the Human Genetics laboratory, Loughborough University in collaboration with the Regional Blood

Transfusion service, Sheffield in 1992, for human population genetic studies (Mastana and Sokol 1998).

Blood samples were taken at blood doning sessions with informed consent. Voluntary self-declaration of ethnic affiliation of the donor and of their parents as well as places of birth were recorded. Related family members and those who were not native to the Leicestershire area for at least two generations were not included in this study.

In the year 1992, DNA was extracted from whole blood, using the salting-out method and quantified spectrophotometrically. DNA was then aliquotted to a final concentration of 100ng/µl in Tris-EDTA pH8.0 buffer and stored at -20°C (refer to appendix for buffer composition).

5.e. Rehydration and Quantification of the Polynesian DNA:

All 175 DNA samples contained 50µl DNA solution and 125µl 100% ethanol. The DNA was recovered by extensive centrifugation for 20 minutes at 10,000 rpm. The ethanol was decanted and the eppendorfs dried to remove residual ethanol before rehydrating in 50µl of Tris-EDTA pH8.0 buffer at room temperature overnight.

Each tube was then briefly spun to collect any condensation and 1µl alliquots of the rehydrated DNA used for quantification.

5.e.i. Agarose 'check gel' preparation

A 1% ethidium bromide agarose 'check-gel' was prepared using the following reagents (Table 5.1) and methodology:

Reagent	Volume / Weight
Agarose (Seakem LE)	0.3g
1X Tris-Borate EDTA buffer (TBE) (see appendix for composition)	30ml
Ethidium Bromide (ETBR)	0.75µl

TABLE 5.1: AGAROSE CHECK-GEL PREPARATION. ALL CONTENTS HEATED IN A CONICAL FLASK (COVERED WITH A LID OF SELLOPHANE) UNTIL THE AGAROSE DISSOLVES.

- The solution was allowed to cool slightly then poured into a 20-well gel mold (Hybaid electrophoresis equipment) and allowed to set at room temperature for approximately 30 minutes.
- The gel was then submerged in a Hybaid electrophoresis tank containing sufficient tris-borate EDTA buffer (with ethidium bromide) to cover the gel by 3-5mm and the combs removed.

N.B Removal of the combs before submerging the gel resulted in air-bubbles forming in the wells.

1μl of each DNA sample was mixed with 3μl loading buffer (Bromophenol blue) and pipetted into consecutive wells of the gel. The gel was then electrophoresized for 30 minutes at 120V, in tris-borate EDTA buffer. To estimate the concentration and integrity of the DNA samples, known concentrations of 25ng, 50ng, 75ng and 100ng of 'K562' DNA (a known control DNA) were included in every gel. The gel was visualized using ultra violet radiation, and the unknown DNA samples sized to the known K562 DNA standards.

A complete listing of the DNA concentrations is given in the appendix.

- The solution was allowed to cool slightly then poured into a 20-well gel mold (Hybaid electrophoresis equipment) and allowed to set at room temperature for approximately 30 minutes.
- The gel was then submerged in a Hybaid electrophoresis tank containing sufficient tris-borate EDTA buffer (with ethidium bromide) to cover the gel by 3-5mm and the combs removed.

N.B Removal of the combs before submerging the gel resulted in air-bubbles forming in the wells.

1μl of each DNA sample was mixed with 3μl loading buffer (Bromophenol blue) and pipetted into consecutive wells of the gel. The gel was then electrophoresized for 30 minutes at 120V, in tris-borate EDTA buffer. To estimate the concentration and integrity of the DNA samples, known concentrations of 25ng, 50ng, 75ng and 100ng of 'K562' DNA were included in every gel. The gel was visualized using ultra violet radiation, and the unknown DNA samples sized to the known K562 DNA standards.

A complete listing of the DNA concentrations is given in the appendix.

Section II: Amplification Methodology

5.f. The efficacy and efficiency of Perkin Elmers' 480 thermocycler

In order to optimize the PCR protocol, one must first understand the limitations of the equipment used. For example, time taken to reach target temperatures (ramping speed), and temperature difference between heating block and PCR reagent mix in the microtube. If these are known then informed adjustments of thermocycle parameters can be made.

The 'ramping' speed of the thermocycler and temperature within a given microtube at a specific heating block temperature were recorded. A total of 50µl PCR reagent mix and two drops of mineral oil overlay in 7, 500µl microtubes were used to imitate an actual PCR run. These microtubes were positioned in the heating block as shown below:

	1	2	3	4	5	6	7	8
A	●	◎	◎	◎	◎	◎	◎	●
B	◎	◎	◎	◎	◎	◎	◎	◎
C	◎	◎	◎	◎	◎	◎	◎	◎
D	●	◎	◎	◎	●	◎	◎	●
E	◎	◎	◎	◎	◎	◎	◎	◎
F	●	◎	◎	◎	◎	◎	◎	●

The programmed thermocycle was Denaturation: 45 seconds @ 94°C
Annealing: 40 seconds @ 62°C
Extension: 100 seconds @ 72°C

The ramping times for the heating block were taken from the end of one heating step to the target temperature of the next.

Ramping from 72°C up to 94°C took 47 seconds (2.1 seconds per degree)
94°C down to 62°C took 23 seconds (0.72 seconds per degree)
62°C up to 72°C took 34 seconds (3.4 seconds per degree)

It was quite clear that this thermocycler was 3-5 fold more efficient at cooling than heating the block. Heating from 72-94°C was 1.3 seconds per degree quicker than heating from 62-72°C. This indicated that the greater the temperature difference the quicker the ramping, and that controlling for a temperature difference of 10°C required the ramping to be more sensitive, and hence slower.

5.f.i. Microtube Temperature

It was observed that the heating block equilibrated at the target temperature and subsequently started timing, before the PCR reaction mix within the microtubes attained the target temperature. The temperatures within the microtube were recorded (using a thermometer) when the heating block had reached its target temperature (see table 5.2).

TABLE 5.2: TEMPERATURE (°C) OF THE PCR REAGENT MIX AT THE MOMENT THE HEATING BLOCK REACHES TARGET TEMPERATURE.

Microtube Position	Microtube Temp. when ramping finishes from 70-94°C (Additional time to reach target temp. in seconds)	Microtube Temp. when ramping finishes from 94-62°C (Additional time to reach target temp. in seconds)	Microtube Temp. when ramping finishes from 62-72°C (Additional time to reach target temp. in seconds)
A1	86 (30)	75 (36)	66 (40)
A8	84 (40)	74 (40)	66 (40)
D1	84 (45)	76 (35)	66 (40)
D5	84 (45)	77 (35)	66 (40)
D8	84 (45)	77 (35)	65 (40)
F1	84 (45)	73 (35)	66 (40)
F8	83 (40)	79 (45)	63 (40)

On average, the microtube temperature did not reach its target until roughly 40 seconds after the heating block had equilibrated. It would also appear that the temperature difference between the heating block and microtube was fairly consistent across the block. However, there was a slight temperature and ramping speed difference at the corners of the heating block. At the end of 'ramping', the microtube placed in the top left corner was observed to be closer to the target temperature than the microtube placed in the bottom right corner. This may have been an anomalous result or a reflection of how the heat was dispersed throughout the thermocycle block.

The difference in temperatures of the heating block and microtube contents has been a point of poignant discussion (Stamm et al. 1991, Linz 1990). In the year 1990, Linz commented that a thermocycler should guarantee temperature homogeneity for all samples of an individual 'run' and run-to-run repeatability. However, Stamm et al. (1991) accepted that a temperature difference between the heating block and microtube contents will be present and so altered the thermocycle programmes to accommodate the delay in heat transfer.

5.g. Loci PCR/ Thermocycle Optimization

The 17 STR loci for preliminary optimization strategies were;

D1S407, D2S262, D20S156, D2S1279, D3S1514, D4S2285, D4S2289, D5S592, D7S1485, D7S618, D7S1517, D9S252, D10S526, D10S507, D10S520, D10S521 and D12S297.

PCR Reaction mix, varied according to the amount of magnesium chloride required.

The basic PCR protocol as stated in the information provided by the Utah Marker Development group, was used as a basis on which modifications were made.

The Utah Marker Development Groups'

Standard Protocol

- 10X Enzyme-specific reaction buffer (*Advanced Biotechnologies formula*):

  750 mM Tris-HCl, pH 8.8 at 25°C

  200mM $(NH_4)_2SO_4$

  25mM $MgCl_2$
- 0.2mM of each dNTP
- 0.2μM of each oligonucleotide primer
- 1.25 units of a thermostable DNA polymerase
- Nucleic Acid Template, 10-20ng.
- Distilled water to a final volume of 25μl

A master mix containing all the PCR components was kept on ice and aliquots of this were added to the DNA samples. After which two drops of mineral oil were added as an overlay to the reagent mix, which was then ready for the thermocycle programme.

The Utah Marker Development Groups' Thermocycling programme:

- 94°C 5 minutes *Initial denaturation*
- 94°C 20 seconds *Denaturation*
- Annealing temp 20 Seconds *Annealing*
- 72°C 40 Seconds *Extension*
- Decrease annealing temp by 1°C per cycle until the lowest temp is reached.
- Repeat the lowest temp for 25 cycles.

The magnesium chloride concentration and estimated T_m or annealing temperature for the loci were estimated by the Utah Marker Development group. In addition, the group had sequenced each locus and noted the repeat motif and allele size range for each locus (Ballard personal communication). This information has been summarized in table 5.3.

added as an overlay to the reagent mix, which was then ready for the thermocycle programme.

The Utah Marker Development Groups' Thermocycling programme:

94°C	5 minutes	*Initial denaturation*
94°C	20 seconds	*Denaturation*
Annealing temp	20 Seconds	*Annealing*
72°C	40 Seconds	*Extension*

- Decrease annealing temp by 1°C per cycle until the lowest temp is reached.
- Repeat the lowest temp for 25 cycles.

The magnesium chloride concentration and estimated Tm or annealing temperature for the loci were estimated by the Utah Marker Development group. In addition, the group had sequenced each locus and noted the repeat motif and allele size range for each locus (Ballard personal communication). This information has been summarized in table 5.3.

TABLE 5.3: BASIC PCR THERMOCYCLE PARAMETERS, EXPECTED ALLELE SIZE RANGE (BASE PAIRS) AND MAGNESIUM CHLORIDE CONCENTRATIONS AS OBSERVED BY THE UTAH MARKER DEVELOPMENT GROUP (BALLARD, L., PERSONAL COMMUNICATION)

Locus	MgCl Conc. (mM)	Anneal Temp Range (°C)	Number of Cycles at Lowest Temp	Repeat Motif	Allele Size Range (base pairs)
D10S507	1.5	58-50	25	AAAT	146-190
D10S520	1.5	64-58	25	AAAG	157-210
D10S521	1.25	60-52	25	AGAT	160-192
D10S526	1.5	60-52	25	AGAT	190-260
D12S297	1.5	60-54	25	AGAT	202-270
D1S407	1.5	60-52	25	AGAT	135-160
D20S156	1.25	60-54	25	AGAT	167-220
D2S1279	1.5	62-52	25	AAAG	147-200
D2S262	1.5	64-56	25	AAAG	174-216
D3S1514	1.5	60-54	25	AAAG	202-240
D4S2285	1.5	58-50	25	AAGG	269-300
D4S2289	1.5	60-54	25	AAAG	272-315
D5S592	1.5	58-50	25	AGAT	162-200
D7S1485	1.25	62-54	25	AAGG	199-225
D7S1517	1.25	58-50	25	AAAG	173-210
D7S618	1.5	58-50	25	AAAG	141-163
D9S252	1.5	60-52	25	AGAT	214-240

The Utah Marker Development group (1995) observed that the most easily interpretable results were those loci with the (AGAT)n and the (AAGG)n repeat motifs and to a lesser extent the (AAAG)n and (AAAT)n repeat motifs. This study has incorporated seven AGAT and two AAGG motifs, also seven AAAG and one AAAT motif. All primer sets were unlabelled and 18 - 21 bases long as listed below:

TABLE 5.4: THE FORWARD AND REVERSE PRIMER SEQUENCES FOR EACH LOCUS. INFORMATION PROVIDED BY THE UTAH MARKER DEVELOPMENT GROUP (BALLARD PERSONAL COMMUNICATION).

PRIMER	FORWARD 5'-3' (Sense strand)	REVERSE 5'-3' (Anti-sense strand)	Conc. nM Forward / Reverse
D10S507	GGGTAATAAGAGCAAATCTGT	CCTTCACTAAAGCAATAAGGA	154/121
D10S520	GCACTCCAGCCTATGCAAC	GTCCTTGTGAGAAACTGGATG	127/108
D10S521	CTCCAGAGAAAACAGACCAA	CCTACCATCAATCAACTGAG	129/142
D10S526	GTGCACTAGCCAGGGTTC	TGAACATCTCAGGTAAGGGA	111/91
D12S297	GTTTGGTATTGGAGTTTTCAG	AAATCATCAGTGGAGTTAGCA	153/79
D1S407	TGCTAACCACATGGAGAGG	GGGCGGGGGATAGAAGGA	155/105
D20S156	TTCTTGGCTTGCAGCTGCA	CGGAAAAGTGCATGCACTG	156/149
D2S1279	GGCAACAAGAGCGAAACTC	GGCTTTGTGGGCTTCTAGTA	164/183
D2S262	ACCCTGCCAAATCCCCCTC	TGCGCCCCCTTTACAGAGG	103/152
D3S1514	GGCAACAGAGCAAGATGC	CCAGCCAGCAGAATTATGA	143/140
D4S2285	ATGAGCTCCTCTGAGAGG	GGAAAGAGGGCAAGACTC	81/101
D4S2289	TTGGAATATCAGATGGAAAGG	GCATGGCATTCCTATGACAC	90/90
D5S592	AGACAGACAGAGAGATTAGA	AGTAAAGTGAGTGGAGAGC	122/92
D7S1485	ACTCCAGCCTGGGTGACAC	ATGATTCTCACCAGTGGCC	91/80
D7S1517	AGCCTGATCATTACCAGGT	CTATTGGGGCCATCTTGC	86/125
D7S618	AAGACCCAGTCTCAAAGAAG	TTTCAGATGATGAAACCGATG	70/162
D9s252	ACCATGATTTGTCAACTCCTA	ACAATGAACATCCATATACCC	155/105

5.h. Primer Dilutions

It was convenient to use 1µl of a 50X concentrated solution of primer to add to the PCR reaction mix of final volume 50µl.

For example;

D9S252 was provided in the concentration 155.4 µMole and rehydrated in 1000µl sterile distilled water.

The final concentration required in a 50µl reaction volume was 0.2µM.

Therefore, it was convenient to make a solution 50X too concentrated and add 1µl of this to the PCR mix.

50 X 0.2µM = 10µM (desired concentration)

The actual stock solution was (155.4µM/10µM) 15.54 X too concentrated.

Therefore, 1µl of stock solution was added to 14.54µl sterile distilled water to give a working solution 50-fold the concentration required and 1µl of this was applied to a PCR reaction mix.

The above procedure was the same for all primer dilutions respective of their initial stock concentration.

As a preliminary measure K562 the known DNA standard for the Forensic Science Service multiplex system, was used at a concentration of 5-10ng in a 50µl PCR reaction volume to optimize the PCR protocols. The aforementioned PCR reagent mix and thermocycle programme were the optimized conditions set by the Utah Marker Development group. This 'optimized' system utilised the MJR thermocycler. This thermocycler was purported to have faster ramping speeds than the Perkin Elmer 480 used in this present study. Therefore, the 'optimized' conditions were only a guideline and starting point.

Amplification success was evaluated by the running of an agarose 'check gel' stained with ethidium bromide (see table 5.1 for reagent mix and methodology). Three microlitre aliquots of the PCR products were mixed with 2µl 'loading buffer' (Bromophenol blue with sucrose) and pipetted into the preformed wells of the submerged agarose gel. The gel was then electrophoresized for 30 minutes at 120V and viewed under ultra violet radiation.

5.i. Preliminary Optimization Results D1S407 using K562 DNA

The tetranucleotide marker D1S407 was chosen as the first marker to optimise. K562 DNA concentrations; 5ng, 10ng, 15ng, 20ng and 50ng were amplified in duplicate to observe the optimal DNA concentration required in a 50μl PCR reaction volume.

The reagent mix was as follows:

Reagent	Volume	Volume	Volume	Volume	Volume	Control	Final
	(μl)	(μl)	(μl)	(μl)	(μl)	**(No**	Conc.
	5ng	**10ng**	**15ng**	**20ng**	**50ng**	**DNA)**	In 50μl
	DNA	DNA	DNA	DNA	DNA		
ddH_20	35.25	34.75	34.25	33.75	30.75	35.75	--
Forward Primer	1	1	1	1	1	1	0.2μM
Reverse Primer	1	1	1	1	1	1	0.2μM
10x Buffer	5	5	5	5	5	5	1 x
$MgCl_2$	3	3	3	3	3	3	1.5mM
dNTP's	4	4	4	4	4	4	200
Taq	0.25	0.25	0.25	0.25	0.25	0.25	1.25 Units
DNA	**0.5**	**1**	**1.5**	**2**	**5**	**0**	
Mineral Oil	1 drop	1 drop	1 drop	1 drop	1 drop	1 drop	
Total	50	50	50	50	50	50	--

The amplification protocol was;

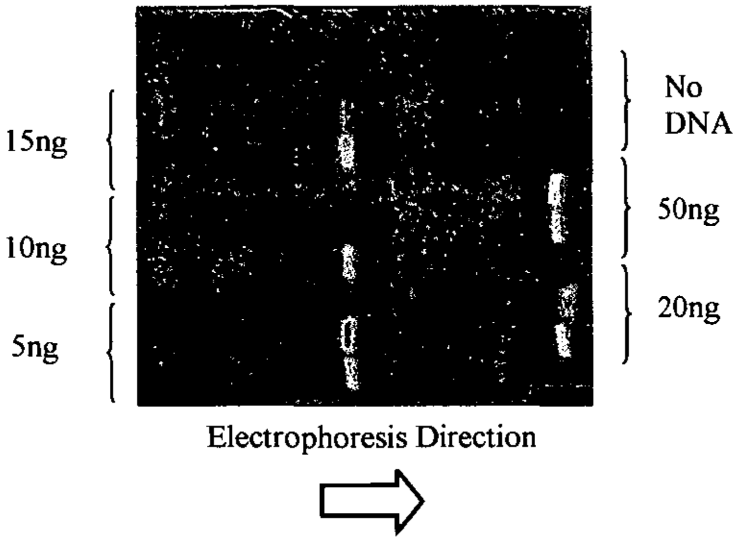
♦ 94°C	3 minutes	*Initial denaturation*
♦ 94°C	45 seconds	*Denaturation*
♦ 64°C	40 seconds	*Annealing*
♦ 94°C	100 seconds	*Extension*

The thermocycle programme was a touchdown protocol of 64-60°C, decreasing by 1°C per cycle with 25 cycles at 60°C.

A 1% agarose check-gel with ethidium bromide in tris-borate EDTA buffer following the protocol as given in this chapter. In this instance 12µl of amplified product and 3µl bromophenol blue loading buffer were pipetted into the preformed wells of the submerged check-gel. The gel was then electrophoresed at 100W for 30-45 minutes. Thereafter the gel was removed from the apparatus and photographed under ultra violet radiation.

Preliminary D1S407 Results

The photograph taken under U.V is as shown below:



The amplifications at specific DNA concentrations were carried out in duplicate to assess the robustness and efficacy of the whole system. The repeatability of the system was not consistent. One of the 10ng DNA concentration amplifications did not produce a result and two further amplifications did not produce as strong a signal as their duplicate partner. The PCR reagents for each sample were added separately to each microtube often in volumes of 1-2µl, instead of combing all reagents as a 'master mix' and pipetting an aliquot to each tube. Therefore, the differences between duplicated samples may be a consequence of a less accurate and reproducible measurement of volumes of reagent. No contamination was observed, as the negative control did not express any sign of amplification. Smearing and high molecular weight bands were seen when DNA concentrations of 15-50ng/50µl were amplified. This was not observed when the DNA concentration was 5-10ng/50µl.

Helpful information from the optimization of D1s407

- DNA concentration using this system was to be kept to 5-10ng per 50µl reaction volume.
- Aliquotting from a 'master mix' containing all reagents except DNA would provide more consistent results, in comparison to adding each reagent separately to each microtube.

5.j. Preliminary Optimization for D9S252 and D10S521 using K562 DNA

The forward and reverse annealing temperatures for the D9S252 primers were given as 58.5°C and 57.5°C respectively. Similarly, the forward and reverse annealing temperatures for D10S521 were 59.4°C and 57.1°C respectively. These temperatures were calculated by the manufacturer '*Genosys*' who custom made these oligonucleotides on the 25th September 1997.

These annealing temperatures were in good agreement with the touchdown thermocycle programme of 60-52°C, reducing by 1°C per cycle with 25 cycles at 52°C, suggested by the Utah Marker Development group (personal communication).

Therefore the above parameters were used for the thermocycle protocol and the reagent mix for PCR was the same as D1S407 with K562 DNA at a concentration of 10ng in a 50µl reaction volume.

Results

A good amplification product was observed using the D9S252 primer set, and there was no high molecular weight smearing. Therefore, the amplification protocol for this locus was kept the same for further amplifications. However, no amplification was observed at locus D10S521 under the same conditions, thus the experiment was repeated.

5.k. Preliminary Optimization for D4S2289, D7S1485, D12s297 and second amplification for D10S521 using K562 DNA

The forward and reverse annealing temperatures of the primers as given by the manufacturers '*Genosys*' were as follows;

Locus	Forward Primer Annealing Temp. (°C)	Reverse Primer Annealing Temp. (°C)
D4S2289	60.3	62.9
D7S1485	65.4	62.6
D12s297	57.6	58.8

The following touchdown protocol was used:

- 94°C 3 minutes *Initial denaturation*
- 94°C 45 seconds *Denaturation*
- 60°C 40 seconds *Annealing*
- 94°C 100 seconds *Extension*

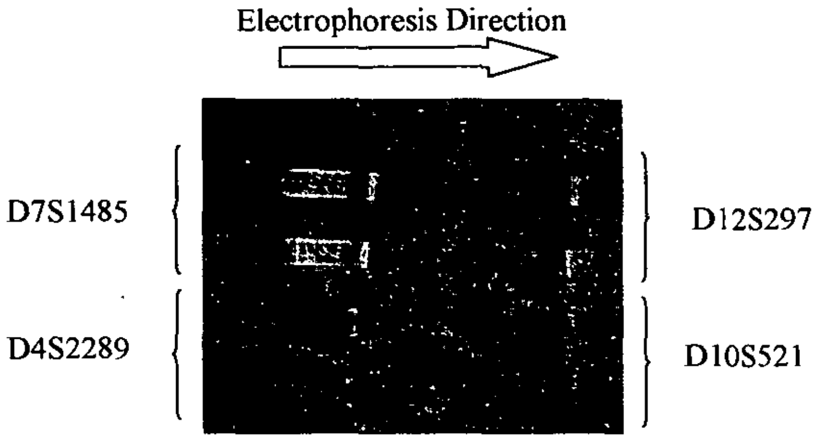
The annealing temperature was lowered 1°C per cycle between 60-52°C and with 25 cycles at 52°C. Locus D10S521 was also repeated. The PCR reagent mix was kept the same.

The efficacy of the amplification at these loci was checked using a 1% agarose 'check-gel' stained with ethidium bromide.

Results

The photograph of the stained check-gel highlighted a variation of amplification efficacy. Both the D4S2289 and D12S297 loci amplified with no high molecular weight smearing, although the D4S2289 amplification was considerably weaker. The D7S1485 locus amplified producing very noticeable extraneous high molecular weight bands. It was observed that the touchdown protocol of 60-52°C may have been too low for the D7S1485 locus. In particular the annealing temperatures for the primers were 65.4°C and 62.6°C for forward and reverse respectively, hence may have required a touchdown thermocycle programme ranging 65-57°C.

Again, the D10S521 locus did not amplify using the aforementioned PCR conditions, see below;



5.1. Preliminary Optimization for D4S2285, D7S618 and D10S507 using K562 DNA

The forward and reverse annealing temperatures of the primers as given by the manufacturers '*Genosys*' were as follows;

Locus	Forward Primer Annealing Temp. (°C)	Reverse Primer Annealing Temp. (°C)
D4S2285	56.9	58.9
D7S618	57.5	62.4
D10S507	56.4	57.6

The touchdown thermocycle programme was lowered from 60-52°C to 58-50°C. To keep the continuity between these amplifications and those already carried out, the PCR reagent concentrations were unaltered. In accordance with previous amplifications the efficacy of the PCR product was observed using the 1% agarose check-gel electrophoresis methodology (see table 5.1 for details).

Results

Amplification produced clear bands with no high molecular weight smearing at loci D10S507 and D7S618. However, D4S2285 and D10S521 did not show any signs of amplification with the above protocol, so further optimizationof these loci was carried out using the protocol below.

5.m. Repeat Preliminary Optimization for D10S521 and D4S2285 using K562 DNA

The touchdown protocol was altered, to accommodate annealing temperatures in the range 60-50°C with 25 cycles at 50°C. The differences in buffer types were also examined. The 'Advanced Biotechnologies Formula' and also 'Promega's Formula' were used. The differences between these two formulae were as listed below:

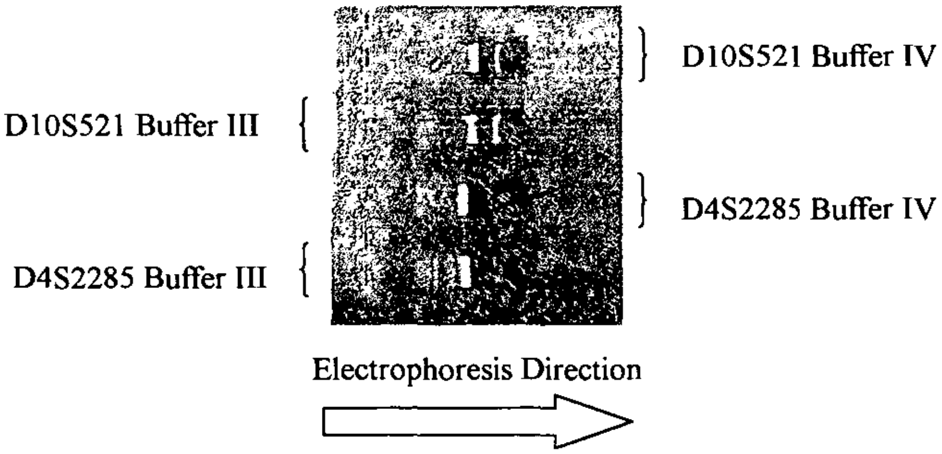
	Buffer III (Promega Formula)		Buffer IV (Advanced Biotechnologies Formula)	
Vial 1	1.25ml of:500mM	KCl	1.25ml of: 200mM	$(NH_4)_2SO_4$
	100mM	Tris-HCl, pH8.8 at 25°C	750mM	Tris-HCl, pH8.8 at 25°C
			0.1%(v/v)	Tween 20
Vial 2	1.25ml of: 25mM	$MgCl_2$	1.25ml of: 25mM	$MgCl_2$
Vial 3	1.25ml of: 2%(v/v)	Triton X-100		

The differences are mainly concentrations and the use of different salts. Tween 20 and triton X-100 are both are supposed reagents to stabilise the PCR reaction. Tween 20 has been unconditionally included in the Advanced Biotechnologies formula, but the Triton X-100 was optional and included as a separate vial. The magnesium chloride was also provided in a separate vial to allow one to deliver different concentrations to PCR reagent mixes. Buffer III without Triton X-100 and Buffer IV with the same concentrations of $MgCl_2$ as before (1.5mM) were used. All other reagent concentrations were the same as previous amplifications.

Results

Both loci amplified using the revised touchdown protocol. Artifacts were observed at the D10S521 locus appearing as feint bands of lower molecular weight (see photograph below). The artifacts may have been the result of too low an annealing temperature at this locus and the primers attaching to an unspecific region of DNA. One should observe that the annealing temperatures of the forward and reverse primers were within the touchdown temperature range of the thermocycle programme. However, the heating block, as demonstrated, would not have reached the target temperature, which in effect lowered the 'working' touchdown temperature range to below the desired optimum for the primers.

The amplifications at D4S2285 produced a large amount of PCR product. High molecular weight artifacts were observed using buffer III, which were not observed using buffer IV. Therefore, the touchdown thermocycle programme of 60-50°C together with buffer IV in the PCR reagent mix were considered suitable for the D4S2285 locus.



5.n. Preliminary Optimization for D3S1514, D2S1279 and D2S262 using K562 DNA

The forward and reverse annealing temperatures for these loci were calculated by the manufacturers '*Genosys*' and are listed below:

Locus	Forward Primer Annealing Temp. (°C)	Reverse Primer Annealing Temp. (°C)
D3S1514	62.2	61.4
D2S1279	62.5	60.8
D2S262	69.4	69.4

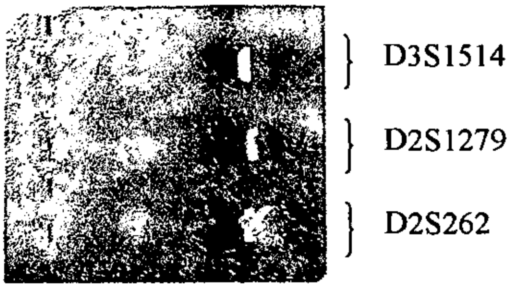
The PCR reagent mix was as follows:

Reagent	Volume (µl)	Final Conc.
Sterile distilled Water	35.55	--
Forward Primer	1	0.2µM
Reverse Primer	1	0.2µM
10 X Buffer (IV)	5	1X
$MgCl_2$	3	1.5mM
dNTP's	4	200
DNA	0.2	~10ng
Taq	0.25	1.25 units
Mineral Oil Overlay	1 drop	--
Total	50	

The thermocycle programme was the standard touchdown protocol as used before, with annealing temperature range of 64-54°C with 25 cycles at 54°C.

Results

Clear bands were observed at the D3S1514 and D2S1279 loci with no high molecular weight smearing or extraneous bands. However, a less definite band was observed at the amplified D2S262 locus. The amplification thermocycle program may have been too low for the primers, causing mis-priming during the annealing phase. Therefore, a touchdown protocol incorporating a higher annealing temperature range was required.



5.o. Preliminary Optimization for D5S592, D7S1517, D10S526, D20S156 and D10S520 Using K562 DNA

The forward and reverse annealing temperatures for these loci were calculated by the manufacturers '*Genosys*':

Locus	Forward Primer	Reverse Primer
	Annealing Temp. (°C)	Annealing Temp. (°C)
D5S592	49.9	53.2
D7S1517	58.5	63.0
D10S526	60.9	59.6
D20S156	68.3	65.3
D10S520	63.4	61.1

The Utah Marker Development group also found that at loci D7S1517and D20S156 1.25mM magnesium chloride was optimal for the PCR, instead of the 'standard' 1.5mM concentration included in the reaction mix of all the other loci.

Therefore the reagent mixes differed as follows:

	Reagent mix for loci: D5S592, D10S526 and D10S520		Reagent mix for loci: D7S1517 and D20S156	
Reagent	Volume (μl)	Final Conc.	Volume (μl)	Final Conc.
Sterile distilled Water	35.55	--	36.05	--
Forward Primer	1	0.2μM	1	0.2μM
Reverse Primer	1	0.2μM	1	0.2μM
10 X Buffer (IV)	5	1X	5	1X
$MgCl_2$	**3**	**1.5mM**	**2.5**	**1.25mM**
dNTP's	4	200	4	200
DNA	0.2	~10ng	0.2	~10ng
Taq	0.25	1.25 units	0.25	1.25 units
Mineral Oil Overlay	1 drop	--	1 drop	--
Total	50		50	

A touchdown thermocycle protocol with annealing temperature range of 64-54°C was tried. This appeased most primer annealing temperatures, although a little low D20S156.

Each locus was amplified in duplicate. Aliquots of the amplified product were electrophoresied through a 1% agarose check gel stained with ethidium bromide.

Results

All loci amplified clearly with little high molecular weight smearing. Therefore the PCR reagent mix and thermocycle programme was to be kept the same for subsequent amplifications at these loci.

A total of nine months had passed before the preliminary optimization had been completed using the K562 DNA standard. The adjustments and alterations to the Utah Marker Development Groups 'optimized' protocol were carried out to accommodate the differences in equipment and reagents between their group and this study.

Section III: Preliminary Analyses Using Maori and Polynesian DNA Samples

5.p. 1st Batch of PCR Amplifications using Maori and Polynesian DNA samples

As a preliminary measure and to check the reproducibility of the PCR protocol, 5 DNA samples in duplicate were amplified using the primer set D3S1514.
5ng sample DNA was used in the first instance.

PCR Reagent Mix:

Reagent	Volume (µl)	Final Conc.
Sterile distilled Water	33.75	--
Forward Primer	1	0.2µM
Reverse Primer	1	0.2µM
10 X Buffer (IV)	5	1X
$MgCl_2$	3	1.5mM
dNTP's	4	200
DNA	2	5ng-10ng
Taq	0.25	1.25 units
Mineral Oil Overlay	1 drop	--
Total	50	

Thermocycle programme:

- ♦ 94°C 3 minutes *Initial denaturation*
- ♦ 94°C 45 seconds *Denaturation*
- ♦ 64°C 40 seconds *Annealing*
- ♦ 94°C 100 seconds *Extension*

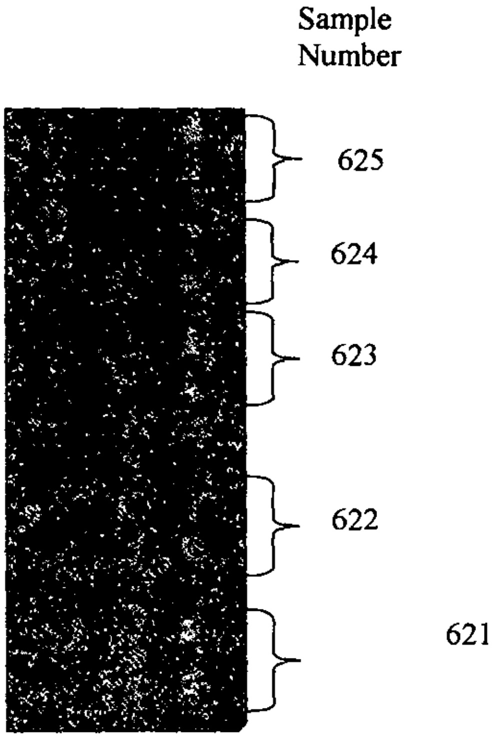
A touchdown protocol of 64 - 54°C, decreasing by 1°C per cycle with 25 cycles at 54°C.

The starting point for this locus was 4°C higher than recommended to ensure high specificity with respect to primer annealing.

A total of 5µl of each amplified product was mixed with 2µl loading buffer and pipetted into the preformed wells of a 1% agarose check gel stained with ethidium bromide (refer to table 5.1). The gel was then electrophoresed for 30 minutes at 120V.

Check gel results

Of the five duplicate amplifications, just one of sample 622 did not show any signs of successful PCR. Sample 621 was observed to have high molecular weight artifacts, possibly caused by too much DNA in the reagent mixture, see below;



The remaining Maori and Polynesian DNA samples were amplified in duplicate following the same protocol as above. A total of 175 Maori and Polynesian DNA samples, or 350 amplifications of which just 4 samples did not amplify at all. These four samples were numbers: 450, 476, 504 and 602. Sample number 602 was observed to have a very low stock DNA concentration (see appendix A for sample concentrations). Interestingly, sample number 467 was observed to have a negligible stock DNA concentration. However, this sample amplified producing a clear band when electrophoresed through an agarose check gel. This indicated that the DNA concentration fell below the detectable threshold of ethidium bromide when quantifying the sample DNA, i.e. less than 1-2ng. However, this was still an acceptable concentration for the sensitive amplification procedures.

5.q. Preliminary Polyacrylamide Gel Electrophoresis and Silver Staining of PCR Products Amplified at the D3S1514 Locus

Initially, polyacrylamide gel electrophoresis (PAGE) was the method of choice to size the amplified alleles. The aim of preliminary PAGE and silver staining was to familiarise oneself with the equipment and protocol.

The PCR products of sample numbers 615, 613 and K562 were electrophoresed following the standard denaturing polyacrylamide gel electrophoresis protocol and silver staining technique. Two allelic ladders were included with the electrophoresis, not only to size the alleles but to observe the relationship between allele size and time required to migrate a specific distance under specific electrophoresis conditions.

5.q.i. Polyacrylamide Gel Preparation and Electrophoresis:

Solutions required:

- 40% acrylamide:bis (19:1) and TEMED
- 0.5% acetic acid in 95% ethanol
- 10X TBE Buffer (Tris-borate EDTA buffer)
- 10% ammonium persulfate

Method

- The glass plate that the gel will be in contact with was etched to distinguish the treated plates. The plates were cleaned with 95% ethanol.
- 3ml of REPELCOTE Vs was evenly spread onto the etched side of the longer plate.
- The repelcote was allowed to dry and the excess removed with a paper towel saturated with deionized water. The plate was dried with a paper towel.
- 3µl of BIND SILANE was added to 1ml of 0.5% acetic acid in 95% ethanol in a 1.5ml microcentrifuge tube and applied to the shorter glass plate wiping over the entire surface area.
- 5 minutes was allowed for the binding solution to dry. To remove excess binding solution the plate was wiped 3-4 times with IMS (95% ethanol).
- **The treated surfaces were not allowed to touch each other.** The glass plates were assembled with 0.4mm side spacers and a bottom spacer and clamped in place.

- Composition of a 4% or 6% acrylamide solution (total of 75ml)

Reagent	4%	6%	Final Conc
Urea	31.50g	31.50g	7M
Deionized Water	40.00ml	36.25ml	-
10X TBE	3.75ml	3.75ml	0.5X
40% acylamide:bis	7.5ml	11.25ml	4% or 6%
Total	75ml	75ml	

- The acrylamide solution was filtered through a 0.2 micron filter.
- 50μl of TEMED and 500μl of 10% ammonium persulfate was added to the acrylamide solution and gently mixed.
- Using a 50ml syringe the acrylamide solution was poured between the glass plates. To prevent bubble formation, pouring was started at one side of the assembled plates and a constant flow of solution was maintained.
- The gel was rested in the horizontal position and the 0.35mm toothed comb was inserted.
- Polymerisation was allowed to proceed overnight.

  The remaining acrylamide solution was used as a polymerisation control.

POLYACRYLAMIDE GEL ELECTROPHORESIS

Using the *Poker Face II Nucleic Acid Sequencer SE 1650 (Hoefer Scientific Instruments)*

- The clamps were removed and the glass plates cleaned using damp paper towels.

  A foam tab was attached to the upper end of each spacer, overlapping the top edge of the shorter plate.

  A thin layer of celloseal grease was applied to all sealing rubber gaskets and the glass plate sandwich with the shorter glass plate facing towards the buffer chamber was clamped in place. 1 litre of 0.5X TBE was then added to the bottom chamber of the electrophoresis apparatus.
- The glass plates plates were then secured in place.

- Roughly 800ml of 0.5X TBE was added to the top buffer chamber of the electrophoresis apparatus.
- With the buffer covering the 30cm long gel the comb was removed to allow buffer to fill the wells.
- A few microlitres of loading dye was added to some of the wells and the upper electrode / lid assembly attached to the upper buffer chamber.
- The gel was pre run to attain a surface area of 50°C.

(For gel running conditions usually 50 watts for the SE1650 apparatus).

- The PCR samples were prepared by mixing 2.5µl of each sample with 2.5µl of 2X Loading Solution.
- Alternating CTT and FFv DNA markers were loaded into separate lanes of the gel.
- The DNA samples were denatured just prior to loading by heating to 95°C for 2 minutes and then were immediately chilled on ice.
- 3µl of each sample was then loaded into the respective cells using feathered micropipette tips.
- The gel was then run as before set at 50 Watts and run until the xylene cyanol has migrated to the base of the gel.
- The glass plate sandwich was disassembled and the glass plates prised apart using a plastic wedge.

5.q.ii. SILVER STAINING Procedure (Promega)

Reagents

Bind Silane

Silver nitrate

Formaldehyde (37%)

Sodium Thiosulfate

Fix / Stop solution (10% Acetic acid)

Staining solution (silver nitrate 37% Formaldehyde and Deionized water)

Developer solution (37% Formaldehyde, Sodium thiosulfate, deionized water and sodium carbonate)

Method:

- The gel was placed in a shallow plastic tray and the following steps were carried out:

Step	**Solution**	**Time**
a	Fix/stop	20 mins
b	Deionized water	2 mins
c	Repeat b twice2X2 mins	
d	Staining solution	30 mins
e	Deionized water	10 seconds
f	Developer solution	2-5 minutes (until alleles and ladders are visible)
g	Fix/stop solution†	5 mins
h	Deionized water	2 minutes

† Added directly to developer solution to stop developing the reaction.

- The gel was allowed to dry overnight and then photographed.

5.q.iii. Preliminary results

In figure 5.1, lanes 2, 4, 6 the TPOX and THO1 ladders ran together and in lane 14, the FES/FPS ladder. Lanes 1,3,5,7 and 15 were kept empty, in case of 'over spill' from adjacent wells. The 'over spill' was observed and the allelic ladders although fainter were clearly defined. The amplified DNA samples K562, 615 and 613 were 'run' in lanes 8, 10 and 12 respectively. The bands that were observed in lanes 7, 9 and 11 were the 'over spill' from the adjacent lanes.

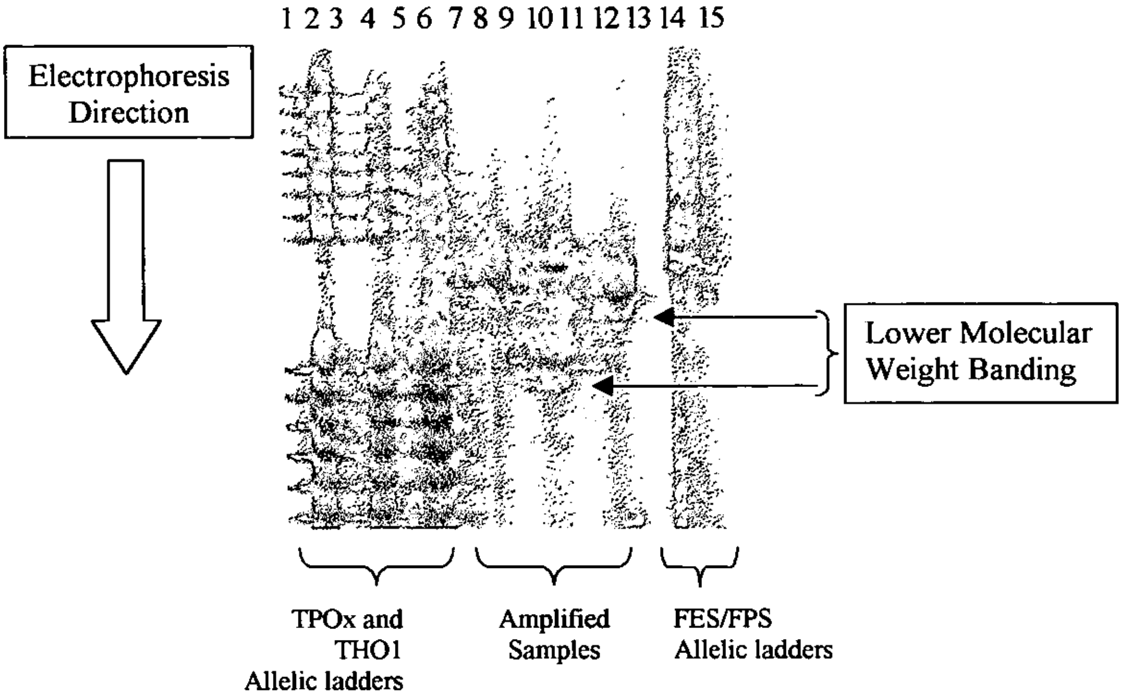

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Electrophoresis Direction
Lower Molecular Weight Banding
TPOx and THO1 Allelic ladders
Amplified Samples
FES/FPS Allelic ladders

FIGURE 5.1: SILVER STAINED POLYACRYLAMIDE GEL OF AMPLIFIED SAMPLES AND THE COMMERCIAL MULTIPLEX ALLELIC LADDERS FES AND FPS.

Pitfalls and Reasons:

Preliminary findings indicated that at least four bands instead of two were present in the heterozygous condition and similarly two bands instead of one in the homozygous condition (see figure 5.1).

Reasons:

- Too great a volume of amplified product loaded into the wells of the gel may have forced the sample to 'run' in two or more consecutive waves, causing the 'ghosting' effect of the additional bands. However, this theory can be discounted, as the additional bands were not observed with the allelic ladders which when loaded were of equal volume as the amplified product.

- The primers were not specific to one locus hence, another area of DNA was amplified which when the PCR product was electrophoresed a second band pattern was observed. However, this too was highly unlikely. The additional bands were mirror images of the prominent band. If the primers were annealing to two distinct loci it would be questionable to observe the alleles of the second locus to have the same base pair size difference to alleles of the first locus.

A review of the observations, led to the summation that if the primers were not annealing at two specific loci and the electrophoresis conditions were correct, then the actual thermocycle programme was not stringent enough. Therefore, the annealing temperature range was raised to increase specificity whilst all other conditions were kept the same.

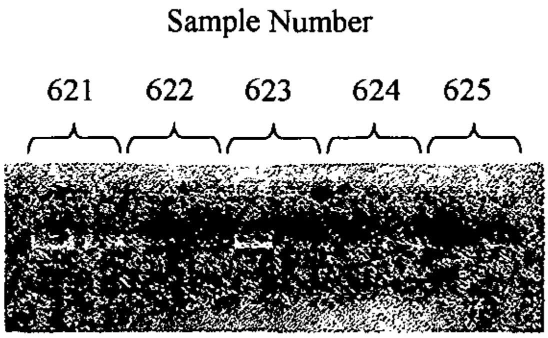
5.r. 1st Revision of the Touchdown Protocol to Remove Extraneous Bands

The annealing temperatures of the touchdown protocol were raised from 64 - 54°C to 64 - 58°C to increase specificity and decrease any spurious banding. All other PCR reaction conditions were the same as the previous experiment.

Check Gel Results

All the samples (except one of the 625 duplicates) amplified well, with little or no high molecular weight smearing. All the PCR products were then run on a polyacrylamide gel.

The polyacrylamide gel electrophoresis conditions were as before.



Results

The silver stained polyacrylamide gel, as shown in figure 5.2 exhibited similar attributes to the previous polyacrylamide gel (see figure 5.1). The PCR products were loaded onto the polyacrylamide gel in the same order as the agarose check gel. The lane numbers and their identification are given in table 5.5.

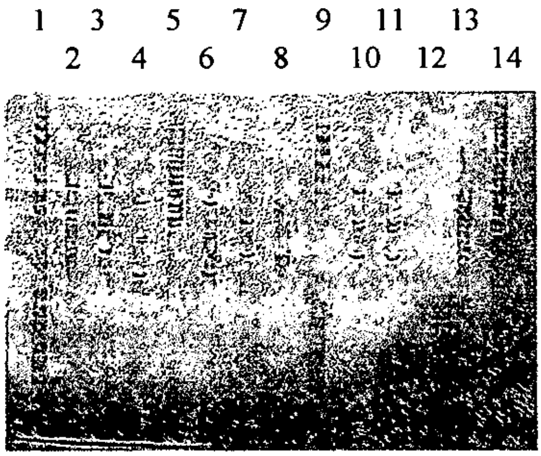


FIGURE 5.2: SILVER STAINED POLYACRYLAMIDE GEL OF THE 'OPTIMIZED' PCR PROTOCOL PCR PRODUCTS AT THE D3S1514 LOCUS. IDENTIFICATION OF THE PRODUCT IN EACH LANE IS GIVEN IN TABLE 5.5 BELOW.

TABLE 5.5: IDENTIFICATION OF EACH PRODUCT OF THE SILVER STAINED POLYACRYLAMIDE GEL FIGURE 5.2 ABOVE.

Lane Number	Identification
1	CTT allelic ladder see appendix for composition
2	PCR product @ D3S1514 for sample number 621
3	Duplicate of Lane 2
4	PCR product @ D3S1514 for sample number 622
5	FFV allelic ladder see appendix for composition
6	Duplicate of Lane 4
7	PCR product @ D3S1514 for sample number 623
8	Duplicate of Lane 7
9	CTT allelic ladder see appendix for composition
10	PCR product @ D3S1514 for sample number 624
11	Duplicate of Lane 10
12	PCR product @ D3S1514 for sample number 625
13	Duplicate of Lane 12
14	FFV allelic ladder see appendix for composition

There was very good resolution of the amplified products although the allelic ladders were less well defined. Similarly to the first polyacrylamide electrophoresis four bands were observed in the heterozygous condition. Only two bands were observed in lanes 7 and 8 corresponding to sample number

623. Either this was the heterozygous state without the extraneous bands or more likely sample number 623 was homozygous at the D3S1514 locus. Further to support the homozygous theory, the distance between the two bands was similar to the distances between the 'mirror images' of the additional bands observed in the heterozygous condition.

In accordance with Promega's PCR Primer manual, if multiple products or a high molecular weight smear is observed, the annealing temperature should be raised and cycles from the bottom end of the temperature range. In addition, the number of cycles per degree should be raised from one to three.

5.s. 2nd Revision of the Touchdown Protocol to Remove Extraneous Bands

The annealing temperature range of the touchdown protocol was raised from 64 - 58°C to 64 - 61°C with 3 cycles per degree and 18 cycles at 61°C. The PCR reagent mix was the same as the 1st revision of the protocol.

An 1% agarose check gel stained with ethidium bromide (refer to table 5.1) was run to check amplification.

Results:

No amplification was observed.

5.t. 3rd Revision of the Touchdown Protocol to Remove Extraneous Bands

The annealing temperature range of the touchdown protocol was altered from 64-61°C to 64 - 60°C, again with 3 cycles per degree and 18 cycles at 60°C. As before the PCR reagent mix was kept the same.

Results:

Again, no amplification was noted after running a check gel.

5.u. 4th Revision of the Touchdown Protocol to Remove Extraneous Bands

The annealing temperature range was lowered again from 64-60°C to 64-59°C, this time with 17 cycles at 59°C.

Results:

However, no amplification was observed.

5.v. 5th Revision of the Touchdown Protocol to Remove Extraneous Bands

Sample number '615' was amplified in duplicate. The touchdown protocol was lowered from 64-59°C to 64-58°C with two cycles per degree and 25 cycles at 58°C, a total of 38 cycles.

Again, a 1% agarose check gel stained with ethidium bromide was used to run the amplified products. On this occasion, the check gel indicated good amplification with no smearing or obvious extraneous bands (see figure 5.3).

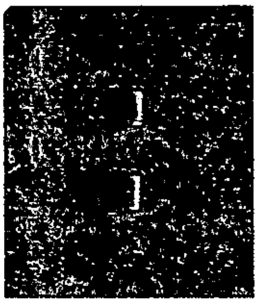


FIGURE 5.3: SAMPLE NUMBER 615 AMPLIFIED IN DUPLICATE. PCR REAGENT CONCENTRATIONS AND THERMOCYCLE PROGRAMME WERE AS LISTED IN THE 5TH REVISION OF THE D3S1514 PROTOCOL

5.v.i. Polyacrylamide Gel Electrophoresis of the fifth revision of the touchdown protocol.

A 6% polyacrylamide gel was run at a higher wattage (50Watts) using the successful DNA amplification from the fifth revision. Amplified DNA samples from the 1st polyacrylamide gel 'run' were included as a comparison (see figure 5.2). Extraneous band distribution between loci was also compared using the D4S2285 and D3S1514 loci.

Extraneous bands were observed for each amplified including at the D4S2289 locus and protocol five (D3S1514) that had been painstakingly revised five times.

5.v.ii. Comparison of polyacrylamide gel results:

It was observed that running the polyacrylamide gels at a higher wattage (50+ Watts), produced sharper clearer bands than at a lower wattage (<30 Watts). Unfortunately, the *Hoefer 1600* electrophoresis equipment had no thermostatic control. Therefore, heat transfer across the polyacrylamide gel was not constant and electrophoresed samples ran at a faster rate in the hotter central portion of the gel and slower at either side of the gel. This was visually observed as a 'smiling effect'. Electrophoresing a lower concentration of PCR product also produced clearer results. However, four bands were still observed.

Why the extraneous banding in a supposed optimised system?

Leicester Universitys' Genetic department was contacted. The genetic department was questioned about the consistent four bands with heterozygotes and two bands with homozygotes. It was suggested that one might expect to detect all 4 strands of DNA from the 2 alleles as silver staining, unlike end-labeling, does not discriminate between the strands of a single molecule. This would explain why this effect was observed regardless of locus, PCR conditions and quantity of product loaded onto the polyacrylamide gel. It would also explain why 'mirror images' have been observed as there would have been an inevitable strand bias in base composition hence mobility. Therefore one set of bands probably corresponds to the 5'-3' strand and the other set to the 3'-5' strand (May,C., personal communication).

Reexamining the photographs of the silver stained polyacrylamide gels, the aforementioned complementary strand bias mobility theory was plausible.

5.w. Base Composition and Mobility

In the year 1972, the effect of base composition in the electrophoresis of Crab DNA in polyacrylamide gels was investigated (Zeiger et al. 1972). It was observed that DNA with progressively greater G-C base contents had progressively greater mobilities in polyacrylamide gels (Zeiger et al. 1972). However, it was suggested that the base composition could affect the secondary structure of the DNA, hence altering mobility. The double helical stability was correlated with G-C content, as G-C base pairs have a higher bonding potential with three hydrogen bonds in comparison to two of the A-T base pair (Zeiger et al. 1972). Furthermore, secondary structures of A-T rich oligonucleotides differed to those with a higher proportion of G-C bases (Zeiger et al. 1972).

The effects of secondary structure on migration time were investigated by Satow et al. (1993). Their expectation that oligonucleotides with the same number of all four bases but in different sequences would migrate at the same rate was incorrect. In the presence of a denaturing polyacrylamide gel, most sequences were separated from each other (Satow et al. 1993). The results indicated a decrease in migration time with an increase in G and C base pairings within the single stranded oligonucleotide (Satow et al. 1993).

The differences in migration rate of oligonucleotides of the same length but different base composition implies that precise sizing of alleles of indeterminate sequence would be difficult. This is because commercial allelic ladders with a different base composition would migrate at a different rate (Satow et al. 1993).

SECTION IV: A NEW SUBMARINE GEL ELECTROPHORESIS SYSTEM:
ELCHROMS' SPREADEX GELS'

On the 6th January 1998 a new submarine gel electrophoresis system was introduced by VH Bio.

The pre made '*Spreadex*' DNA gels were characterized by a DNA exclusion limit, whereby any molecules above a particular size were excluded and unable to migrate through the gel.

The resolving power of the spreadex gels was two times greater than polyacrylamide and capable of resolving four base pair differences in the 100 - 300 base pair range in a gel shorter than 4 cm. Oligonucleotide size estimation was also more precise than conventional polyacrylamide gels, as sequence dependent mobility's were largely eliminated at 55°C. Recovery of PCR fragments can be achieved by merely excising the desired band from the gel and used directly for re-amplification.

Apparatus:

Submarine gel tank

Power supply

Forceps

Shallow tray for staining (12'X 5')

Rotashaker

UVP Image store 7500

Reagents:

30mM TAE buffer (Running buffer) (See appendix for buffer composition)

Sterilised distilled water

Ethidium bromide

Loading buffer (5X Bromophenol blue and xylene cyanol)

5.x. Preparation of precast gels and DNA samples for electrophoresis:

The precast 'spreadex' gels were available in varying sizes and for varying optimal base pair range for the best resolution. The 'EL600' spreadex gel was chosen which had an optimal resolution range of 150-300 base pairs.

Prior to running the EL600 spreadex gel at 55°C it was equilibrated by placing the gel in its protective wrapping on the lid of the electrophoresis apparatus for 10 minutes, whilst running a current through the buffer in the gel tank to heat it.

The DNA samples were prepared by mixing 3-4 parts of sample with 1 part of elchrom Scientific's loading buffer. The CTT and FFv allelic ladders were used to compare the unknown band sizes to. 1μl of the ladder was mixed with 2μl loading buffer for use with the gel (for a composition of the allele sizes of the ladders in base pairs see appendix).

5.y. Electrophoresis of gels without temperature control in a non-circulating submarine electrophoresis apparatus

Elchrom purported that the spreadex gels could be used in conventional non-circulating submarine electrophoresis tanks although the following may occur:

- Gel may become warmer at the edges than the middle, thus as observed with PAGE, migration across the gel differed and the 'smiling effect' would visible.
- The reproducibility of runs is low unless room temperature and running buffer temperature are identical.

At first, the spreadex gels were run in a non-circulating *Hybaid* submarine electrophoresis tank. A total of 2-3 liters of 30mM running buffer was added to the submarine electrophoresis apparatus and the gel removed from the protective covering and submerged into the apparatus. A 'catamaran' was specifically designed to hold the submerged gel in place in the buffer. This device pinned down the sides of the gel, without interfering with the flow of electric current across the gel.

The PCR products (3μl) were mixed with loading buffer (2μl) and were pipetted sequentially into the wells of the submerged preformed spreadex gel.

Six PCR products and two allelic ladders were provisionally used with the sample spreadex gel. The products were loaded onto the gel in the order:

TABLE 5.6: A MIX OF PCR PRODUCTS PREVIOUSLY RUN ON DENATURING POLYACRYLAMIDE GELS WERE CHOSEN FOR SPREADEX GEL ELECTROPHORESIS.

Lane Number	Identification
1	1st amplification of 615 @ D3S1514 (re: figure 5.2)
2	PCR amplification @ D4S2285, sample: K562
3	PCR amplification @ D3S1514, sample: 623
4	CTT allelic ladder (see appendix for composition)
5	PCR amplification @ D4S2285, sample 615
6	5th Revised amplification of 615 @ D3S1514
7	FFv allelic ladder (see appendix composition)
8	PCR amplification @ D3S1514, sample K562

The voltage was set to 300 (equivalent to 10V/cm between electrodes) and run for a total of 4 hours 15 minutes.

It was observed that as the temperature of the solution increased the voltage decreased to a minimum of 150 V.

Andrews (1988), observed that the more concentrated the buffer the lower the electrical resistance, thus the higher the current at a particular voltage, hence the greater the heat generated. Similarly, this study observed that following the relationship 'V = I X R', as the heat increased, voltage reduced to compensate for the lower resistance hence keeping the ampage the same.

The temperature of the buffer was raised from room temperature to over 50°C during the 4 hours 15 minutes running time.

After electrophoresis had been completed the gel was removed from the apparatus using the forceps provided and the gel removed from its plastic backing by running a nylon thread between the plastic backing and the gel.

The gel was then placed face down into a shallow tray for staining with 2μl ethidium bromide (10mg/ml concentration) in 50ml sterile distilled water.

The tray was left to gently rock on a rotashaker to evenly stain the gel for 45 minutes.

The staining solution was carefully discarded and destaining of the gel was then carried out for 30 minutes using sterile distilled water only.

The bands were visualized under ultra violet irradiation and photograph taken.

5.y.i. Preliminary Results

Sample number 623 and K562 DNA (lanes 3 and 8) both were homozygous at the D3S1514 locus (figure 5.4). No additional extraneous bands or high molecular weight smearing was observed. However, the heterozygotes did show fainter extraneous bands typically of a higher molecular weight in comparison to the two distinctive brighter bands. These observations contradicted the results of the polyacrylamide gels using the exact same samples. Furthermore, the FFv and CTT ladders were not at all clear enough to enable any sizing of the D3S1514 and D4S2285 alleles.

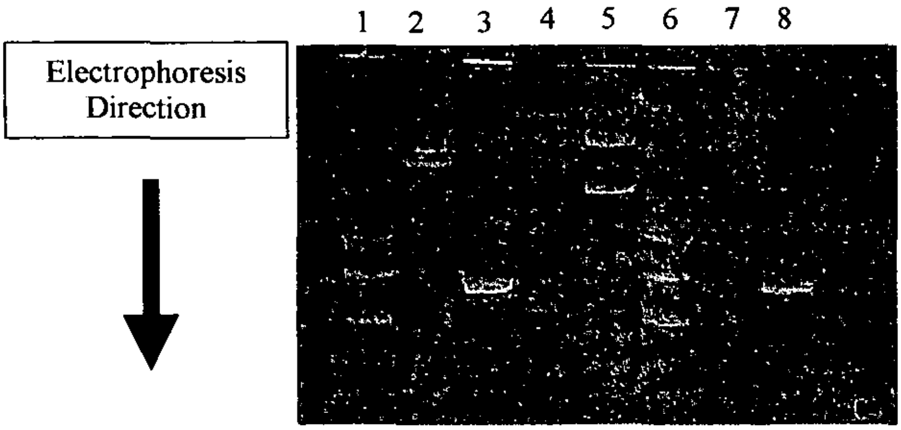


FIGURE 5.4: EL600 SPREADEX GEL STAINED WITH ETHIDIUM BROMIDE. LANE NUMBERS AND SAMPLE IDENTITY CORRESPOND TO THOSE GIVEN IN TABLE 5.6

The same samples were then re-analysed using a 6% denaturing polyacrylamide gel.

A pre run was set up for 30 minutes. A different power supply was used and the running conditions were 40W, 320V and 125mA.

The actual running settings after pre heating the gel were 2500V, 60 Watts and 2-3mA. The voltage was very high as this was the only way 60 Watts could be achieved, as the mAmpage was so low.

Silver staining of the polyacrylamide gel expressed two bands for the homozygous condition and four bands for the heterozygous condition.

Comparing the results of the spreadex gel and those of the polyacrylamide gel one could presume that indeed the two bands of the homozygous condition were the forward and reverse strands of the homozygous allele.

The fainter extraneous bands observed in the heterozygous condition when run through the spreadex gel (see figure 5.4) cannot be attributable to the forward and reverse DNA strands as the DNA is kept in its double stranded conformation. However, the fainter bands appear to be of higher molecular weight than the true bands.

5.z. The Heteroduplex Theory

The spreadex gels electrophorese double stranded DNA, therefore one would not expect to observe the complementary strands separately. However, the amplification of heterozygotes tended to 'amplify' additional oligonucleotide bands, not observed with homozygotes. The assumption was that synthesised oligonucleotides of the heterozygote were forming a heteroduplex. In other words, the longer allele paired to the shorter allele causing a conformational change of the double strand retarding the mobility of this molecule through the gel. If this was true, then purposefully mixing two sample DNAs of known homozygosity (although different allele sizes) and amplifying the mixed DNA sample, one would observe three bands instead of two. The theory was put into experimental procedure. Sample numbers 528 and 524 were both homozygous at the D10S520 locus. Amplification conditions were as previously described. The amplification was carried out in duplicate and the PCR products electrophoresed for 1 hour 15 minutes through an EL600 spreadex gel.

5.z.i. Results

Ethidium bromide stained the remainder of the oligonucleotide that did not electrophorese in the well of the spreadex gel.

The two homozygous alleles were clearly amplified and in addition, another band was clearly defined in one of the duplicate amplifications and fainter in the other (figure 5.5). The additional band was therefore a natured double stranded oligonucleotide, with a sense-strand from one allele and the complementary or antisense-strand from the other allele. Hence, the conformation of the heteroduplex would differ to the homozygous alleles.

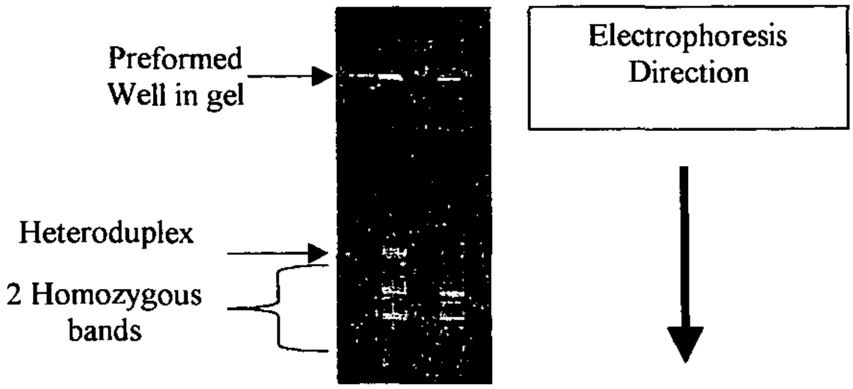


FIGURE 5.5: SPREADEX GEL ELECTROPHORESIS OF AN AMPLIFIED MIXTURE OF TWO KNOWN HOMOZYGOUS DNA SAMPLES IN DUPLICATE (NUMBERS 528 AND 524) AT THE D10S520 LOCUS

To ultimately confirm the heteroduplex theory, sequencing of all the 'bands' of a heterozygous sample should not reveal any sequence difference between them. In particular, sequencing of both the 3'-5' and 5'-3' strands of the purported heteroduplex, one would observe sequence homology to the amplified locus, albeit one strand would have a greater number of base pairs in comparison to its complementary strand.

5.z.ii. Sequencing 'heteroduplex' bands

AltaBio were chosen to carry out the sequencing due to their cheap, fast and reliable service.

A selection of homozygous and heterozygous samples were re-run on a spreadex gel, optimised for the greatest separation of the bands (see figure 5.6 and table 5.7).

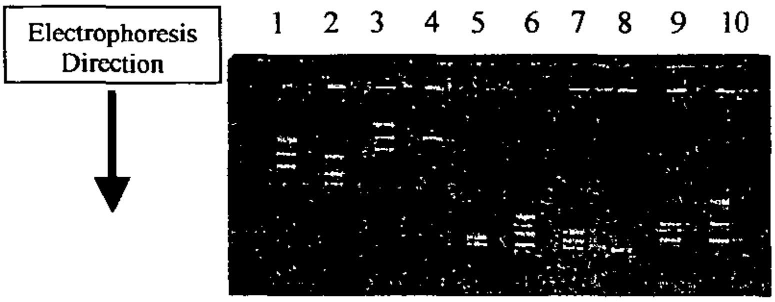


FIGURE 5.6: A SELECTION OF HETEROZYGOUS AND HOMOZYGOUS AMPLIFIED SAMPLES AT SIX DIFFERENT TETRANUCLEOTIDE SHORT TANDEM REPEAT LOCI. AN EL600 SPREADEX GEL WITH OPTIMAL RESOLUTION WITHIN THE RANGE 150-300 BASE PAIRS WAS CHOSEN.

(REFER TO TABLE 5.7 FOR LANE NUMBER: SAMPLE/LOCUS INFORMATION).

Samples were electrophoresed for 1 hour 15 minutes at 300 Volts (10V/cm between electrodes) in 1X TAE buffer, prepared from 40X concentrate (see appendix for composition).

TABLE 5.7: THIS INFORMATION CORRESPONDS TO THE SAMPLES RUN ON THE SPREADEX GEL IN FIGURE 5.7.

Gel Lane Number	Sample Number	Locus	Homozygous/ Heterozygous	Total number of bands observed
1	518	D12S297	Heterozygous	3
2	82	D3S1514	Heterozygous	3
3	608	D4S2285	Heterozygous	3
4	615	D4S2285	Homozygous	1
5	508	D5S592	Heterozygous	2
6	505	D5S592	Heterozygous	3
7	515	D10S520	Heterozygous	3
8	521	D10S520	Homozygous	1
9	528 + 524	D10S526	Heterozygous	3
10	528 + 524	D10S526	Heterozygous	3

The ethidium bromide stained gel was placed onto the UV transilluminator. The bands were carefully excised using a sterile scalpel and placed into clean 500µl microtubes.

To amplify the excised bands, the PCR reagent mix was added directly to the gel slice. No crushing or dissolving of the gel was required. These samples were amplified using the relevant optimised PCR protocols.

The PCR products were then purified following the QIAquick spin methodology in QIAGENS handbook. This protocol purify's single- or double-stranded PCR products ranging 100bp to 10kb, removing primers, nucleotides, polymerases and salts (see appendix for details). However, since a 50µl reaction PCR volume was used instead of 100µl as stated in the protocol, 200µl of buffer PB and 40µl of PCR reaction was used instead of 500µl of buffer PB and 100µl PCR reaction as the methodology stated. The purified product was concentrated to 50µl in buffer 'EB' (10mM Tris-HCl, pH 8.5).

5.z.iii. Preparation for sequencing

In preparation, samples to be sequenced at *Alta Bioscience* had their DNA concentration evaluated, as specific concentrations were required for sequencing. The DNA DipStick Kit by INVITROGEN was used to estimate the DNA concentrations in the range of 0.1 to >10.0 ng/µl. The protocol for this kit was followed exactly, without any alteration to volume or concentrations (see table 5.8 for results).

TABLE 5.8: CONCENTRATION OF AMPLIFIED PRODUCT IN 40-50µL PCR REACTION VOLUME.

Sample Number	Locus	DNA Conc. (ng/µl)
611	D12S297	5
611	"	5
611	"	5
518	D12S297	10
518	"	10
518	"	10
82	D3S1514	3
82	"	5
82	"	5
608	D4S2285	6
608	"	3
608	"	5
615	D5S592	5
508		4
508		4
505		10
505	D10S520	10
515		50
515		20
515		20
521		4
528+521		4
528 +521		6

Alta Bioscience required a concentration of 10-20ng/µl in 80µl, thus the 1st PCR did not yield a high enough concentration. Therefore, an one microlitre aliquot of the cleaned up sample was re-amplified in a 50µl volume and this product cleaned and added to the first clean up trial to raise the concentration of the

DNA to 500ng – 1μg in 80μl, as stipulated by Alta Bioscience for their sequencing protocols.

In addition to the known DNA concentration of DNA, also 18μl of the primers needed to be provided at a concentration of 5 picomole/μl. Respective dilution factors were calculated and the primers diluted accordingly.

Primer	Conc (nM)	Dilution factor to gain 100 pM/20μl
D12S297 Forward	152.6	1μl stock to 1525μl ddH_20
D12S297 Reverse	78.9	1μl stock to 789μl ddH_20
D3S1514 Forward	143.4	1μl stock to 1433μl ddH_20
D3S1514 Reverse	140.7	1μl stock to 1406μl ddH_20
D5S592 Forward	122.0	1μl stock to 1219μl ddH_20
D5S592 Reverse	92.0	1μl stock to 919μl ddH_20
D10S520 Forward	127.1	1μl stock to 1270μl ddH_20
D10S520 Reverse	108.0	1μl stock to 1079μl ddH_20
D4S2285 Forward	80.6	1μl stock to 805μl ddH_20
D4S2285 Reverse	101.2	1μl stock to 1011μl ddH_20

20μl of these diluted primers were aliquoted labelled and sent with the primers to Alta Biosciences.

AltaBio Sequencing service

The cycle sequencing service incorporated the AmpliTAQ DNA polymerase FS and used the *ABI Prism, BigDye terminator cycle sequencing ready reaction kit (PE Applied Biosystems – Perkin Elmer).* The flourescent dyes needed for the 377 sequencer were linked to the di-deoxy nucleotides. These were incorporated into DNA during the sequencing extension reactions the resultant different sized fragments are detected in the usual manner through a polyacrylamide gel. The bases were coloured differently hence recognition of a particular base at a particular region within the fragment.

The results of the sequencing reactions varied with respect to the quality and quantity of the PCR fragment to be sequenced. In general, if too much DNA template was used for the sequencing reaction then the polymerase enzyme was exhausted after just 50-70 bases thus reducing the amount of product extending to the full length (refer to raw sequencing data appendix).

However, the critical point of the sequencing results was to confirm that the extraneous bands believed to be heteroduplex structures, were of the same nucleotide sequences as the 'true' alleles. Sequencing the extraneous bands from a PCR reaction containing two different DNA samples both of known homozygosity at the D10S526 locus, indeed confirmed heteroduplex formation within the PCR reaction (re: figure 5.5 this chapter).

Perhaps the formation of the heteroduplex structure occurred during amplification, or possibly during the final cool down period, at the end of the thermocycle programme.

Additionally, the sequencing results were compared to the basic locus information supplied by the Utah Marker Development group. There was very good sequence homology between the sequence results in this present study and those of the Utah Marker Development group (see autosomal results chapter for a comparison of sequenced loci). Interestingly, base substitutions were observed between the Utah Marker Development group sequence information and that generated by Alta Bioscience, for example, D4S2285 had a T to C substituiton at 191 base pairs comparing Utah information to Alta Bioscience respectively (see autosomal results chapter for more information). Similarly, a G to T substitution was observed at D10S520 (position 131 bp) and a G to C substitution at D7S1485 (position 160 base pairs). This would be an interesting area to expand upon for further research, however, is not an integral part of this present thesis and hence will not be discussed at length.

5.z.iv. Summary of the preliminary methodology

In summary, K562 DNA was chosen in the first instance to provide a suitable DNA control for the optimisation of the amplification of the tetranucleotide short tandem repeat loci. The thermocycle programmes for these loci varied according to the 'Tm' or melting temperature of the primers.

Two separate systems were tested for the analysis of the amplified PCR product. Firstly, denaturing polyacrylamide gel electrophoresis and secondly the new 'spreadex' submarine gel electrophoresis system. The denaturing polyacrylamide gel electrophoresis was inappropriate as the denatured DNA strands 'ran' to different lengths of the gel due to conformational differences of the four DNA strands. Therefore, one was unable to confidently size the alleles, as the silver staining technique could not discriminate between the paired strands of each allele. The new 'spreadex gel' system from Elchrom allowing double stranded DNA to be analysed, eliminated the ambiguity of discriminating between the four strands. However, heteroduplex formation was believed to be present as extraneous bands (running to a position on the gel associated with a higher molecular weight than the actual alleles) were observed in only the heterozygous condition. A similar observation was made by purposefully blending and amplifying, two DNA samples which were both homozygous at a particular locus although differed in allele length. To finally validate the heteroduplex formation, samples were sent to Alta Biosciences for sequencing. There was a definite sequence homology between the 'heteroduplex' band to either allele, thus confirming the heteroduplex theory. Therefore, confident allele calling of the two lower molecular weight bands excluding the supposed 'higher molecular weight' band was carried out and Elchroms spreadex gel electrophoresis was chosen as the method of choice.

Chapter 6

Section I: The Actual Autosomal Methodology

The Handle Crank

The preferred equipment and reagents as suggested by Elchrom for use with the spreadex gels included:

- the *SEA2000 submarine electrophoresis* unit with pump delay controller,
- catamarans to accommodate all sizes of *spreadex* gel,
- 40X Running buffer (Tris-acetate EDTA) and
- forceps to remove the gel from the apparatus after electrophoresis.

Additional equipment not provided by Elchrom:

- A *Fisons HAAKE D1* thermostatically controlled thermal circulator was used to sustain the electrophoresis apparatus at the required temperature (55°C).

Additional equipment used to carry out the staining procedure and the recording of results.

- A *R100 Luckham Rotatest Shaker* to gently shake the gel in the staining solution.
- An Ultra Violet *transilluminator* and a *UVP Image store 7500* to process and record the photograph of the stained gel. A hard copy of the photograph could be processed using the *Video Graphic Printer.*

The SEA2000 apparatus was superior to the HYBAID standard submarine electrophoresis equipment for a number of reasons.

1. The SEA2000 generated an electric field linear with the electrodes at the same level as the gel. In comparison HYBAID generated an arc-like electric field passing between anode and cathode as the electrodes were positioned lower than the gel. Furthermore, the electrodes were closer to the gel in the SEA2000 allowing higher field strengths than in standard submarine units.
2. Buffer circulation in the SEA2000 was perpendicular to the electrical flow to remove air bubbles forming at the electrodes and to renew the buffer to reduce the pH differential between the electrodes.

Were such adjustments merely a costly addition or were they improvements to the standard submarine electrophoresis system?

Spreadex gels have been used in both systems, with very different results. Samples electrophoresed using the standard equipment have been observed to 'smile'. This indicated an inconsistent heat flow across the gel. Furthermore, the running times of the gels were hours longer and the apparatus towards the end of the electrophoresis run was overheating with a profuse amount of air bubbles forming at the electrodes and around the gel. In direct comparison the SEA2000 equipment controlling for heat transfer across gels, allowed an even heat distribution with quicker electrophoresis time without overheating or distorting the gels. Therefore, the SEA2000 was indeed an improvement to using the standard electrophoresis system.

6.a. The Optimised Complete Methodological Procedure

The PCR reagent mix concentrations were kept the same as the optimized preliminary methodologies. The thermocycle programme using the *Perkin Elmer 480* thermal cycler, remained as a touchdown protocol with the following parameters;

	Number of cycles	Temperature (°C)	Time (Secs)
Initial denaturation	Pre cycles	96	120
Denaturation	Each cycle	96	30
Annealing	1 cycle per degree	59-52	30
	25 cycles	51	30
Chain extension	Each cycle	72	45

The annealing temperature range appeased the majority of the quoted temperature range of the primers (see table 5.3). Each locus was amplified sequentially, with all 175 of the Polynesian and 50 Leicestershire U.K. Caucasian DNA samples. The efficacy of amplification was assessed using 1% agarose gels stained with ethidium bromide. Aliquots of PCR product (2µl) were mixed with 3µl loading buffer loaded into the preformed wells of the 'check gel' and electrophoresed following the protocol as given, in Chapter 5.

Locus D10S507 did not amplify either the Polynesian or the Leicestershire DNA samples using the optimised DNA protocol. After two further amplification attempts, analyses using this locus were stopped. Similarly, loci D2S1279, D10S520, D10S521, D4S2289, D20S156 and D7S1517 did not amplify at all well using the optimised PCR protocol, therefore further amplification attempts were continued with caution. Of these unproductive loci, D10S521 and D4S2289 were both problematical at the preliminary optimisation stage using K562 DNA. Loci D10S521, D20S156 and D7S1517 were purported to require an optimal magnesium concentration of 1.25mM, 0.25mM less than the standard magnesium concentrations (Utah Marker Development group 1998 personal communication).

6.b. Allelic Ladders – The sizing of unknown bands- Protocol I

The CTT and FFv allelic ladders used with the spreadex gels were not clearly defined (see figure 6.1), and so were not used as size standards using the SEA2000 apparatus.

The CTT allelic ladder appeared as smearing with no definition to the bands at all. The FFv allelic ladders were a little clearer although, only alleles of the FES/FPS ladder were clearly defined. However, in comparison to the allelic ladders, the PCR products produced sharp, brightly fluorescing bands under ultra violet irradiation. Heteroduplex formation was observed as fainter bands situated above the two prominent 'lowest' molecular weight bands. The term 'lowest' has been used here to describe the difference between the actual alleles to be sized and the heteroduplex. It should be observed that the heteroduplex is not of higher molecular weight even though appearing to have electrophoresed more slowly than the alleles.

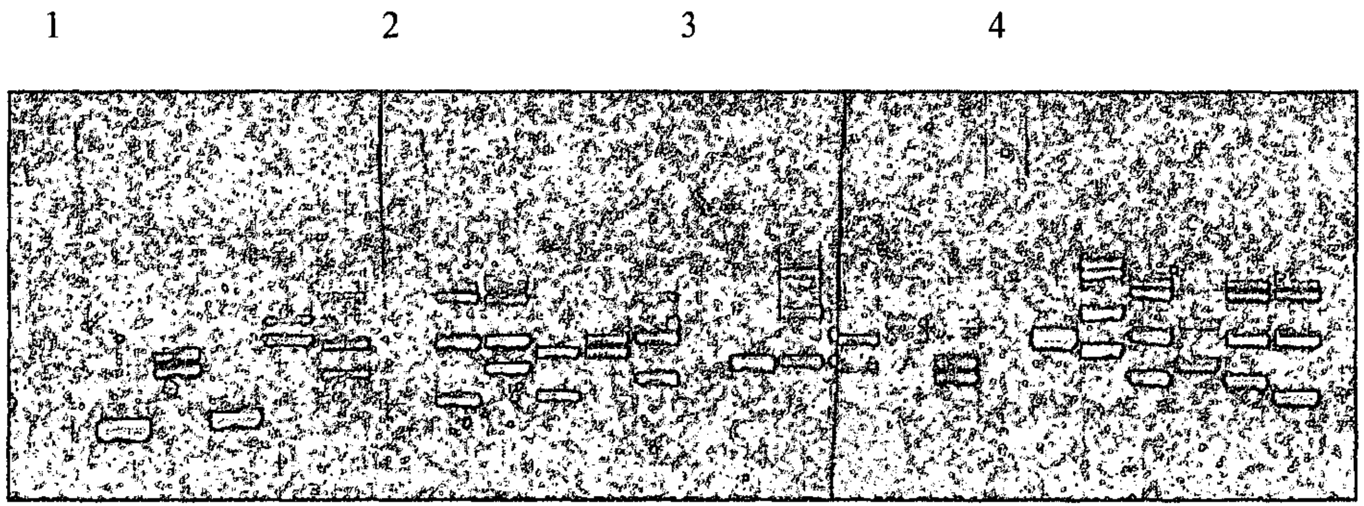


FIGURE 6.1: COMPARISON OF RESOLUTION BETWEEN PCR PRODUCTS AT THE D3S1514 LOCUS AND THE FFV AND CCT ALLELIC LADDERS. NUMBERS 1,3 AND 4 CORRESPOND TO FFV ALLELIC LADDERS AND NUMBER 2 CORRESPONDS TO THE FFV ALLELIC LADDER (SEE APPENDIX XX FOR LADDER COMPOSITION).

6.b.i. The Arbitrary Allelic Ladder

Allelic ladders of randomly chosen samples used to compare allelic bands were designed. These incorporated PCR products from 4-5 different samples which were concentrated by ethanol precipitation, spinning and rehydrating in a smaller volume of TE buffer. When run on a gel these were designated an arbitrary number starting at 1, given to the smallest allele or lowest molecular weight band. If alleles were found outside the ladder range, they were coded as for example, 0 or –1 for lower molecular weight bands. The arbitrary allelic ladder had the advantage that it was specific to the locus and the bands could be easily identified. However, it had the disadvantage that the exact base pair size of the alleles was unknown. Furthermore, heteroduplex formations could interfere with the 'actual' alleles.

The exact methodology for the preparation of the allelic ladders is given below:

Four or five Polynesian DNA samples were chosen that amplified producing a strong band with little or no high molecular weight smearing. This ensured the best resolution of allelic ladders possible. The samples chosen as listed overage (table 6.1);

TABLE 6.1: SAMPLES USED TO PRODUCE ARBITRARY ALLELIC LADDERS AT EACH LOCUS, LADDER NUMBERS REFER TO FIGURE 5.8

Locus	Ladder Number	Sample Numbers
D12S297	A1	534, 525, 517, 528, 526
D12S297	A2	611, 521, 520, 516
D12s297	A3	531,607,604, 523
D7S1485	B1	410, 436, 401, 112, 109
D7S1485	B2	114, 422, 440, 404
D7S1485	B3	102, 100, 61, 420
D4S2285	C1	115, 116, 418, 419, 401
D4S2285	C2	420, 423, 114, 406
D4S2285	C3	421, 410, 432, 417
D5S592	D1	513, 503, 475, 488, 515
D5S592	D2	477, 482, 507, 501
D5S592	D3	500, 512, 475, 509
D1S407	E1	489, 477, 502, 488
D1S407	E2	623, 612, 621, 617, 624
D1S407	E3	618, 616, 625, 614, 622
D9S252	F1	451, 458, 454, 457, 469
D9S252	F2	449, 443, 455, 464
D9S252	F3	442, 448, 456, 444
D7S618	G1	465, 479, 475, 470, 460
D7S618	G2	476, 464, 468, 463
D7S618	G3	489, 478, 482, 471
D4S2289	H1	605, 604, 603, 602, 601
D4S2289	Not shown	609, 608, 607, 606
D4S2289	Not shown	613, 612, 611, 610
D3S1514	Figure 5.9 '1'	451, 489, 502, 506, 509
D3S1514	Figure 5.9 '2'	454, 453, 451, 449
D3S1514	Figure 5.9 '3'	433, 425, 424, 423, 422

A total of 15μl of each amplified DNA sample was pipetted into a 1.5ml microtube. This was kept on ice and 400μl of chilled absolute ethanol added. The ethanol/DNA solution was gently mixed by repeatedly inverting the microtube for 3 minutes to precipitate the DNA.

The microtubes were then centrifuged for 3 minutes @ 10,000 RPM to form an oligonucleotide pellet. The solution above the pellet was carefully decanted and the pellet allowed to dry for 5-10 minutes in a 30°C sterile incubator.

Tris-acetate EDTA buffer pH8.0 was pipetted into the microtube to rehydrate the pellet to a final volume of 40μl. The microtube was briefly vortexed and spun for a few seconds to collect the contents at the base of the tube before placing into a 56°C water bath for 1 hour.

The microtubes were then briefly spun to collect any condensed solution around the walls of the tubes.

6.c. Spreadex Gel Electrophoresis

Aliquots (5µl) of each of the prepared 'allelic ladders' were mixed with 3µl bromophenol blue loading dye and run on an EL600 'spreadex' gel. The ladders were loaded sequentially leaving a lane between ladders of different loci. Elchroms submarine gel electrophoresis tank set at 120 Volts 55°C for 90 minutes. The spreadex gel was then stained (30 minutes) using ethidium bromide (0.75µl in 50ml distilled sterile water) in sterile distilled water and destained for 30-40 minutes in sterile distilled water.

Results

In general, the protocol devised to construct an arbitrary allelic ladder was a success. The notations given in figures 6.2 and 6.3 corresponded to the information in table 6.1.

The alleles at loci, D9S252 and D7S618 were not as clearly defined as those at the other loci. The third ladder of locus D1S407 had not worked very well, possibly because the oligonucleotides were probably lost at the ethanol precipitation stage. The resolution of B and C (loci D7S1485 and D4S2285 respectively) would have been improved with a longer electrophoresis time. However, for the purposes of examining the ladders at these loci, the running time was adjusted to accommodate the fastest migrating locus, or locus with the smallest sized alleles.

6.c.i. Arbitrary Sizing

The arbitrary sizing of the ladders at a particular locus was determined by allocating the number '1' to the smallest allele. Allelic ladder '3' locus D5S592, was observed to have the smallest allele of the three ladders. This was termed '1'. This ladder also had allele numbers 2,3,4 and 6. This locus similarly to all the other loci has a tetranucleotide repeat, therefore the allele sizes differ by 4 base pairs. If all alleles were present in one ladder, one would observe regularly spaced bands. However, as observed with ladder '3' locus D5S592, allele 5 was not present. Alternatively, ladder '1' at locus D5S592 had alleles 2, 4, 5 and 6 and ladder '2' had alleles 2,3 and 4. Therefore, these ladders loaded at regular intervals onto a gel, e.g. six lanes apart act as suitable markers to which the amplified DNA samples could be arbitrarily sized.

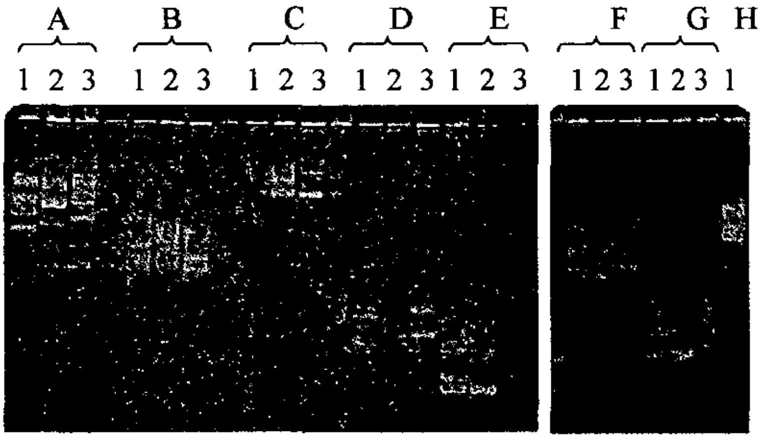


FIGURE 6.2: ALLELIC LADDERS OF EIGHT LOCI ELECTROPHORESED THROUGH AN EL600 SPREADEX GEL FOR 90 MINUTES AND STAINED WITH ETHIDIUM BROMIDE. LABELS A-H WERE REFERRED TO IN TABLE 5.10.

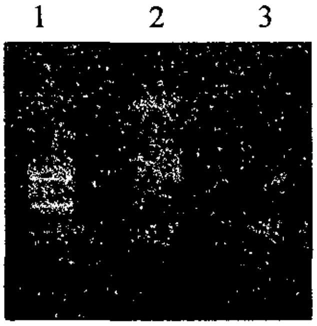


FIGURE 6.3: THE THREE ALLELIC LADDERS OF LOCUS D3S1514. THE BANDS ALTHOUGH VISIBLE DID NOT STAIN VERY WELL WITH THE ETHIDIUM BROMIDE. LONGER DESTAINING MAY HAVE BEEN REQUIRED.

6.d. An alternative approach to sizing alleles – *Protocol II*

A second approach was to run the HAEIII digestion ladder with the samples (see figure 6.4). The HAEIII ladder is most commonly used to size the restriction enzyme digest using HAEIII. However, the allele sizes in base pairs in this ladder were 310, 281, 271, 234, 194, 118 and 72. Although these did not match the allele sizes directly, once the gel had been photographed, the image was scanned onto the computer.

The computer programme (UVItech) was written specifically to size the alleles. The alleles of the ladders were compared to the lengths of the unknown samples and an exact base pair size was recorded (see figure 6.4 and table 6.2).

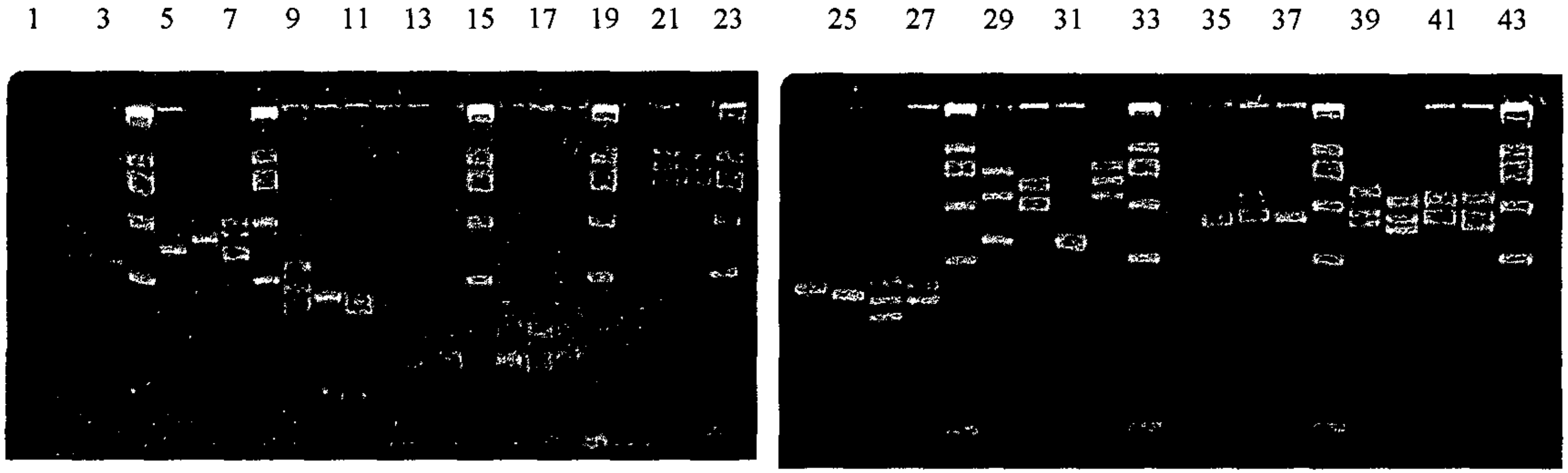


FIGURE 6.4: TWO SPREADEX GELS WERE ELECTROPHORESED SEPARATELY. THE HAEIII LADDER WAS CHOSEN TO SIZE THE ALLELES IN CONJUNCTION WITH THE SEMI-AUTOMATED COMPUTER PROGRAMME 'UVITECH'. REPRESENTATIVE AMPLIFIED PRODUCTS AT THE TEN CLEAREST SYSTEMS WERE ELECTROPHORESED FOR 95 MINUTES @ 55°C IN THE ORDER AS SHOWN IN TABLE 5.11(OVERPAGE).

TABLE 6.2: PRELIMINARY SPREADEX GEL ANALYSIS USING THE ALLELIC LADDER HAEIII AND THE COMPUTER PROGRAM UVITECH, TO SIZE THE ARBITRARY ALLELES. LANE NUMBERS CORRESPOND TO THOSE LISTED IN FIGURE 6.4.
SAMPLE NUMBERS WITH THE PREFIX 'L' CORRESPOND TO THE LEICESTERSHIRE UK CAUCASIAN SAMPLE POPULATION.

Lane Number	Locus/ Ladder	Sample number	Homozygous/ Heterozygous	Allele sizes (Base Pairs)
4, 8, 15, 19, 23, 28, 33, 38 and 43	HAE III allelic ladder	------------- -----------	-------------------- ----------------	Allele sizes (bp) 310, 281, 271, 234, 194 and 118
1	D2S262	L161	Heterozygous	199, 203
2	D2S262	L144	Heterozygous	207, 211
3	D2S262	98	Homozygous	203, 203
5	D7S1485	489	Homozygous	212, 212
6	D7S1485	473	Homozygous	220, 220
7	D7S1485	621	Heterozygous	208, 212
9	D5S592	L24	Heterozygous	178, 186
10	D5S592	525	Homozygous	182, 182
11	D5S592	453	Heterozygous	178, 182
12	D7S618	107	No Result	----------
13	D7S618	78	Heterozygous	142, 150
14	D7S618	L2	Homozygous	154, 154
16	D1S407	413	Homozygous	152, 152
17	D1S407	464	Heterozygous	148, 152
18	D1S407	112	Homozygous	152, 152
20	D4S2285	601	No Result	----------
21	D4S2285	L180	Heterozygous	277, 297
22	D4S2285	113	Heterozygous	277, 293
24	D10S520	L104	Homozygous	179, 179
25	D10S520	L188	Homozygous	175, 175
26	D10S520	104	Heterozygous	163, 171
27	D10S520	419	Heterozygous	171, 171
29	D12S296	611	Heterozygous	205, 245
30	D12S296	L125	Heterozygous	233, 241
31	D12S296	454	Homozygous	205, 205
32	D12S296	518	Heterozygous	245, 265
34	D3S1514	617	Heterozygous	210, 230
35	D3S1514	91	Heterozygous	218, 222
36	D3S1514	84	Heterozygous	222, 226
37	D3S1514	L163	Homozygous	222, 222
39	D9S252	611	Heterozygous	218, 226
40	D9S252	480	Heterozygous	214, 222
41	D9S252	106	Heterozygous	222, 226
42	D9S252	L1	Heterozygous	218, 226

It was observed that the HAEIII ladder produced clearer, sharper allelic bands, and the resolution of these bands was superior to the CCT and FFV allelic ladders previously used (see figures 5.2 and 5.4).

The results using the semi-automatic computer aided sizing method could be compared to the arbitrary technique. Assigning the arbitrary alleles of the allelic ladder an exact base pair size using the *UVItech* programme and HAEIII ladder, it was possible to re-size the alleles to base pairs. If the two methods did not agree then the sample was electrophoresed again using the HAEIII ladder and re-sized using the *UVItech* system.
These methods in concert, had the additional advantage of acting as a control on sizing the bands, and as a check on the two systems used. Furthermore, an unbiased second person oversaw these methods and checks, to corroborate the results and methodology used.

6.e. Methodology for Elchroms complete spreadex gel system

The spreadex gels were allowed to equilibrate to the running temperature by placing the sealed gel package on top of the already equilibrated electrophoresis equipment (50°C) 10 minutes before use.

The gel was then removed from its' packaging, placed into the buffer and held down by the catamaran.

The buffer pump used to circulate the heated buffer around the tank, was turned off when loading the gel. A mix of 2µl loading buffer and 3µl of the PCR product was prepared for each sample and 2µl allelic ladder with 2µl loading buffer. If a 25 well gel was used then the samples were pipetted one at a time. However, if a 100 well gel was used the samples were first added to a microtitre plate and mixed with the loading buffer, so that a multichannel pipette could be used to load the sample/buffer mix onto the gel. The advantage of using this system was that the quicker the gel was 'loaded' less diffusion occurred in the wells before an electric current was applied. If the diffusion in the well was too great then the PCR product or ladder did not produce neat, sharp bands when the gel was stained with ethidium bromide.

After all the samples had been 'loaded' into the wells of the gel, the pump delay was switched on to 1.5min. This allowed the samples to pass through into the gel before the buffer circulation started, otherwise the well contents were flushed out.

The voltage was set to 10V/cm (120 volts) between electrodes, temperature to 55°C and the duration of the electrophoresis dependent of the size of the allele fragments. Between 90-135 minutes was usual, although the exact timings for each locus are given below;

Locus	Optimal Electrophoresis Time (Mins)	Locus	Optimal Electrophoresis Time (Mins)
D1S407	75	D7S1485	105
D2S262	105	D7S618	95
D3S1514	105	D9S252	135
D4S2285	90	D10S520	90
D5S592	80	D12S297	135

After the appropriate length of time the gel was removed from the apparatus using the forceps provided. The plastic backing was removed, by sliding a nylon thread between gel and plastic to break the seal. The plastic backing was then peeled away and the gel placed into the staining tray.

Ethidium bromide (0.75µl) was mixed with 50ml of sterile distilled water and poured gently over the gel. The gel was then allowed to gently shake for 30 minutes to stain it.

The staining solution was then poured away, to an ethidium bromide waste disposal flask. 50ml of distilled sterile water was added as a destaining agent to remove excess ethidium bromide and the gel shaken for a further 30 minutes to destain.

6.f. Summary

Elchrom's spreadex gel technology for use in their submarine gel electrophoresis tank provided a faster, more efficient resource to analyse the amplified PCR products in comparison to the silver staining procedure. The ability to analyse double stranded DNA removed the ambiguity of sizing the single stranded DNA observed using denaturing polyacrylamide gels. The quality of the precast 'spreadex' gels was validated by the manufacturer, thus reassurance that the gels gave consistent and reliable results.

The initial 17 possible marker systems were reduced to a reliable and robust 10. These were; D1S407, D2S262, D3S1514, D4S2285, D5S592, D7S1485, D7S618, D9S252, D10S520 and D12S297. The raw data for each population at each of the ten loci has been listed in the appendix. The ten loci amplified products varying in length in accordance with the primer sites. These loci may be investigated further to assess their potential as multiplex markers for use in flourescent technology. Analysis of multiplex systems with the spreadex system would be problematic as the resolution of PCR products of three or more combined systems across a typical 4cm gel for a range of 150-350 base pairs would be highly unreliable.

Section II

Y chromosome Methodology

Y chromosome analyses were carried out under the supervision of Dr Thomas at University College London following his published protocol (Thomas et al. 1999).

Therefore, for the purposes of this thesis the Y chromosome methodology follows the published protocol, although with specific ammendments.

Thomas et al.'s (1999) methodology described the same 11 diallelic polymorphisms as used in the present study and 10 Y chromosome microsatellites of which 6 were used in the present study.

One microsatellite multiplex 'kit' and two diallelic (or UEP) multiplex kits were designed to accommodate all the marker systems.

6.g. Microsatellite Multiplex

The microsatellite kit (MS) contained four primer pairs to amplify DYS19, DYS388, DYS390 and DYS393 (Kayser et al. 1997) and two primer pairs that were redesigned to amplify DYS391 and DYS392 (Thomas et al. 1999).

All PCR products from this multiplex fell within a 100- 230bp range and discriminated using a combination of size and fluorescent dye labels.

6.h. Apparatus

ABI-310 genetic analyser (Perkin Elmer)

ABI-377 automated sequencer (Perkin Elmer)

Biometra – Uno II thermal cycler

384- Microtiter plates

0.2 ml thermocycle tubes

6.i. Microsatellite PCR Reagents

PCR mix

Reagent	Final concentration in 10μl volume
dNTPs	200μM
Tris-HCl (pH9.0)	10mM
Triton X-100	0.1%
Gelatin	0.1%
KCl	50mM
MgCl	2.2mM
Taq polymerase (HT Biotech)	0.13U
TaqStart Monoclonal antibody	9.3nM
Primers	As listed in table 6.3

The primer sequences for each locus together with the dye label for the appropriate primer of each locus pair is given below (table 6.3);

Primer Name	Primer Sequence	Dye Label	Final Conc. (μM)
DYS19-L	CTACTGAGTTTCTGTTATAGT	TET	0.236
DYS19-R	ATGGCATGTAGTGAGGACA		0.236
DYS388-L	GTGAGTTAGCCGTTTAGCGA	TET	0.318
DYS388-R	CAGATCGCAACCACTGCG		0.318
DYS390-L	TATATTTTACACATTTTTGGGCC		0.127
DYS390-R	TGACAGTAAAATGAACACATTGC	FAM	0.127
DYS391-L-N[a]	CTATTCATTCAATCATACACCCATAT	FAM	0.384
DYS391-R-N[a]	ACATAGCCAAATATCTCCTGGG		0.384
DYS392-L-N[a]	AAAAGCCAAGAAGGAAAACAAA		0.155
DYS392-R-N[a]	CAGTCAAAGTGGAAAGTAGTCTGG	HEX	0.155
DYS393-L	GTGGTCTTCTACTTGTGTCAATAC		0.180
DYS393-R	AACTCAAGTCCAAAAAATGAGG	HEX	0.088

TABLE 6.3: MICROSATELLITE PRIMER SEQUENCES, THEIR DYE LABEL AND CONCENTRATIONS.

Microsatellite Thermocycle Programme
Cycling parameters were;

♦ Initial denaturation	5 minutes	95°C

Then 38 cycles of;

♦ Denaturation	1 minute	94°C
♦ Annealing (Tm)	1 minute	57°C
♦ Extension	1 minute	72°C

A final incubation step of 72°C for 10 minutes was included.

6.j. Preparation of microsatellite multiplex PCR mix

All PCR reagents except the Taq polymerase and TaqStart monoclonal antibody were pre-mixed and stored at -20°C. The Taq and TaqStart were added to the other reagent just prior to carrying out the PCR. Primers were also mixed and stored as a 10X stock, to be added only as required.

To minimise the time that the Taq enzyme was in contact with primers and the other reagents, 1µl of template DNA was placed at the base of the 0.2 ml PCR tube. Then 9µl of the PCR mix with the Taq/TaqStart added, was pipetted into the lid of each PCR tube. The lid of each PCR tube was then carefully closed and the reaction tubes were placed in a thermal cycler.

6.k. Analysis of the microsatellite PCR product

1.2µl aliquots of the PCR product were mixed with 0.5µl size standard labelled with the fluorescent dye TAMRA (PE-Applied Biosystems) and 12µl of de-ionised formamide, for use with the ABI-310 genetic analyser.

Samples were denatured for 3 min @ 96°C and chilled on ice for 5 minutes before being run.

6.I. 1st UEP Multiplex

UEP Multiplex

The 1st UEP (or diallelic) multiplex kit contained primers to amplify loci 92R7 (Mathias et al. 1994), sY81 (Seielstad et al. 1994), SRY+465 (personal communication to M. Thomas), SRY 4064 (Whitfield et al. 1995), Tat (Zerjal et al. 1997) and YAP (Hammer 1994). All the markers with the exception of YAP (an alu insertion polymorhism) were single nucleotide substitutions.

Apparatus

ABI-310 genetic analyser (Perkin Elmer)

ABI-377 automated sequencer (Perkin Elmer)

Biometra – Uno II thermal cycler

384- Microtiter plates

0.2 ml thermocycle tubes

1st UEP PCR Reagents

PCR mix

Reagent	Final concentration in 10μl volume
dNTPs	200μM
Tris-HCl (pH9.0)	10mM
Triton X-100	0.1%
Gelatin	0.01%
KCl	50mM
MgCl	1.5mM
Taq polymerase (HT Biotech)	0.13U
TaqStart Monoclonal antibody	9.3nM
Primers	As listed in table 6.4

The primer sequences for each locus together with the dye label for the appropriate primer of each locus pair is given below (table 6.4);

Primer Name	Primer Sequence	Dye Label	Final Conc. (μM)
92R7-A	TGCATGAACACAAAAGACGTA		0.125
92R7-R	GCATTGTTAAATATGACCAGC	HEX	0.125
TAT-L	GACTCTGAGTGTAGACTTGTGA		0.078
TAT-R	GAAGGTGCCGTAAAAGTGTGAA	TET	0.078
sY81-L	ATGGGAGAAGAACGGAAGGA	FAM	0.125
sY81-R	TGGAAAATACAGCTCCCCCT		0.125
SRY+465-L	GCCGAAGAATTGCAGTTTGC		0.055
SRY+465-R	GTTGATGGGCGGTAAGTGGC	HEX	0.055
SRY4064-L	GGTATGACAGGGGATGATGTGA	TET	0.095
SRY4064-R	CCACGCCCAGCTAATTTTTTGT		0.095
YAP-C	AGGACTAGCAATAGCAGGGGA AGA	TET	0.100
YAP-D	CAGGGCCAACTCCAACCAAG		0.100

TABLE 6.4: 1ST UEP MULTIPLEX PRIMER SEQUENCES, THEIR DYE LABEL AND CONCENTRATIONS.

1st UEP multiplex Thermocycle Programme
Cycling parameters were;

♦ Initial denaturation	5 minutes	95°C

Then 38 cycles of;

♦ Denaturation	1 minute	94°C
♦ Annealing (Tm)	1 minute	58°C
♦ Extension	1 minute	72°C

A final incubation step of 72°C for 10 minutes was included.

Preparation of 1st UEP multiplex PCR mix

The same methodology as for the microsatellite multiplex was used.

1st UEP multiplex restriction enzyme digestion

Digestions were performed in 384-well microtiter plates in a final volume of 8µl. Each reaction contained;

Reagent	Concentration/volume
PCR product	2µl
NEB buffer 4 (Biolabs)	1 X
Acetlyated BSA	0.01 µg/µl
*Bsr*BI	0.3U
*Fnu*4HI	0.3U
*Nla*III	0.3U
*Hind*III	1.8U

The predicted sizes of the digested product together with its associated polymorphic status for each dye-labelled PCR product, were as given below;

Poly-morphism	Enzyme	Other enzymes that also cut	Labelled Primer	Dye Label	PCR product size	Size of labelled cut fragment
92R7	*Hind*III	*Nla*III	R	HEX	55	28 (C)
TAT	*Nla*III		R	TET	112	83 (T)
SY81	*Nla*III		L	FAM	142	105 (A)
SRY+465	*Fnu*4HI		R	HEX	148	98 (C)
SRY4064	*Bsr*Bi	*Nla*III	L	TET	225	135 (G)
YAP	NA		L	TET	99/413	NA

N.B. the letters in parentheses of the cut fragment size (bp) refer to the base that is present at the cut site. This was not an indication of whether the fragment contained the ancestral or derived form of the polymorphism. Primer 'R' and 'L' (right and left) refer to the direction of chain extension of the sense and anitsense strands of the DNA.

Analysis of the 1st UEP multiplex PCR product

1.0μl aliquots of the digestion product were mixed with 2.0μl of a loading buffer (formamide: dextran blue: TAMRA (PE-Applied Biosystems) size standard, in the ratio 23: 4: 2), for use with the ABI-377 automated sequencer.

Samples were denatured for 3 min @ 96°C and chilled on ice for 5 minutes before being run.

Samples required electrophoresis on a 12cm 6.5% gel for 2.5 hours. (The acrylamide was set at 6.5% to resolve the 28bp HEX-labelled digestion product of the 92R7 digest).

6.m. 2nd UEP Multiplex

UEP Multiplex

The 2nd UEP (or diallelic) multiplex kit contained primers to amplify loci the single nucleotide substitutions; M9, M13, M20 (Underhill et al. 1997) and SRY10,831 (Whitfield et al. 1995) and the single base pair deletion polymorphism, M17 (Underhill et al. 1997). No restriction enzyme was found that would discriminate between the two allelic forms of M17. Thus, one primer for this locus was redesigned to mismatch at a single position thereby creating a restriction enzyme recognition site.

2nd UEP Multiplex; PCR Reagents

The PCR reagent mix was the same as for the 1st UEP multiplex system.

The primer sequences for each locus together with the dye label for the appropriate primer of each locus pair is given below (table 6.5);

Primer Name	Primer Sequence	Dye Label	Final Conc. (μM)
M9-L	TCAGGACCCTGAAATACAGAACT	TET	0.125
M9-R	TTGAAGCTCGTGAAACAGATTAG		0.125
M13-L	TAGTTTATGCCCAGGAATGAAC	HEX	0.078
M13-R	ATCCAACCACATTTGCAAAA		0.078
M17-L	GTGGTTGCTGGTTGTTACGT		0.125
M17-R	AGCTGACCACAAACTGATGTAGA	TET	0.125
M20-L	AGTTGGCCCTTTGTGTCTGT	FAM	0.055
M20-R	CATGTTCAGTGCAAATGCAAC		0.055
SRY10831-L	TCATTCAGTATCTGGCCTCTTG	FAM	0.095
SRY10831-R	CACCACATAGGTGAACCTTGAA		0.095

TABLE 6.5: 2ND UEP MULTIPLEX PRIMER SEQUENCES, THEIR DYE LABEL AND CONCENTRATIONS.

2ND UEP multiplex Thermocycle Programme

Cycling parameters were;

- Initial denaturation 5 minutes 95°C

Then 38 cycles of;

- Denaturation 1 minute 94°C
- Annealing (Tm) 1 minute 56°C
- Extension 1 minute 72°C

A final incubation step of 72°C for 10 minutes was included.

Preparation of 2ND UEP multiplex PCR mix

The same methodology as for the microsatellite multiplex was used.

2ND UEP multiplex restriction enzyme digestion

Digestions were performed in 384-well microtiter plates in a final volume of 8µl. Each reaction contained;

Reagent	Concentration/volume
PCR product	2µl
NEB buffer 4 (Biolabs)	1 X
Acetlyated BSA	0.01 µg/µl
*Hinf*I	0.32U
*Bsp*143I	0.32U
*Afl*III	0.32U
*Dra*III	0.32U
*Ssp*I	0.32U

The predicted sizes of the digested product together with its associated polymorphic status for each dye-labelled PCR product, were as given below;

Poly-morphism	Enzyme	Other enzymes that also cut	Labelled Primer	Dye Label	PCR product size	Size of labelled cut fragment
M9	*Hinf*I	*Afl*III, *Ssp*I	L	TET	214	48 (C)
M13	*Bsp* 143I		L	HEX	119	56 (G)
M17	*Afl*III		R	TET	124	101 (-G)
M20	*Ssp*I	*Afl*III	L	FAM	106	62 (A)
SRY 10831	*Dra*III		L	FAM	73	41 (G)

N.B. the letters in parentheses of the cut fragment size (bp) refer to the base that is present at the cut site. This was not an indication of whether the fragment contained the ancestral or derived form of the polymorphism.

Analysis of the 2^{ND} UEP multiplex PCR product

The same protocol as for the 1^{st} UEP multiplex was used.

Chapter 7

Statistical Methods

7.a. Population Genetic Statistics

Very many statistics can be applied to genetics. However, for the purposes of this study this chapter has concentrated on statistical issues involving polymorphic marker loci, from the basic summary of the population data to assessing complex population interactions assuming specific models of mutation.

7.a.i. Summary statistics

Basic summary statistics, such as;

i) Allele frequency distributions,

ii) Hardy-Weinberg equilibria,

iii) heterozygosity/ gene diversities and

iv) multidimensional geometric considerations,

Summary statistics describe population data at polymorphic loci without reference to any particular evolutionary model.

In addition to these summary statistics further analyses may be used to investigate the genetic differentiation between populations, for example; gene flow and fixation indicies, as well as molecular variances. Measures of genetic distances between populations incorporate various theoretical models to infer relationships between data sets or populations.

This chapter concentrates on the aforementioned statistics as applied to microsatellite data in the present study.

Summary statistics

i) Allele frequency distributions

Estimates of allele frequencies within a specific population, calculated using the 'gene counting' method, are a record of the number of times a specific allele is observed in a population, divided by the total number of alleles.

Graphical representations of allele frequencies express normal or multimodal distributions and are of use to describe basic differences and similarities between populations.

ii) Hardy-Weinberg equilibria

The Hardy-Weinberg principle was introduced in 1908 by two independent workers: G.H.Hardy and Wilhelm Weinberg. These researchers derived a mathematical equation that expressed the relationship between allele frequencies at a given locus and the frequencies of the resultant genotypes. The fundamental assumption governing this relationship was that the population studied was randomly mating, with negligible migration and immigration and no selection of genotypes. (Mc Conkey 1993).

To detect departures from the Hardy-Weinberg equilibrium (HWE) a variety of measures can be used including; a chi-square test, exact test (Guo and Thompsom 1992; Sjerps et al. 1995) and likelihood ratio test (Weir 1992). Likelihood – ratio tests have been considered preferable to the chi-square test when interpreting microsatellite data, as the high level of variation associated with the chi-square test reduces the statistical power (Hoelzel and Bancroft 1992).

The exact test (Guo and Thompson 1992) and likelihood ratio test (Weir 1992) were the first statistics to incorporate multi-allele loci.

For the purposes of this study, the exact test has been chosen for use with the genotypic data. Separate tests of HWE carried out at each locus, are analogous to Fisher's exact test on a two-by-two contingency table but extended to a contingency table of arbitrary size. A Markov-chain is implemented for use with such tables. Nei (1987) described the Markov chain as a mathematical process incorporating natural selection and random sampling of alleles from generation to generation, until the allele is either lost or fixed in the population.

These expected values can be compared to the observed values and a 'p' - value (or probability) estimated (Sjerps et al. 1995; Hartl and Clark 1989).

iii) Heterozygosity / Gene Diversity

Heterozygosity or gene diversity has been defined in terms of gene frequencies.

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

The gene diversity for a particular locus is hence;

m is the number of alleles

x_i is the frequency of the ith allele squared

However, this is not an unbiased estimate, as it does not incorporate the number of individuals sampled. Therefore, an unbiased estimate of *h* is given by:

$$\hat{h} = 2n(1 - \sum_{i=1}^{m} X_i^2)/(2n-1)$$

where, n is the number of individuals sampled (equation 8.4 Nei 1987).

Two sampling variances are associated with the estimated gene diversity. These are the *interlocus* and *intralocus* variances.

Intralocus variance

This variance is associated with the sampling of individuals at each locus. Essentially, it is the variance of the gene diversity among samples for a specific population at a specific locus, and is given by:

$$V_s(h) = (2n/2n-1)^2 V(\sum X_i^2)$$

This equation incorporates the variance of the sum of all the squared allele frequencies, where n is the number of individuals sampled.

Interlocus variance

This variance is associated with the sampling of loci from the genome and follows the equation:

$V(h)$=$V(h)$+$Vs(h)$

Which is the variance of the unbiased estimate of gene diversity equaling the sum of the interlocus ($V(h)$) and intralocus ($V_s(h)$) variances.

Thus the interlocus can be calculated by subtracting the intralocus variance from the total gene diversity variance ($V(h)$).

Hence, the variance of the unbiased estimate of gene diversity, and is given by:

$V(h)= \Sigma(h_j - H)^2/(r-1)$

Where is obtained by:

$H = \Sigma h_j j/r$

hj is the value of the unbiased estimate of gene diversity at the jth locus from r loci.

(equations 8.7 and 8.8 from Nei 1987).

iv) Multidimensional Geometric considerations

These geometric distances do not involve any evolutionary concepts other than similar allele frequencies that imply a closer genetic similarity. These distances may be regarded as a data reduction device so that comparisons of populations is kept as simplistic as possible (Nei 1987; Weir 1996)

$\hat{h}= 2\ n\ (1- \Sigma mXi2)/(2n-1)$

where, n is the number of individuals sampled (equation 8.4 Nei 1987).

Two sampling variances are associated with the estimated gene diversity. These are the *interlocus* and *intralocus* variances.

Intralocus variance

This variance is associated with the sampling of individuals at each locus. Essentially, it is the variance of the gene diversity among samples for a specific population at a specific locus, and is given by:

$Vs1(h)=(2n/2n-1)^2\ V(\Sigma^\wedge xi^2)$

Where $V(\Sigma^\wedge xi^2)$ is the variance of the sum of all the squared allele frequencies and n is the number of individuals sampled.

Interlocus variance

This variance is associated with the sampling of loci from the genome and follows the equation:

$V(\hat{h})=V(h)+Vs(h)$

Which is the variance of the unbiased estimate of gene diversity equaling the sum of the interlocus ($V(h)$) and intralocus ($V_s(h)$) variances.

Thus the interlocus can be calculated by subtracting the intralocus variance from the total gene diversity variance ($V(\hat{h})$).

Hence, the variance of the unbiased estimate of gene diversity, and is given by:

$V(\hat{h})= \Sigma(\hat{h}_j - \hat{H})^2/(r-1)$

Where $\hat{h}$ is obtained by:

$\hat{H} = \Sigma\hat{h}_j\, j/r$

$\hat{h}j$ is the value of the unbiased estimate of gene diversity at the *j*th locus from *r* loci.

(equations 8.7 and 8.8 from Nei 1987).

iv) Multidimensional Geometric considerations

These geometric distances do not involve any evolutionary concepts other than similar allele frequencies that imply a closer genetic similarity. These distances may be regarded as a data reduction device so that comparisons of populations is kept as simplistic as possible (Nei 1987; Weir 1996)

7.b. Measures of Genetic Differentiation

Statistics used to infer genetic structure within and between populations are not universal to all the discriminatory systems available. Indeed, statistics used for protein data are modeled according to the properties of the protein and hence may not be suitable for other systems, for example, variable number of tandem repeat (VNTR) data (Nei 1987). Therefore, the statistics proposed for use with microsatellite data will only be discussed further.

The informativeness of microsatellite loci is of use in the investigation of genetic structure within and between populations. However, factors such as genetic drift, admixture and fluctuating population size complicate the construction of the genetic relationships between populations (Bowcock et al. 1991).

Statistical models proposed to infer ancestral relationships from today's gene pool are used in conjunction with the physical remains of yesteryear uncovered by archaeologists and anthropologists. This provides a highly informative view of not only the genetic structure today, but also our links with ancestors.

The statistical analyses to be reviewed that measure genetic differentiation between populations include;

i) gene flow (Nm;Slatkin 1995),

ii) fixation index (Weir 1996), and

iii) the Analysis of Molecular Variance (AMOVA) (Schneider et al. 1997)

i) Gene Flow

Santos et al. (1997) used polymorphic microsatellite DNA markers to investigate gene flow in human populations. The estimated number of individuals exchanged per generation (Nm) is a measure of gene flow. High values of Nm were interpreted as high gene flow between geographically close populations. Gene flow influences the statistical inferences made about the ancestral relationships between populations. Hartl (1981) commented that relatively little gene flow is required to prevent significant genetic divergence among subpopulations due to random genetic drift.

Gene flow may be calculated in four different ways:

- Nei's method, where Nm is a function of the level of genetic differentiation, as estimated by GST. However, the GST estimate is used for protein data and not microsatellite data hence a bias may be present as an infinite alleles model (IAM) is employed instead of a stepwise mutation model (SMM) which is perhaps more informative as previously described for microsatellite data (Santos et al. 1997). The GST estimate used to obtain Nm values expressed higher results than using alternative methods such as the Rst index (Slatkin 1995). Santos et al. (1997) regarded this as a reflection that there were differing mutation rates and/or IAM was inadequate for microsatellite systems.
- The private-allele method incorporates allelic frequencies from populations as a function of the migration rate (Nm). However, was not sensitive to the number of demes or subpopulations within a sample.
- A qualitative measure of the different classes of alleles encountered in a set of populations.
- A method developed to use the stepwise mutation model, based on the Rst index (Slatkin 1995). Here Nm is calculated as

Nm =[(d-1)/4d][(1/Rst-1],

d is the number of populations.

However, Santos et al. (1997) argued that there was no strong evidence that all microsatellite loci follow the stepwise mutation model. In particular it was noted that dinucleotide repeat loci do not consistently follow the SMM. However, tetranucleotide markers express a greater potential to follow the SMM as there was a lower potential for enzyme slippage, in comparison to dinucleotide markers and tetranucleotide loci were not observed to be readily linked to disease markers unlike some trinucleotide repeats.

The limitations of Nm were made clear. If populations are not in equilibrium, inaccurate gene flow estimates will be calculated. Recent populations, severe bottlenecks, high rates of extinction and recolonization all contribute to such inaccuracies (Santos et al. 1997).

Therefore, it is necessary to first detect the equilibrium status of the populations studied before inferences on gene flow can be definitely made (Santos et al. 1997).

7.c. Fixation Index

The fixation index symbolized as 'F_{ST}' measures the effects of population subdivision resulting in a reduction of heterozygosity of a subpopulation due to random genetic drift. F_{ST} was used to determine whether natural selection or neutral variation was present in populations. Neutral variation is the product of drift. and one expects drift to be equal for all genes. Drift is explained to depend only on demographic properties of the populations and not on the particular gene being studied. F_{ST} will vary from gene to gene, but the extent of variation will be predictable (Bowcock et al. 1991).

F_{ST} can further be anticipated in terms of identity by descent (Hartl 1981). One may observe that inbreeding may be reflected in the statistic as substantial, although mating has been random. Therefore the observed inbreeding may be the consequence of a small population size and not through non-random mating (Hartl 1981). Natural selection favors particular alleles or genes in some environments. Hence F_{ST} for this allele/gene is expected to be higher than for a allele/gene affected by drift alone. However, natural selection may favor the heterozygote more than the homozygote, and F_{ST} will be lower than average (Bowcock et al. 1991).

To qualitatively guide one through the interpretation of F_{ST} results; a range of 0 to 0.05 F_{ST} may be considered as an indicator of little genetic differentiation, 0.05 to 0.15 as moderate differentiation, 0.15 to 0.25 as great differentiaiton and above 0.25 as very great differentiation (Hartl 1981).

Interestingly, of the total genetic variation found in three major races (i.e., caucasoid, negroid and mongoloid), only 0.07 (7%) was ascribed to variation between populations and 0.93 (93%) was observed within populations (Hartl 1981).

7.d. Variance

The F-statistic computations involve analysis of variance of allele frequencies and have been adapted for use with microsatellite data (Michalakis and Excoffier 1996). Between group and within group variances can be established. Michalakis and Excoffier (1996) incorporated Slatkins (1995) Rst measure in terms of mean squared deviations of between and within populations (Goldstein and Schlotterer 1999 Pp114). The interlocus and intralocus variances associated with gene diversity have previously been discussed (see heterozygosity / gene diversity section this chapter) (Nei 1987).[1]

7.e. Measures of Genetic Distances

There are two major models evoked by population geneticists to explain mutation properties and their subsequent effect on allele frequencies, these are the infinite allele model and the stepwise mutation model.

7.e.i. Infinite allele model (IAM)

The infinite allele model (IAM) was first proposed in 1964 by Kimura and Crow. This model assumed that there was no migration of genes, but every mutational event formed a new allelic type and genetic drift was the prominent force (Weir 1996; Hartl and Clark 1989). Equilibrium was established between the loss of variation by drift and the introduction of variation by mutation (Weir 1996).

The use of the IAM with microsatellite data is questionable. Microsatellite repeats have been found to mutate to a longer or shorter length not in keeping with the IAM (Goldstein et al. 1995). Furthermore, the IAM does not control for natural selection, gene flow between populations and stochastic events (Santos et al. 1997).

Nei's (1972) standard genetic distance assumes an IAM. The basic principal is that any difference in electrophoretic mobility or immunological reaction is caused by a mutation at the gene level. The Standard genetic distance is defined as

$D=-\log_e I$, where I is the extent of genetic similarity between populations, known as normalised genetic distance or genetic identity. This however, is a biased test, as it has a theoretical inference that the genetic distance is calculated in terms of allele frequencies for all loci in the genome (Nei 1987). In the year 1978, Nei proposed the 'unbiased estimate' for genetic distance which incorporated the realistic situation of sampling a proportion of the population and testing these samples at a proportion of all loci within the genome (Nei 1987).

[1] Further information regarding variances has been included in the Y-chromosome section on analyses of molecular variance (AMOVA), this chapter.

7.e.ii. Stepwise mutation model (SMM)

The stepwise mutation model (SMM) proposed by Ohta and Kimura (1973), assumes mutations to occur in discrete steps altering the electrophoretic mobility of an allele by one unit (Hartl and Clark 1989; Rousset 1996). In the first instance this model was used for the electrophoretic charge on protein polymorphisms and later was used as a model of evolution of microsatellite allele sizes (Rousset 1996). The SMM model has been found better suited to analysing genetic distances of microsatellite multiallelic data. It should be noted that the assumptions of the SMM differ sharply from the assumptions of the IAM. Hence distances designed to increase linearly under the IAM, such as Nei's standard distance, are both nonlinear and supposedly inaccurate for microsatellite loci (Goldstein et al. 1995; Takezaki and Nei 1996).

7.e.iii. The Coancestry / Coalescent Model

Understanding the ancestry among genes in a sample, the ‘genealogical processes’, can be best explained using a 'coalescent' model, together with a stepwise mutation model (Wilson and Balding 1998). This model incorporates a complex Markov chain Monte Carlo (MCMC) simulation algorithm.

The Markov chain, describes and predicts allele frequency distributions among populations and the ‘drift’ of allele frequency (or transition probability) from generation to generation (Hartl and Clark 1989; see also section on Hardy-Weinberg equilibria this chapter).

The coalescent assumes neutrality, random mating and a constant large population size. The coalescent also indicates that all alleles present in the population at any particular time can be traced back to one common ancestral allele (Hartl and Clark 1989).

Time in the coalescent model is measured in units of N generations, where N is the (fixed and large) population size. The time until the first coalescence of two lineages at a common ancestor has an exponential distribution with mean 2/n(n-1). The time between the first and second coalescences has an exponential distribution with mean 2/(n-1)(n-2). This continues until time tn-1, where exactly two ancestors coalesce, here the exponential distribution has mean of 1 (Weir 1996).

If a genetic bottleneck (a severe temporary reduction in population size) forces a very low level of polymorphism, (for example emigrants from an established population form a new subpopulation) the accompanying random genetic drift is termed the 'founder effect' (Hartl and Clark 1989).

In such cases as the founder effect, there may also be unequal numbers of males and females. This bias in sex ratio may also enhance random genetic drift. Weir (1996) notes that the coancestry distance measure is appropriate in such instances when estimating divergence due to drift. However, one should note that the coancestry distance has no assumptions about the ancestral population (Weir 1996).

7.f. F_{ST} as a coancestry coefficient

F_{ST} is described as the coancestry coefficient (Weir 1996), used in Slatkin's RST distance measure (Schneider 1998; Slatkin 1995).

R_{ST}, is a measure of population subdivision based on microsatellite allele frequencies (Slatkin 1995). It considers a demographic model where two haploid populations diverged from a population identical in size. The two populations then remained isolated without immigration or emigration (Schneider et al. 1998). R_{ST} can be used to estimate the effective migration rate between populations (Freimer and Slatkin 1996). The F_{ST} is expressed, (using the analysis of variance approach) in terms of the coalescence times. F_{ST} on its own, does not apply to microsatellite data as the F_{ST} model assumes low mutation rates applicable for allozyme data but not microsatellites (Slatkin 1995). Slatkin (1995) further compared F_{ST} to R_{ST} and found estimates based on F_{ST} expressed too much genetic similarity with a large difference in coalescence times is predicted. In comparison, estimates using R_{ST} expressed no bias. In general, the better fit of the R_{ST} model to microsatellite data in comparison to the F_{ST} model was due to the R_{ST} statistic being based on the stepwise mutation model (Slatkin 1995).

7.g. Average Square Distance (ASD) and D1 distance measures following the Stepwise Mutation Model

Goldstein et al. (1995) developed a genetic distance based on the stepwise mutation model (SMM), that incorporated allelic repeat scores. The reliability of the distance measure was evaluated with computer simulations and comparisons made with allele sharing and Nei's distance that followed the IAM. It was found that no one distance measure could precisely reconstruct genetic relationships between populations / taxa , but for phylogenetic reconstruction of taxa that are sufficiently diverged the distance measure proposed by Goldstein et al. (1995) was preferable. Goldstein et al.'s and (1995) and Slatkin's (1995) distance measures (ASD or D1) increase linearly with time, following the unconstrained SMM model. However, there were problems associated with its use. It had a high variance, caused partly by its dependence on the variation within populations. Moreover, due to the population sizes varying among taxa in any phylogeny, the inclusion of the intrapopulation variance term obscured the relationship between separation time and the observed value of ASD (Goldstein et al. 1995).

The genetic distance 'D1' averages the squared difference in repeat numbers for two alleles drawn one each from different populations isolated r generations in the past.

The D1 distance measure was compared to the allele sharing model and Nei's distance. After about 1000 generations has passed, both the allele-sharing model and Nei's distance are beginning to asymptote, but D1 remains linear (Slatkin 1995).

The estimation of 'D0', averaged the squared difference in repeat numbers for two alleles drawn from the same population.

The between-individual component of the sum of squares reflected either 'D0' or 'D1', dependent on whether the two individuals came from the same or different populations (Bowcock et al. 1994). A matrix of these between-individual squared differences therefore contained elements of either D0 or D1. If these estimates were sufficiently different, a clustering program would group individuals into their correct populations, an approach taken by Bowcock et al. (1994) who used a distance measure based on the proportion of shared alleles between individuals. This suggested that microsatellite loci could be used to assign individuals to the populations from which they came.

7.g.i. Goldsteins $(\delta\mu)^2$ distance measure

The distance measure delta mu squared $[(\delta\mu)^2]$ (Goldstein et al. 1995) has been specifically designed to overcome problems associated with large variances associated with genetic distances. Delta mu squared conveniently simplifies to $(mx - my)^2$ where mx and my are the means of allele sizes in populations x and y. This model has been shown to be fairly robust to changes in population size, defined in terms of allele frequencies (Takezaki and Nei 1996). Delta mu squared based distances are not sensitive to levels of allelic variance in the absence of range constraints, however, with range constraints, loci with the lowest mutation rates will remain accurate longer.

It should be noted that if population size changes linearity will be lost until the within population variance assumes a new equilibrium value (Goldstein et al. 1995).

7.h. Phenogram construction methods

The phenogram construction methods of the Unweighted Pair Group Method Arithmetic mean (UPGMA) of Sneath and Sokal (1973) and the Neighbor Joining Method (NJ) of Saitou and Nei (1987).

7.h.i. The average distance method or UPGMA method

The unweighted pair-group method with arithmetic mean (UPGMA) was originally developed for constructing a phenogram (Sokal and Michner 1958) but has also been used for phylogenetic trees (Nei 1985). The UPGMA method assumes that the expected rate of gene substitution is constant. If the measure is linear with evolutionary time without error, it gives the correct topology and correct branch lengths of the genetic relationship between populations / taxa.

The UPGMA is constructed by the sequential clustering of populations with the smallest genetic distance. The two closest populations cluster with the lengths of the branches assumed the same. New distances are then calculated between the clustered and the remaining populations. The underlying assumption of the UPGMA method is that the expected rate of gene substitution is constant (Nei 1987).

7.h.ii. The Neighbor-Joining method (NJ)

The NJ method was developed by Saitou and Nei (1987) as a method for estimating phylogenetic trees.

This method does not obtain the shortest possible tree for a data set, but a tree that is as close as possible to the 'true' phylogenetic tree (Rholf 1993).

To construct the tree, the closest pair of distances merge which results in the greatest reduction in length of the graph (Rholf 1993). Whereby, a pair is defined as two units joining at a single node (Weir 1996). At each node several equally close neighbors may exist, hence more than one tree may be constructed (Rholf 1993).

Y Chromosome Statistics

7.i. Haplotypic Data Analysis – Specific to diallelic and STR Loci

Polymorphic microsatellite loci have been used to analyse the genetic ancestry of not only autosomal chromosome regions but also male (Y chromosome) and female (mitochondrial DNA) lineages. The linked polymorphic loci are analysed and together the alleles form a haplotype. Once different haplotypes have been identified the frequencies of the haplotype (not each locus) are calculated in the same way as the allele frequencies of autosomal marker systems.

The statistical analyses performed on haplotypic data include; gene diversity (similar to the autosomal diversities), different hierarchical Analyses of Molecular Variance (AMOVA) which evaluates population genetic structure, pairwise genetic distances, FST based genetic distances for short divergence time (as previously mentioned), exact test of population differentiation and network analyses (Schneider et al. 1997).

These analyses have been included in the present study to determine male lineage relationships within and between Polynesian and U.K. Leicestershire populations and so are discussed further.

7.j. Gene Diversity

The gene diversity for haploid data as stated in the 'Arelquin' computer program, is defined as the probability that two randomly chosen haplotypes are different in the sample data set. The gene diversity and sampling variance follows Nei's equations 8.5 and 8.12 (1987) and are given as:

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

$$V(\hat{H}) = \frac{2n}{n(n-1)}\left\{2(n-2)\left[\sum_{i=1}^{k} p_i^3 - (\sum_{i=1}^{k} p_i^2)^2\right] + \sum_{i=1}^{k} p_i^2 - (\sum_{i=1}^{k} p_i^2)^2\right\}$$

Where *n* is the number of gene copies, *k* is the number of haplotypes and P*i* is the frequency of the i-th haplotype.

The difference between these statistics and the autosomal gene diversities, lie in the number of gene copies in the sample. The autosomal unbiased estimates of gene diversity and the sampling variance incorporate 2n (the diploid condition) and the Y-chromosome data incorporates n (the haploid condition) (Nei 1987).

7.k. Analyses of Molecular Variances (AMOVAs)

The variance values obtained using this method reflect the proportion of the molecular variance explicable in terms of population differences of haplotype frequencies, hence an index of population differentiation (Hagelberg et al. 1999).

The variance components are estimated on three hierarchical levels; within populations within groups, between populations within groups and between groups. The significance of the results is attained via comparison of observed values to the empirical null distribution attained from 1,000 randomizations or permutations (Kittles et al. 1998).

7.k.i. Analysis of variance for haplotypic data:- one group of populations

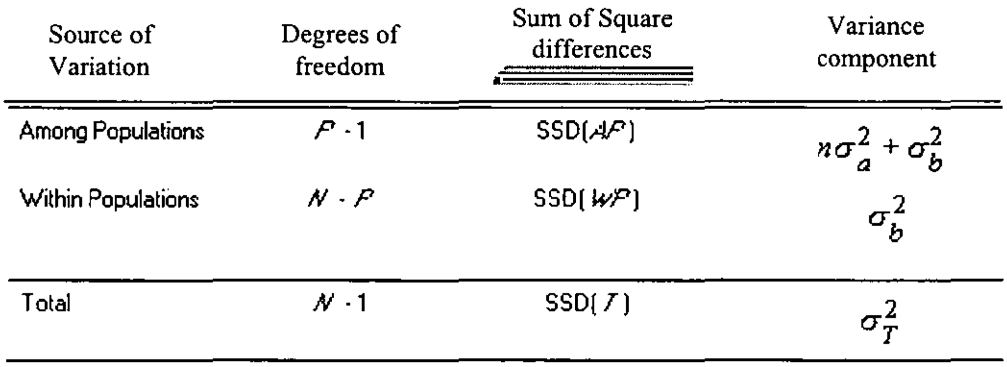
The Arlequin program incorporates variance statistics for haplotypic data. Variation is calculated among and within populations for one group of populations, also among populations within groups for several groups of populations. The among population variance component and F_{ST} are tested by permuting haplotypes among populations (see table 7.1).

Analyses of variance (Excoffier and Smouse 1994) use a simple hierarchical model of population genetic structure.

Assumptions of the properties of the haplotypes within the population are that they are; additive, random, independent and have associated variance components (expected squared deviations) within and among populations (Michalakis and Excoffier 1996). The differences among the haplotypes are assumed to arise by point mutations and not recombination as would be more frequent with autosomal loci (Excoffier and Smouse 1994).

TABLE 7.1: NOTATION USED BY ARELQUIN IN THEIR ANALYSIS OF MOLECULAR VARIANCE OF HAPLOTYPIC DATA FOR ONE GROUP.

[P is the total number of populations, N is the total number of gene copies, SSD is the sum of squared deviations, AG is among groups, WP is within population and T is the total].



Source of Variation	Degrees of freedom	Sum of Square differences	Variance component
Among Populations	$P - 1$	SSD(AP)	$n\sigma_a^2 + \sigma_b^2$
Within Populations	$N - P$	SSD(WP)	σ_b^2
Total	$N - 1$	SSD(T)	σ_T^2

Where n and Fst are defined by;

$$n = \frac{N - \sum_{p}^{N^{2}\frac{p}{N}}}{P-1} \qquad F_{st} = \frac{\sigma_a^2}{\sigma_T^2}$$

7.l. Network Construction

Direct relationships exist between haplotype frequencies and their time since common ancestry. High frequency haplotypes have probably been present in the population for many generations, thus accruing substantial copy numbers (Excoffier and Smouse 1994). In general, the new mutants are derived from the common haplotypes and so are more closely linked to these than other rare variants (Excoffier and Smouse 1994). Evolution embraces emigration and geographic diffusion of the gene pool. Thus a complicated network of haplotypes can be sampled within any defined population.

Minimum spanning unique event polymorphism networks and microsatellite networks of specific haplogroups can be constructed (Hurles et al. 1998; Excoffier and Smouse 1994). These assume a single-step mutation process, by linking haplotypes differing by a single mutation and then 'adjacent' haplotypes (one repeat difference over the six microsatellite loci) until all the haplotypes are incorporated into the network (Hurles et al. 1998).

7.m. Coancestry Coefficients for haploid populations

For the purposes of this thesis, coancestry coefficients were constructed using the Y chromosome haplotype data and the computer program 'Arlequin'. The coancestry coefficient was calculated incorporating the F_{ST} measure as;

$F_{ST} = 1 - (1\text{-}1/N)^t \cong 1 - e^{-t/N}$

Here F_{ST} between pairs of haploid populations is a function of population size N with a divergence of t generations ago. The genetic distance

$D = -\log(1 - F_{ST})$ is therefore approximately proportional to t/N for short divergence times.

7.n. Dating the common ancestry of 'modal' haplotypes

This dating method follows the work of Thomas et al. (personal communication). The common ancestry date (*t*) in generations for a cluster of closely related Y chromosomes, centred around a modal haplotype, is assessed by calculating the average squared difference (ASD) in allele size between all chromosomes included the cluster and the ancestral haplotype (assumed to be the same as the modal haplotype), averaged over all loci. This has the expectation μt where μ is the microsatellite mutation rate (Goldstein et al. 1995, Slatkin 1995).

A point estimate of μ=4/3155=0.00127 (Thomas et al. personal communication), based on data from 3 published studies (Heyer et al. 1997, Kayser et al. 1997, Bianchi et al. 1998), is used.

The point estimate of the mutation rate based on previous studies (Heyer et al. 1997, Kayser et al. 1997, Bianchi et al. 1998), was restricted to the same microsatellite loci as those used in the present study (giving 4 observed single-step mutations in 3155 separate meiotic events).

In order to test the reliability of the estimate, a narrow confidence interval (CI) on t can be calculated according to the method of Thomas et al. (1998). This method uses 100,000 replications to compute the confidence interval. The interval assumes μ is known without error and reflects sampling variance of mutations on a star genealogy only. Therefore, only modal haplotype clusters can be dated with any accuracy using this method.

7.o. Summary of the choice of statistical methods

Chu et al. (1998) note that the use of microsatellites in the reconstruction of closely related populations is problematic due to the large numbers of loci that need to be examined. If the variance of the genetic distance between loci is larger than the variance due to sampling error in the estimation of genetic distance, then a sufficient number of samples have been examined (Chu et al. 1998).

However, if the number of repeats at any locus is restricted then the accuracy and linearity with time of all distances will be strongly affected (Feldman et al. 1996; Goldstein et al. 1995). If one is to assess divergence times accurately the locus range and mutation rate need to be precise. The most accurate distance measurements in recently separated populations are those that use the product of allele frequencies, however in this case distance is not linear with time (Takezaki and Nei 1996). Goldstein and Pollock (1997) note that if a distance is used to estimate relative times of divergence, then the expectation should increase linearly with time and the coefficient of variance ideally should be low.

Chu et al. (1998) observed that the microsatellite markers used in their study would only detect major genetic contribution from particular sources and a haplotype-based analysis may have detected minor genetic contribution from closely related populations. Finally it was also commented on that phenogram constructions the neighbor-joining method was supposedly more robust in the presence of genetic admixture.

Forensic statistics summarised

The statistics used in forensic analyses are modified extensions of population genetic statistics (Saferstein 1988). Forensic genetics incorporate statistics that directly reflect the present population structure without any reference to the prehistory or evolutionary population structure.

This section of the chapter concentrates on the widely used and accepted forensic statistical analyses, including;

- testing population structure using Hardy-Weinberg expectations,
- discrimination power of loci,
- matching probability,
- polymorphic information content and
- paternity identity (Saferstein 1982, Weir 1990, NRC 1996, Sjerps et al. 1995).

In criminal investigations involving DNA evidence a number of key questions are posed;

i) Does the recovered DNA sample match a suspect and if so,

ii) was this sample left by the suspect or left by another with the same DNA profile.

In order to assess ii), one must assess the frequency with which the profile occurs in the general population. Here, knowledge of the population structure is used, based on population genetics.

7.p. Hardy-Weinberg Equilibria

Hardy-Weinberg expectations were discussed in the population genetic section of this chapter and so will not be re-introduced here.

7.q. Discriminatory Forensic Statistics

The value of genetic markers for the discriminatory potential between individuals is assessed using the probability of identity statistic (Pi). The 'Pi' measures the probability that two individuals selected at random will be identical at that locus.

Pi is equated using the formulae;

$Pi = \sum x_i^2$

Where, xi is the frequency of the ith phenotype/genotype in the polymorphic system.

Pi values can be cumulated to include other polymorphic systems by multiplying the Pi values for each system. This process can be used to assess the potential of alternative analytical protocols (Sensabaugh 1982).

Power of exclusion = $h^2(1-2*h*H^2)$, as described by the Powerstat Excel package, following the protocol of Brenner and Morris (1990), incorporates the percentage of homozygosity (h) and heterozygosity (H). This statistic does not use genotypic or expected frequencies, instead a direct measure of the population structure is measured.

The power of exclusion can be extended to incorporate more than one locus. If there are n loci, and the sum of the squares of the genotype frequencies at locus I is Pi, then the exclusion power is 1- (P1, P2, P3, ... Pn) (NRC 1996).

Microsatellite loci show varying allele frequencies between different racial or geographic populations. Using the Pi value, a statistic has been developed to quantify the informative a marker system is in differentiating between populations.

Here, the probability of selecting two individuals at random one from population X and the other from population Y, who have the same phenotype is given as Pi(X,Y) = Σ xiyi, where xi and yi are the frequencies of the ith phenotype. If the populations are nearly identical then Pi(X,Y) approaches the individualization potential values for X and Y;

Therefore,

D = ((Pi (X) + Pi(Y))/2) - (Pi(X,Y))

The larger the index value, the greater the racial differentiation potential of the marker system (Sensabaugh 1982).

The aforementioned statistics do not assume any prior knowledge of ancestral affiliations and only account for the population structure and relationships in the present gene pool.

7.r. Matching probability

The match probability is calculated from the frequencies of DNA markers given in the database of samples. Thus dependent on the size and ethinic origin of the database, the match probability may vary (NRC 1996).

For the purposes of this study, the match probability was calculated using the 'powerstat' macro, following the protocol of Brenner and Morris (1990).

A frequency table is constructed for genotypes, as shown below;

	Alleles (given in base pairs)				
	100	104	108	112	116
100	2		1	2	
104		1	3		2
108				5	
112					
116					

The average genotype frequency squared, is calculated by;

[genotype number/Total sample number]^2

Finally, these figures are added together giving the 'Matching Probability' for a specific locus. In this way many loci can be compared to each other to ascertain the loci systems with the greatest discriminatory potential for use in Forensic analytical systems.

7.s. Polymorphic Information Content

Is used as an indicator of the usefulness of the marker in discriminatory systems. The equation, given by Botstein et al. (1980) is;

$$1-\sum_{i=a}^{n}\sum_{j\geq i}^{n}P_{ij}^{2}-\left(\left(\sum_{i=a}^{n}\sum_{j\geq i}^{n}P_{ij}^{2}\right)^{2}+\sum_{i=a}^{n}\sum_{j\geq i}^{n}P_{ij}^{4}\right)$$

Essentially this is a summation of squared genotype frequencies (Pi) and a classic tool in the description of a particular marker system (Botstein et al. 1980).

7.t. Paternity Index

Within forensic science if a putative father is not excluded by serological techniques, for example, blood-group and enzyme analyses then a paternity index is calculated (NRC 1996). The paternity index makes use of an appropriate ethnic database to calculate the probability that the putative father is the true father against any other male in the population (Saferstein 1988).

Included within this thesis is the typical paternity index, which has been calculated using the 'Powerstat' Excel package, following the equation given by Brenner and Morris (1990).

Typical Paternity Index = H+h/2H

Where H is percentage heterozygosity and h is percentage homozygosity (Brenner and Morris 1990). This typical paternity index calculation allows a quantitative measure of how discriminatory a locus is in determining paternity. Therefore, it follows that the more loci studied increases the potential for discriminating between 'alleged' fathers and the ability to exclude individuals from the investigation.

7.u. F_{ST} revisited: the Forensic aspect

Gill and Evett (1995) and Balding et al. (1996), investigated the population genetics of STR polymorphisms for use in Forensic science laboratories. Wrights F statistics were used to describe the inbreeding coefficient (F_{IT}), coancestry (F_{ST}) and average within population inbreeding coefficient (F_{IS}). These workers believe the coancestry coefficient or $F_{ST,}$ is a key population genetics parameter. F_{ST} correlates two alleles sampled from distinct individuals within a subpopulation. Using a quadraplex marker system, Gill and Evett (1995) observed greater divergences between than within ethnic group comparisons with an F_{ST} estimated at 0.002 - 0.003.

In forensic casework the significance of a matching profile is described in terms of a 'likelihood ratio' (LR). A likelihood-based method of estimating F_{ST} was purported to be of more use than other methods, as it allowed a range of plausible values to be assessed rather than a single point estimate. It also allowed a subpopulation to be compared to a larger population (database) required for forensic work (Balding et al. 1996).

The likelihood based probability has two hypotheses, either the crime sample was left by the suspect or, the crime sample was left by an unknown person. If STRs are used as the marker system of choice and independence is assumed, the LR is the inverse of the relative frequency of the observed genotype in the relevant population (Gill and Evett 1995). As an example; given a LR of 1,000, the probability that the profiles are the same is 1,000 times as great if the samples came from the same person as opposed to someone else (National Research Council Report 1996). However, Balding et al. (1996) observed that combining loci information to produce a LR, may falsely assume that the mutation rates of the loci are equal, hence creating a bias in the results.

Weir (1994) assumed a LR as the inverse of the frequency of the genotype;

LR = 1/PA, where PA is the genotype frequency.

However, it was quickly observed that this ignored the effects of inbreeding and relatedness. Furthermore, the effect of relatedness was greatest when small allelic frequencies and high values of F_{ST} were observed. In such instances, a low LR was recorded.

Thompson (1995) reviewed the use of likelihood ratios to present forensic DNA evidence in the courtroom, with some disturbing results. The likelihood ratio was reduced when the exact DNA profile was uncertain and laboratory errors were found to exist. In particular, with the occurrence of extra bands in an evidentiary sample that was not observed in the accuseds profile. Either the additional bands in the evidentiary sample profile were artifacts, or, the evidentiary sample contained a mixture of DNA from more than one person. In situations such as these, it is important to assess both thesc theories subjectively incorporating arbitrarily assigned probabilities concerning the relationship between the victim and defendant and the victim's background (Thompson 1995).

7.v. Forensic Analyses using the Y chromosome

The Y chromosome has a place in forensic and paternity studies, although its use would be select and very specific. The obvious application of haplotype analyses would be the selective PCR of male/female fractions of DNA in rape cases complementing the discriminatory autosomal marker systems. However, using the Y chromosome does have its disadvantages. Firstly, substructuring within populations is such that within any male lineage many males will share the same Y chromosome haplotype. This is because little recombination exists other than at the pseudoautosomal regions of the Y chromosome (Jobling et al. 1997; see also Chapter 2). Hence the exclusion of males is easier when the profiles are clearly different. However, when the profiles are similar or the same, further analyses incorporating autosomal loci would be carried out, to further include or exclude the suspect with certainty (Jobling et al. 1997).

Chapter 8
The Autosomal STR Variation:
The Results

8.a. Introduction

This results chapter statistically evaluates 10 autosomal polymorphic tetranucleotide short tandem repeat (STR) loci individually and together among five defined populations. Numerous statistical analyses have been carried out using Genetic Data Analysis (GDA), and NT-SYSpc version 1.80 (Rholf 1993) software packages. In some instances two programs calculated similar analyses, for completeness these have been compared and any differences noted. All of the analyses included here have been used by other researchers analysing STR data and have been referenced accordingly.

The statistical analyses have described how informative the loci are singularly and as a collective for use as potential population/forensic genetic markers. The genetically diverse U.K. Leicestershire population (Mastana and Sokol 1998) provided an interesting comparison to the purported less diverse Polynesians (Hurles et al. 1998, Murray-McIntosh et al. 1998). The known prehistory of the U.K. Leicestershire and Polynesian populations, together with a clear cultural and geographical divide, provided an a priori test to evaluate if the 10 markers could clearly distinguish between the populations.

The Polynesian populations were separated by the individuals own admission (Chambers personal communication) into; admixed Polynesian Islanders, non-admixed Polynesian Islanders, admixed Maoris and non-admixed Maoris.

The STR markers have also been investigated to asses whether the admixed and non admixed populations can be distinguished.

It was not the intention of this study to trace the evolutionary origins of the populations. At most, the relationship between the populations was observed and compared to existing referenced investigations of known evolutionary / migrational 'trends', where appropriate, in Polynesia and the U.K.

The majority of the purported admixture in the Polynesian samples was European. Therefore, one could hypothesize that, if the loci were sufficiently discriminatory then the admixed Polynesian populations may be observed to have a closer similarity to the U.K. Leicestershire population, than the non-admixed Polynesian populations.

Abbreviations of Population names used in the results sections:

U.K. = U.K. Leicestershire Caucasian

AM = New Zealand Maoris with known 'foreign' admixture

NAM = New Zealand Maoris no reported admixture

NAI = Polynesian Islanders with no reported admixture

AI = Polynesian Islanders with known 'foreign' admixture (mainly European in origin)

8.b. Basic Descriptive statistics

The number of samples per locus used in the statistical analyses varied (Table 8.1). This variation was the direct result of the success of either amplification procedures or the inability to correctly designate an allele size with absolute certainty. 'Spreadex' gels which could not be 'read' clearly were either re-run or discounted (see section on materials and methods). All the raw data has been included in the appendix.

TABLE 8.1. ALLELE NUMBERS FOR EACH POPULATION AT 10 AUTOSOMAL STR LOCI.

	U.K.	AM	NAM	AI	NAI	TOTAL
D10s520	98	94	128	26	46	392
D12s297	84	74	112	30	50	350
D1s407	92	64	82	20	30	288
D2s262	66	76	112	26	44	324
D3s1514	108	80	110	24	36	358
D4s2285	98	106	124	22	48	398
D5s592	100	96	110	26	42	374
D7s1485	40	96	140	26	52	354
D7s618	102	100	128	30	56	418
D9s252	74	104	130	28	48	384

8.c. Repeat sequence motif of the 10 autosomal STR loci

The Utah Marker development group isolated and sequenced each tetranucleotide marker, although had not carried out detailed population genetics on the systems. The sequence containing the repeat motif is listed in Table 8.2 (i). DNA samples from this study were also sequenced at AltaBioscience (University of Birmingham) providing not only a comparison to the Utah Marker Development groups sequence information but a study to determine the base sequence of spurious bands (a suspected heteroduplex formation) observed whilst analysing PCR products. Comparisons made between alleles sequenced by the Utah Marker development group and the *AltaBioscience* company highlighted variations not only of repeat number but also base substitutions, insertions and deletions (all raw sequence data for the present study is given in the appendix).

TABLE 8.2: COMPARISON OF LOCI SEQUENCE STRUCTURE FROM THE UTAH MARKER DEVELOPMENT GROUP AND ALTABIOSCIENCE

(i) repeat sequence motif as sequenced by Utah Marker development group(Ballard personal communication)

(ii) repeat sequence motif as sequenced by AltaBioscience (see appendix for sequence information)

Locus	Repeat sequence motif
D1S407	(i) (AGAT)2(AGGA)(AGGT)2(ACAT)(AGAT)8
	(ii) (AGAT)2(AGGT)4(ACAT)(AGAT)11
D2S262	(i) (AAAG)14(A)(AAAG)3
D3S1514	(i) (AAAG) 2 (AGAG) 3(AAAG) 3 (AG)(AAAG)12
	(ii) (AAAG)3(G)(AGAG)2(G)(AAAG)2
D4S2285	(i) (AAAG)3(AAGG)2(AGGG)6(AAGG)2(AAGC)
	(ii) (AAAG)(AAGG)3(AGGG)5(AAGG)9(AAGC)
D5S592	(i) (AGAT)12
	(ii) (AGAT)11
D7S618	(i) (AAAG)14(AA)(AAAG)4(AGAG)(AAAT)
	(ii) (AAAG)13(AA)(AAAG)4(AGAG)(AAAT)
D9S252	(i) (AGAT)9(AATT)(AATA)(CAT)(AGAT)(GAT)
	(ii) (AGAT)11(AATT)(AATA)(CAT)(AGAT)2(GAT)
D10S520	(i) (AAAG)4(AAAT)(AAA)(AAAG)4(AA)(AAAG)11(AAAC)6
	(ii) (AAAG)4(AAAT)2(AAAG)5(AG)(AAAG)5(AAAT)(AG)(AAAG)5
D12S297	(i) (AGAT)3(GGAT)(AGAT)4(GATG)(GAT)(AGAT)3(GTT)(AGAT)4
	(ii) (AGAT)3(GGAT)(AGAT)3(GATG)(GAT)(AGAT)3(GTT)(AGAT)4
D7S1485	(i) (AAGG)15(AAAA)(AGAA)2(AAAG)
	(ii) (AAGG)12(AGGG)(AAGG)(AAAA)(AGAA)2(AAAG)

In general the longer the allele the more complex the sequence. The AAAG motif was observed at four loci; D2S262, D3S1514, D7S618 and D10S520. The AGAT motif was observed at four loci; D1S407, D5S592, D9S252 and D12S297 and finally the AAGG motif was only observed at two loci; D7S1485 and D4S2285.

The simplest repeat sequence was at locus D5S592 with an allele length range of 162-198 base pairs. Similarly, the most complex repeat sequence was observed at locus D4S2285 with an allele range of 265-305 base pairs (refer to the appendix for raw sequencing data).

8.d. A comparison of sequenced alleles between the Utah Marker Development group and the present study

D1S407: Other than the actual repeat number difference, the allele sequenced by the Utah Marker development group (UMDG) had an AGGA insertion that was not observed in an allele of the present study sequenced by AltaBioscience.

D2S262: A positive sequence result was only obtained by the Utah Marker development group (UMDG), as sequencing was unsuccessful at Alta Bioscience, hence there were no sequences for comparison.

D3S1514: Variation not only in the number of AAAG motifs was observed, but also the AGAG motif inserted among the AAAG motifs. It was also interesting to observe an AG insertion in the Utah Marker Development groups' sequenced allele and a G insertion in AltaBiosciences sequenced allele.

D4S2285: Similarly to D3s1514, both the UMDG and AltaBioscience sequences contained two motifs (AAGG and AGGG) which were tandemly repeated, although the AAGG motif was more frequent. Complex sequence variations were observed either side to the central core of repeats, consisting of base substitutions and deletions/insertions.

D5S592: The UMDG and AltaBioscience sequences isolated a simple clear repeat structure with little variation.

D7S618: The repeat structure in both alleles was the same other than the initial AAAG repeat number as expected between different sized alleles.

D7S1485: The UMDG and AltaBioscience sequenced alleles had a similar sequence. However, the second allele sequence (AltaBioscience, this study) had a base substitution in the AAGG repeat motif to AGGG. This tetranulceotide separated the tandem AAGG repeat motifs.

D9S252: A similar sequence was observed between the two sequenced alleles. The only difference between the two alleles was the number of tandemly repeated AGAT motifs, as one would observe from different sized alleles.

D10S520: The two alleles sequenced at this locus had quite different repeat patterns. The principle repeat motif was AAAG. The first allele had an 'AAA' and an 'AA' insertion, between the AAAG repeats. In addition, a secondary repeat motif AAAC, was sequenced that differed from the original motif by a base substitution from G to C.

The G to C substitution was not observed in the second sequenced allele, instead an 'AAAT' motif was isolated, indicating a base substitution from 'G' to 'T'.

D12S297: Similarly to D9S252, the only difference between the two sequenced alleles at the D12S297 locus was the number of tandemly repeated AGAT motifs.

8.e. Allelic Variation

All 10 tetranucleotide short tandem repeat loci were polymorphic across all the populations of the present study. Interesting distributions of allele frequencies were observed among the different loci. The number of different alleles varied from six alleles at the D1S407 locus to sixteen different alleles at D12S297 locus (across all populations). In general, greater numbers of alleles were isolated within the U.K. Leicestershire population than the Polynesian populations. Although at specific loci, lower numbers of alleles were observed in the U.K. Leicestershire population than the Polynesian populations (NAI and NAM), in particular at D4S2285 and D9S252 even though larger numbers of U.K. chromosomes were analysed.

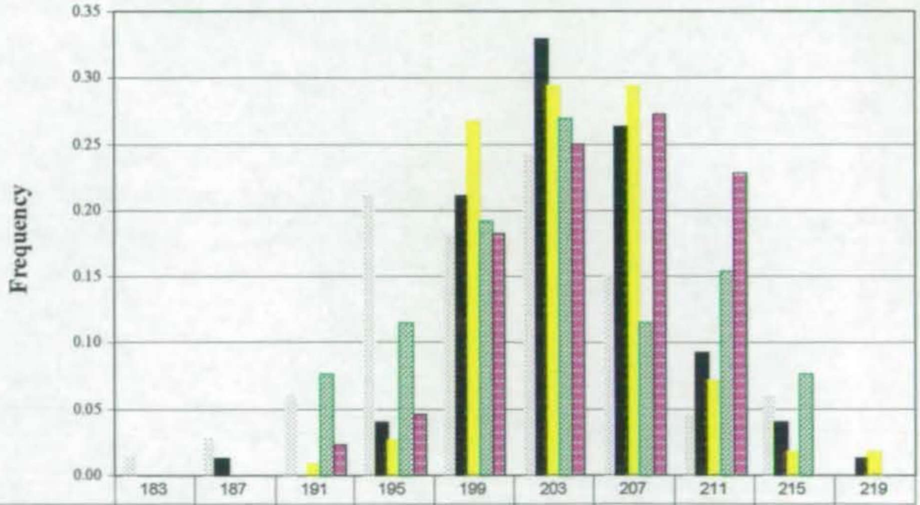
Allele frequency Distributions

The allele frequency distributions at each locus across all populations were calculated (figures 8.1 – 8.10). The allele size in base pairs can be used to describe the allele when a repeat motif count can not be provided (Goldstein and Schlotterer 1999). The allele size in base pairs was used in the present study. The allele frequency tables were listed below the graphs, where no alleles were observed the table is left blank. The results for individual polymorphic loci are discussed overpage:

D2S262

In general, the allele frequencies followed a normal distribution with the 203 base pair (bp) allele the most frequent allele in the AM, AI, U.K. populations and equally as frequent with 207bp allele of the NAM population. Over 75% of the NAM data was described by 3 alleles (199bp, 203bp and 207bp). Six alleles were isolated with frequencies above 5%, within the U.K. Leicestershire population. In the NAI and AI populations, 4 and 7 alleles respectively had frequencies above 5% (Figure 8.1). The smallest (183bp) allele was only observed in the U.K. Leicestershire population.

FIGURE 8.1: ALLELE FREQUENCY DISTRIBUTION AT THE D2S262 LOCUS

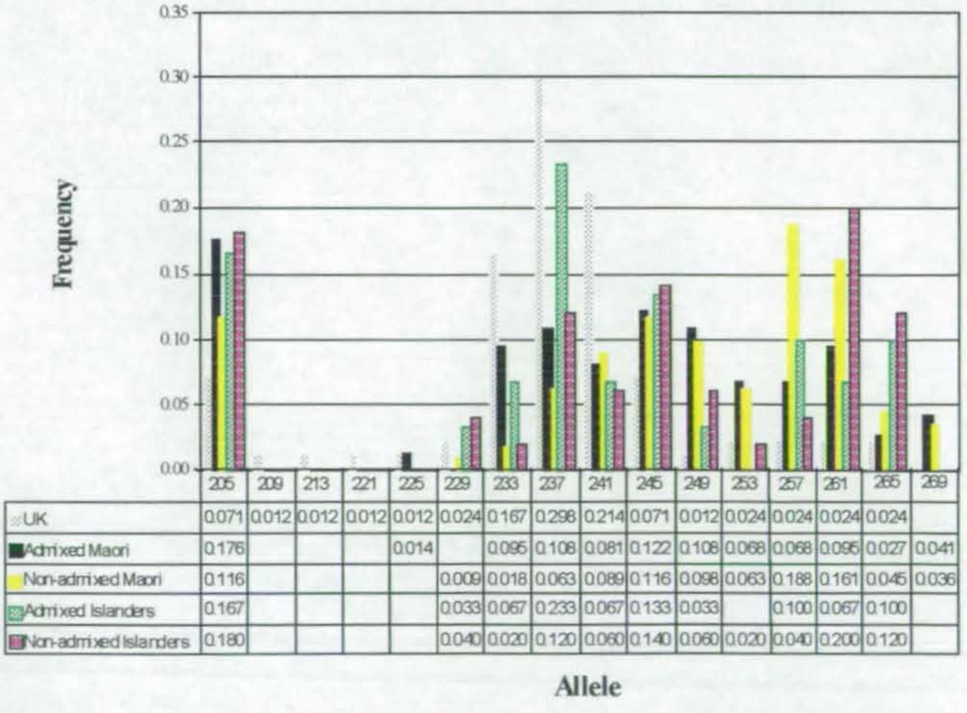


	183	187	191	195	199	203	207	211	215	219
▨ U.K.	0.015	0.030	0.061	0.212	0.182	0.242	0.152	0.045	0.061	
■ Admixed Maori		0.013		0.039	0.211	0.329	0.263	0.092	0.039	0.013
■ Non-admixed Maori			0.009	0.027	0.268	0.295	0.295	0.071	0.018	0.018
▨ Admixed Islander			0.077	0.115	0.192	0.269	0.115	0.154	0.077	
▨ Non-admixed Islander			0.023	0.045	0.182	0.250	0.273	0.227		

D12S297

Sixteen different alleles were observed at this locus across all the populations (see figure 8.2) and a bimodal distribution of allele frequencies was observed. The shortest allele (205bp) was found in all the populations with frequencies between 7-18%. No alleles were observed between 209bp and 225bp in the Polynesian Islander and non-admixed Maori populations and only at low frequency (<5%) between alleles of 221bp and 229bp within the U.K. Leicestershire population. There was a reversal of gene frequencies between the U.K. Leicestershire population, admixed Polynesians and non-admixed Polynesians, where 233bp and 237bp alleles were most frequent in the admixed Polynesian and U.K. Leicestershire populations and least frequent in non-admixed Polynesian populations. Similarly, the 261bp allele was most frequent in the non-admixed Polynesian populations and least frequent in admixed and U.K. Leicestershire populations.

FIGURE 8.2: ALLELE FREQUENCY DISTRIBUTION AT THE D12S297 LOCUS

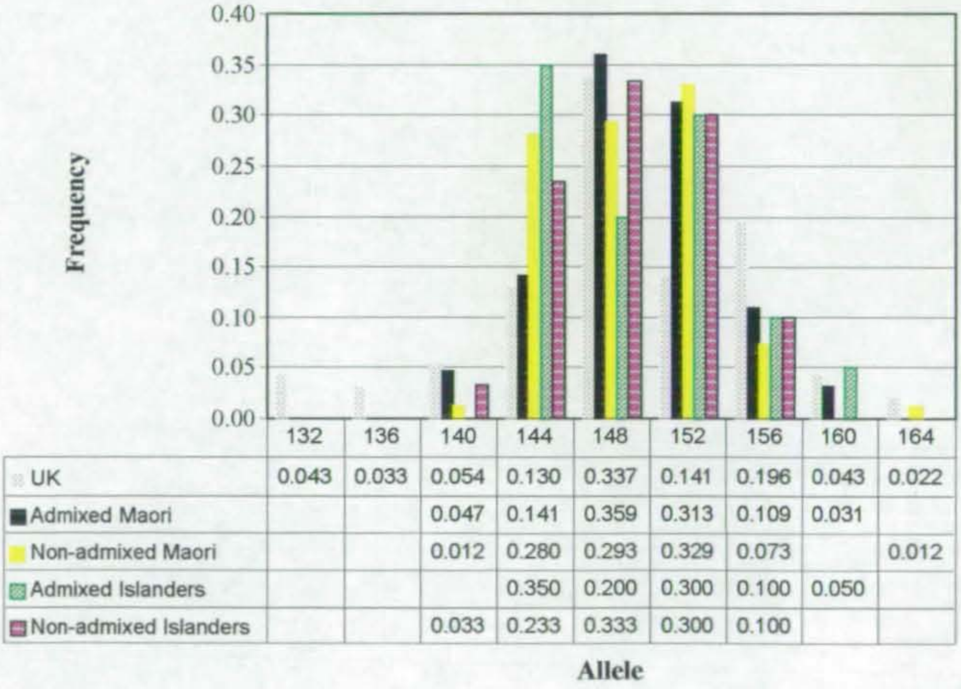


	205	209	213	221	225	229	233	237	241	245	249	253	257	261	265	269
▨ UK	0.071	0.012	0.012	0.012	0.012	0.024	0.167	0.298	0.214	0.071	0.012	0.024	0.024	0.024	0.024	
■ Admixed Maori	0.176				0.014		0.095	0.108	0.081	0.122	0.108	0.068	0.068	0.095	0.027	0.041
■ Non-admixed Maori	0.116					0.009	0.018	0.063	0.089	0.116	0.098	0.063	0.188	0.161	0.045	0.036
▨ Admixed Islanders	0.167					0.033	0.067	0.233	0.067	0.133	0.033		0.100	0.067	0.100	
▨ Non-admixed Islanders	0.180					0.040	0.020	0.120	0.060	0.140	0.060	0.020	0.040	0.200	0.120	

D1S407

The allele frequency distribution was normally distributed among all the populations (see figure 8.3). The Polynesian populations had a maximum of 6 alleles in comparison to the 9 alleles of the U.K. Leicestershire population. Over 75% of the total data (all populations) was described by three alleles (144bp, 148bp and 152bp) within the Polynesian populations, with the remaining alleles at frequencies below 11%. The 132bp allele and 136bp alleles were only observed at low frequencies in the U.K. Leicestershire population (<5%).

FIGURE 8.3: ALLELE FREQUENCY DISTRIBUTION AT THE D1S407 LOCUS

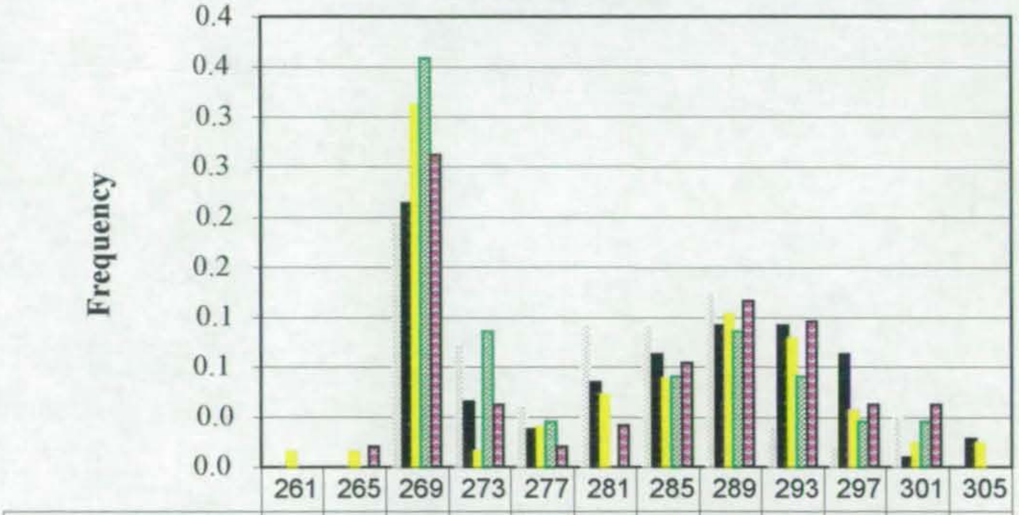


	132	136	140	144	148	152	156	160	164
░ UK	0.043	0.033	0.054	0.130	0.337	0.141	0.196	0.043	0.022
■ Admixed Maori			0.047	0.141	0.359	0.313	0.109	0.031	
■ Non-admixed Maori			0.012	0.280	0.293	0.329	0.073		0.012
▩ Admixed Islanders				0.350	0.200	0.300	0.100	0.050	
▦ Non-admixed Islanders			0.033	0.233	0.333	0.300	0.100		

D4S2285

A bimodal distribution of allele frequencies was observed at this locus (see figure 8.4). The 269bp allele was the most common in all the populations. The U.K. Leicestershire population had the lowest number of different alleles (9 different alleles) compared to the Polynesian populations (10-11 different alleles). The 265bp allele and 305bp allele were only found in Polynesian DNA samples.

FIGURE 8.4: ALLELE FREQUENCY DISTRIBUTION AT THE D4S2285 LOCUS



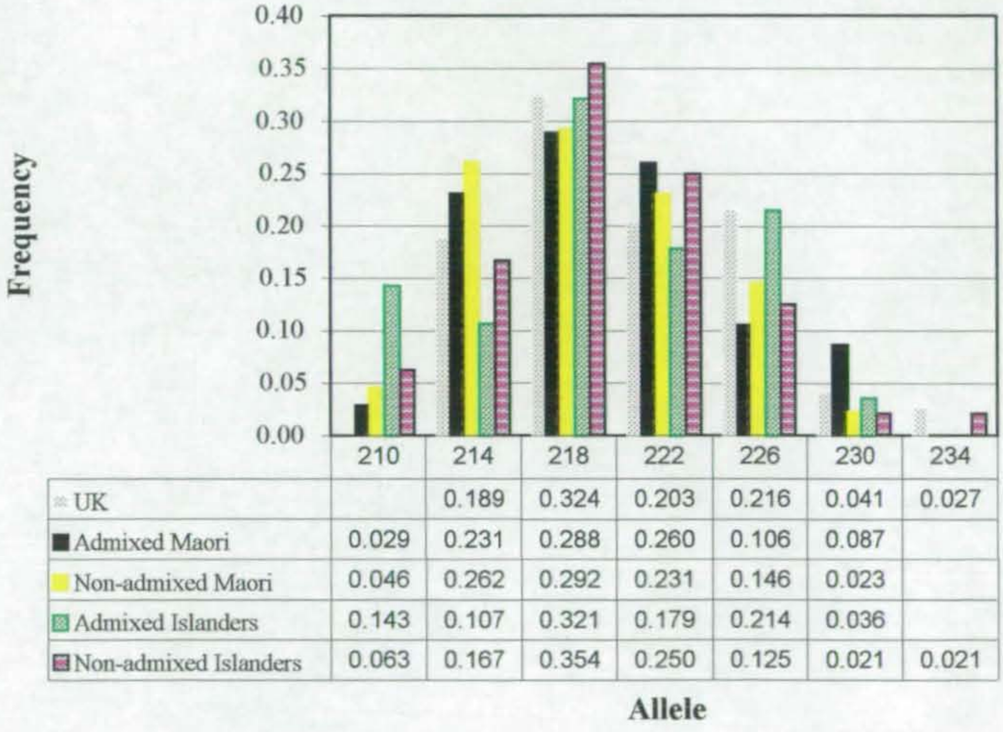
	261	265	269	273	277	281	285	289	293	297	301	305
░ UK			0.24	0.12	0.06	0.14	0.14	0.17	0.04	0.02	0.05	
■ Admixed Maori			0.26	0.06	0.03	0.08	0.11	0.14	0.14	0.11	0.00	0.02
🟨 Non-admixed Maori	0.01	0.01	0.36	0.01	0.04	0.07	0.08	0.15	0.12	0.05	0.02	0.02
▨ Admixed Islanders			0.40	0.13	0.04		0.09	0.13	0.09	0.04	0.04	
▦ Non-admixed Islanders		0.02	0.31	0.06	0.02	0.04	0.10	0.16	0.14	0.06	0.06	

Allele

D9S252

The allele frequency distribution at this locus is slightly negatively skewed (see figure 8.5). The 218bp allele was observed in over one quarter of all population samples tested. The 210bp allele was only observed in the Polynesian populations and the 234bp allele in U.K. Leicestershire and non-admixed Islander populations. Overall, 80% of the total data was described by 4 alleles (214bp, 218bp, 222bp and 226bp) a sharp contrast to that of D12S297 where 80% of the total data was described by 6 or more alleles.

Figure 8.5: Allele frequency distribution at the D9S252 locus

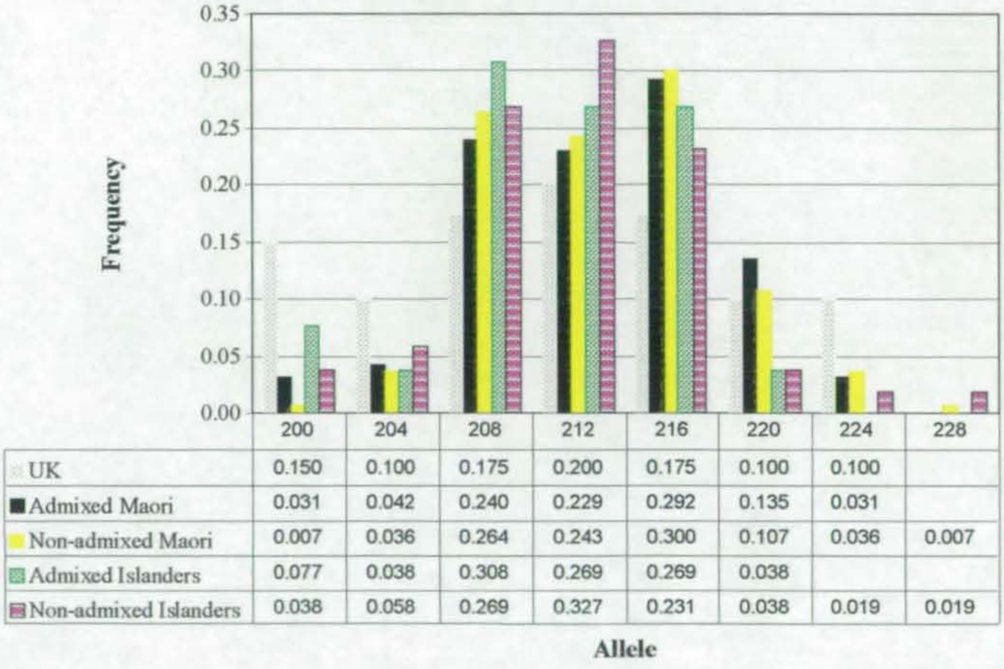


	210	214	218	222	226	230	234
▨ UK		0.189	0.324	0.203	0.216	0.041	0.027
■ Admixed Maori	0.029	0.231	0.288	0.260	0.106	0.087	
■ Non-admixed Maori	0.046	0.262	0.292	0.231	0.146	0.023	
▨ Admixed Islanders	0.143	0.107	0.321	0.179	0.214	0.036	
▤ Non-admixed Islanders	0.063	0.167	0.354	0.250	0.125	0.021	0.021

D7S1485

A normal distribution of 8 alleles was observed at this locus (see figure 8.6). Over 75% of the Polynesian data was described by three alleles (208bp, 212bp and 216bp). Of these three alleles, the 216bp allele was most frequent in the admixed and non-admixed Maori populations and the 208bp and 212bp alleles most frequent in the admixed and non-admixed Islander populations respectively. In contrast, the U.K. Leicestershire population had an even distribution of allele frequencies, with the most frequent allele of length 212bp describing 20% of the data.

FIGURE 8.6: ALLELE FREQUENCY DISTRIBUTION AT THE D7S1485 LOCUS

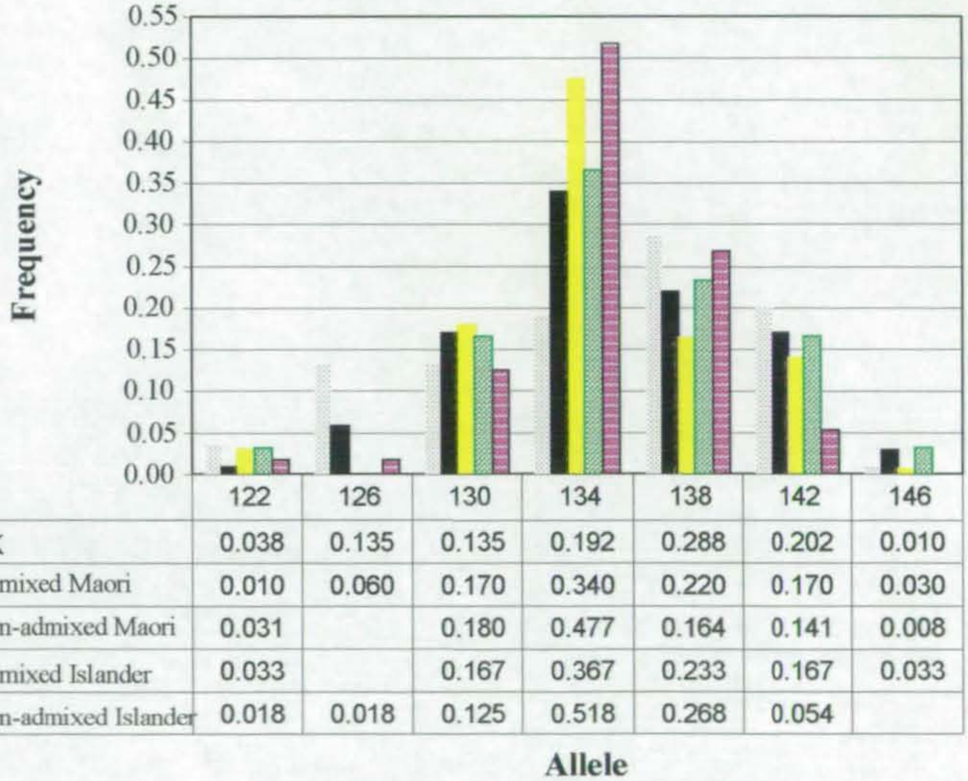


	200	204	208	212	216	220	224	228
░ UK	0.150	0.100	0.175	0.200	0.175	0.100	0.100	
■ Admixed Maori	0.031	0.042	0.240	0.229	0.292	0.135	0.031	
🟨 Non-admixed Maori	0.007	0.036	0.264	0.243	0.300	0.107	0.036	0.007
▨ Admixed Islanders	0.077	0.038	0.308	0.269	0.269	0.038		
▦ Non-admixed Islanders	0.038	0.058	0.269	0.327	0.231	0.038	0.019	0.019

D7S618

The distribution of alleles at this locus indicated a slight positive skew (see figure 8.7). Over 45% of the samples typed in the non-admixed Islander and non-admixed Maori populations had the 134bp allele. This was also the most frequent allele in the admixed Islander (36%) and admixed Maori (34%) populations. However, the 138bp allele was observed as the most frequent in the U.K. Leicestershire population (29%). Interestingly, the 126bp allele was not isolated in either the non-admixed Maori or the admixed Islander populations, even though the 122bp allele (1 repeat motif smaller) was isolated, albeit at a low frequency (<3%).

FIGURE 8.7: ALLELE FREQUENCY DISTRIBUTION AT THE D7S618 LOCUS

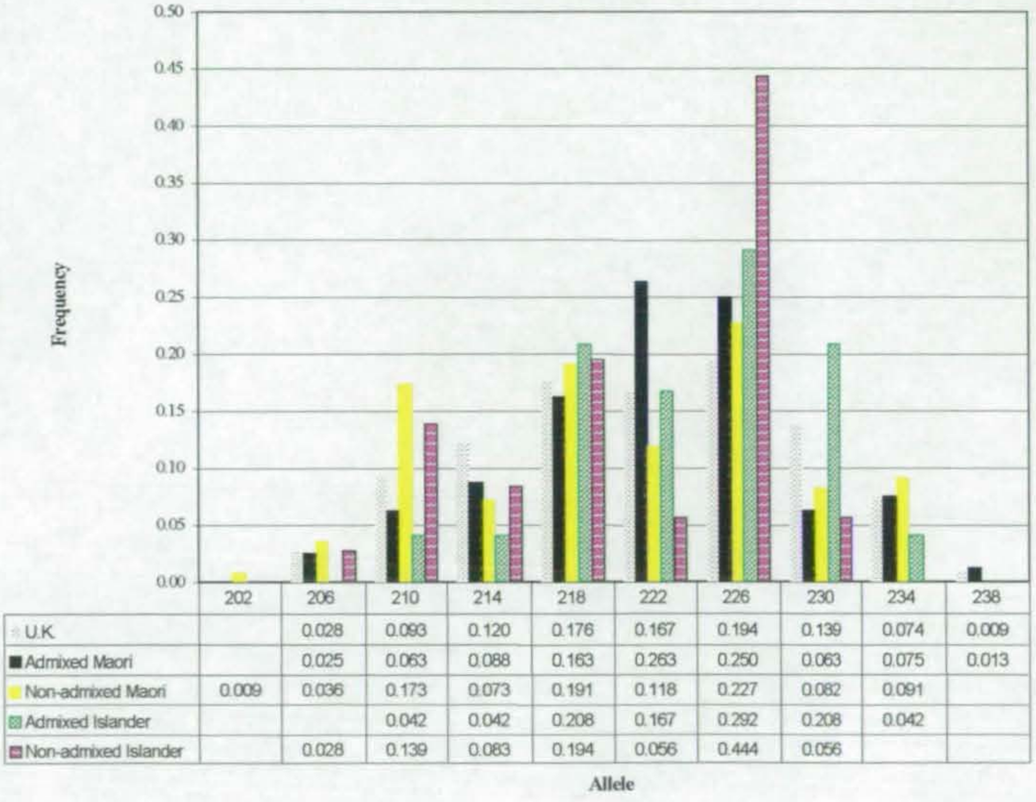


	122	126	130	134	138	142	146
░ UK	0.038	0.135	0.135	0.192	0.288	0.202	0.010
■ Admixed Maori	0.010	0.060	0.170	0.340	0.220	0.170	0.030
■ Non-admixed Maori	0.031		0.180	0.477	0.164	0.141	0.008
▨ Admixed Islander	0.033		0.167	0.367	0.233	0.167	0.033
▤ Non-admixed Islander	0.018	0.018	0.125	0.518	0.268	0.054	

D3S1514

The distribution of allele frequencies at this locus was positively skewed (see figure 8.8). Ten different alleles were observed across all the populations. The modal allele (226bp) had a frequency of 44% in the non-admixed Islander population. This allele was also most frequent in admixed Islander, non-admixed Maori and U.K. Leicestershire populations with frequencies of 29%, 23% and 19% of the respective population data. However, the modal allele (222bp) within the admixed Maori population, was observed at a frequency of 26.3%.

FIGURE 8.8: ALLELE FREQUENCY DISTRIBUTION AT THE D3S1514 LOCUS



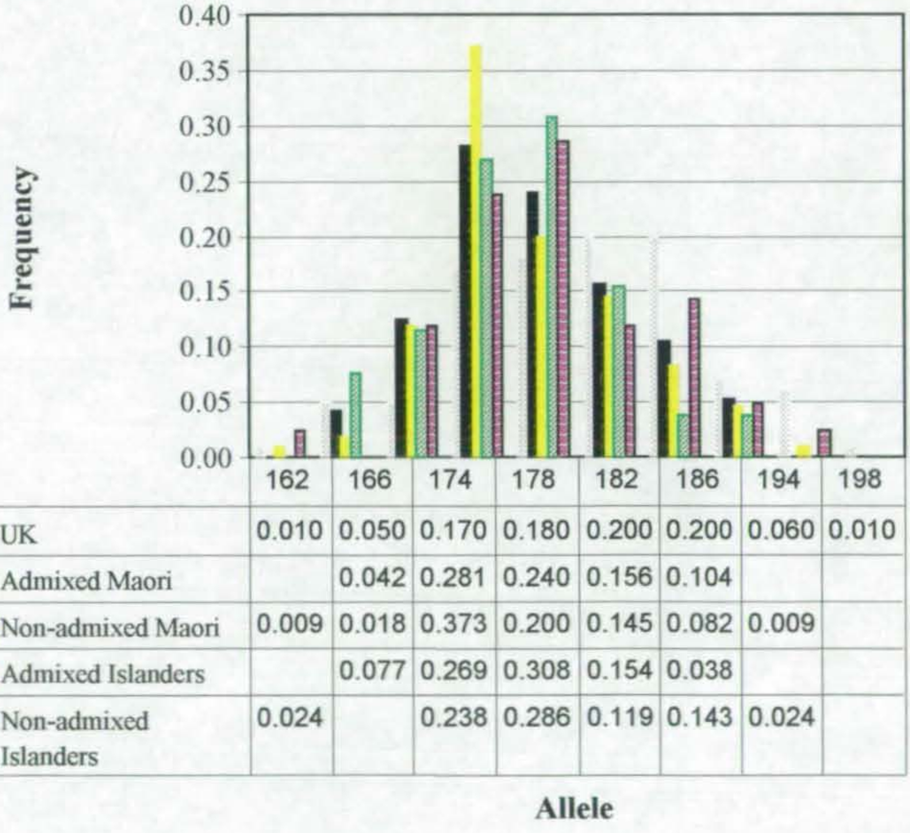
	202	206	210	214	218	222	226	230	234	238
▨ U.K		0.028	0.093	0.120	0.176	0.167	0.194	0.139	0.074	0.009
■ Admixed Maori		0.025	0.063	0.088	0.163	0.263	0.250	0.063	0.075	0.013
■ Non-admixed Maori	0.009	0.036	0.173	0.073	0.191	0.118	0.227	0.082	0.091	
▨ Admixed Islander			0.042	0.042	0.208	0.167	0.292	0.208	0.042	
▨ Non-admixed Islander		0.028	0.139	0.083	0.194	0.056	0.444	0.056		

D5S592

Ten different alleles were isolated among the five populations, with the U.K. Leicestershire population having the greatest number of different alleles (see figure 8.9). The allele frequencies of all the populations were normally distributed.

The 174bp allele was most frequent in the admixed and non-admixed Maori populations with respective frequencies of 28% and 37% of the population data. Similarly, the 178bp allele was modal in the non-admixed Islander and admixed Islander populations with frequencies of 29%, and 31% of the respective population data.

FIGURE 8.9: ALLELE FREQUENCY DISTRIBUTION AT THE D5S592 LOCUS



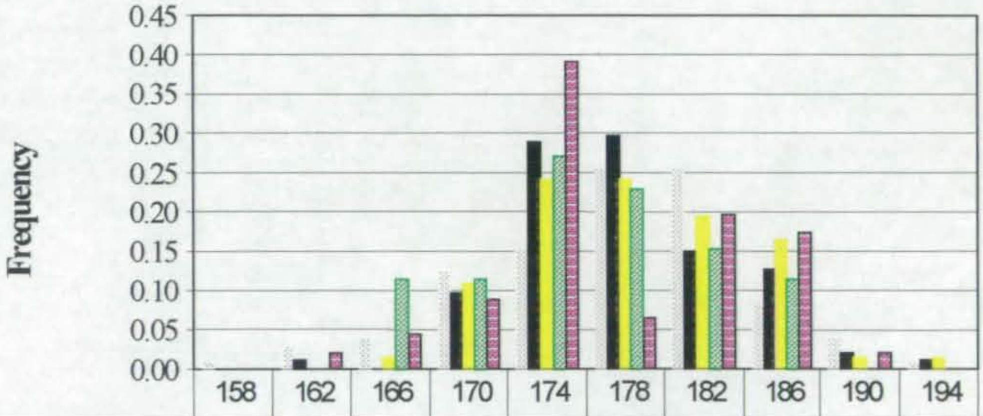
	162	166	174	178	182	186	194	198
░ UK	0.010	0.050	0.170	0.180	0.200	0.200	0.060	0.010
■ Admixed Maori		0.042	0.281	0.240	0.156	0.104		
■ Non-admixed Maori	0.009	0.018	0.373	0.200	0.145	0.082	0.009	
▨ Admixed Islanders		0.077	0.269	0.308	0.154	0.038		
▤ Non-admixed Islanders	0.024		0.238	0.286	0.119	0.143	0.024	

D10S520

The allele frequencies at this locus were positively skewed. Hence, the frequency distribution was positively skewed (alleles 182 and 186bp) about the modal alleles (174 and 178bp).

Similarly, to D5S592 this locus also had a total of ten different alleles across the five populations ranging from 158bp to 194bp in length. The 174bp to 186bp alleles described over 70% of the total population data. Of the highest allele frequencies, over 35% of the AI population had the 174bp allele. The 178bp allele was observed at 30% of the AM population and the 178bp and 182bp alleles together described 50% of the U.K. Leicestershire population data.

FIGURE 8.10: ALLELE FREQUENCY DISTRIBUTION AT THE D10S520 LOCUS



	158	162	166	170	174	178	182	186	190	194
░ UK	0.010	0.031	0.041	0.122	0.153	0.255	0.255	0.082	0.041	0.010
■ Admixed Maori		0.011		0.096	0.287	0.298	0.149	0.128	0.021	0.011
🟨 Non-admixed Maori			0.016	0.109	0.242	0.242	0.195	0.164	0.016	0.016
▨ Admixed Islanders			0.115	0.115	0.269	0.231	0.154	0.115		
▦ Non-admixed Islanders		0.022	0.043	0.087	0.391	0.065	0.196	0.174	0.022	

Allele

8.f. Testing for Hardy Weinberg equilibrium

Two independent tests were computed using the GDA software.

The probability, or '*p*-' value indicated the extent of which the data failed to reject the hypothesis that the population was in Hardy-Weinberg equilibrium (HWE). Values below the standard 0.05 level indicated a significant departure from HWE (Weir 1996).

The exact test compared observed and expected genotypic frequencies. This method assessed the probability of observed genotypic frequencies, assuming HWE, conditional on the observed allele frequencies.

The fisher measure evaluated HWE by shuffling the results of the data set. Then the genotypes generated were assessed to be more or less probable than the original data. For the purposes of assessing the significance of this data, the probabilities were calculated from 5000 re-shufflings.

The Chi-square measure followed the standard chi-square formula, a 'success' occurred when a shuffling produced a more extreme chi-square value than the observed value.

These tests produced slightly different results. Of the 100 HWE tests (50 exact and 50 chi-square) only 10 departures were observed among the populations, which was slightly higher than one would expect by chance alone (Table 8.3). However, the departures from HWE were not consistent across loci or consistent within any one population. This could be due to a chance factor or may be indicative of a population substructure. The non-admixed Islander population, at locus D12S297, was not in HWE using either test statistic. While the non-admixed Maori and U.K Leicestershire populations departed from HWE at the D1S407 and D9S252 loci respectively. Further departures from HWE using the exact test were observed at D9S252 (non-admixed Maori) and D2S262 (admixed Islander), while departures from HWE using the chi-square test were observed at D1S407 (U.K. Leicestershire) and D7S618 (non-admixed Maori). There were no observed deviations from HWE within the admixed Maori population using either the exact test or the chi-square test.

Table 8.3: Two measures to test for Hardy-Weinberg equilibrium across all populations and all loci.

	Fisher's Exact Test					**Chi-Square Measure**				
	U.K.	**AM**	**NAM**	**NAI**	**AI**	**U.K.**	**AM**	**NAM**	**NAI**	**AI**
Locus	*p-*	*p-*	*p-*	*p-*	*p-*	*p-*	*p-*	*p-*	*p-*	*p-*
D10S520	0.557	0.606	0.149	0.289	0.177	0.380	0.548	0.201	0.492	0.291
D12S297	0.853	0.341	0.494	**0.023**	0.861	0.813	0.510	0.625	**0.040**	0.780
D1S407	0.074	0.154	**0.002**	0.541	0.750	**0.002**	0.289	**0.027**	0.599	0.704
D4S2285	**0.010**	0.205	0.276	0.159	0.804	**0.001**	0.364	0.334	0.152	0.690
D9S252	0.196	0.216	**0.013**	0.156	0.065	0.212	0.295	0.059	0.239	0.064
D3S1514	0.588	0.620	0.462	0.642	0.895	0.639	0.735	0.157	0.708	0.127
D2S262	0.074	0.223	0.304	0.755	**0.016**	0.075	0.515	0.423	0.726	0.061
D7S618	0.379	0.097	0.334	0.090	0.157	0.571	0.208	**0.029**	0.282	0.619
D7S1485	0.618	0.491	0.649	0.261	0.952	0.383	0.437	0.138	0.149	0.746
D5S592	0.160	0.588	0.439	0.842	0.872	0.156	0.691	0.701	0.428	0.668

Values in **bold type** indicate a significant departure from HWE (Significance level <0.05 used as proposed by Weir 1996).

8.g. Heterozygosity and Nei's unbiased gene diversity

The percentage of observed heterozygotes, the heterozygosity index (HI) and Nei's unbiased gene diversity statistic were calculated for all loci and populations (refer to table 8.4).

The non-admixed Islander population had the lowest observed heterozygosity and HI at all loci except D1S407, D2S262 and D4S2285. At these loci, the sample numbers tested were lower than the U.K. Leicestershire and Maori populations, hence a sampling bias may have affected estimates.

The highest HI values were observed in the U.K. Leicestershire population at seven out of ten loci D1S407, D2S262, D3S1514, D5S592, D7S1485, D7S618, and D10S520.

Nei's unbiased gene diversity, was also calculated (Table 8.4). This incorporated sample sizes which reduced any bias due to either very small or very large sample numbers. Nei's unbiased gene diversity in comparison to the heterozygosity index produced slightly different results. In general, the unbiased gene diversity values were higher than the heterozygosity indicies. At the D10S520 locus, the U.K. Leicestershire population had the highest heterozygosity index value (0.820), while the unbiased gene diversity statistic was highest in the admixed Islander population (0.843). Similarly, at the D2S262 locus, the U.K. Leicestershire population had the highest heterozygosity index value (0.834), while the unbiased gene diversity statistic was higher in the admixed Islander population (0.861).

TABLE 8.4: CALCULATIONS OF HETEROZYGOSITIES AND UNBIASED GENE DIVERSITIES AT ALL LOCI ACROSS ALL POPULATIONS

		U.K.	**AM**	**NAM**	**AI**	**NAI**
D10S252	% Heterozygotes	81.6%	83.0%	79.7%	76.9%	73.9%
	Allele Numbers	98	94	128	26	46
	HI	**0.820**	0.780	0.805	0.811	0.764
	Unbiased *h*	0.828	0.788	0.811	**0.843**	0.780
D12S297	% Heterozygotes	88.1%	81.10%	94.60%	93.30%	80.00%
	Allele Numbers	84	74	112	30	50
	HI	0.823	0.923	**0.951**	0.889	0.924
	Unbiased *h*	0.832	0.936	**0.959**	0.919	0.943
D1S407	% Heterozygotes	65.2%	59.40%	65.90%	80.00%	80.00%
	Allele Numbers	92	64	82	20	30
	HI	**0.803**	0.738	0.722	0.735	0.733
	Unbiased *h*	**0.812**	0.749	0.731	0.774	0.758
D2S262	% Heterozygotes	76.70%	65.80%	75.00%	69.20%	81.80%
	Allele Numbers	60	76	112	26	44
	HI	**0.834**	0.766	0.748	0.828	0.776
	Unbiased *h*	0.848	0.776	0.755	**0.861**	0.794
D3S1514	% Heterozygotes	88.9%	82.50%	85.50%	91.70%	72.20%
	Allele Numbers	108	80	110	24	36
	HI	**0.855**	0.820	0.846	0.795	0.731
	Unbiased *h*	**0.863**	0.830	0.854	0.829	0.752
D4S2285	% Heterozygotes	81.6%	92.50%	90.30%	81.80%	83.30%
	Allele Numbers	98	106	124	22	48
	HI	0.846	**0.851**	0.808	0.773	0.828
	Unbiased *h*	0.855	**0.859**	0.815	0.810	0.845
D5S592	% Heterozygotes	96.0%	95.80%	89.10%	92.30%	81.00%
	Allele Numbers	100	96	110	26	42
	HI	**0.845**	0.808	0.777	0.787	0.810
	Unbiased *h*	**0.854**	0.817	0.784	0.818	0.830
D7S1485	% Heterozygotes	85.0%	79.20%	84.30%	92.30%	80.80%
	Allele Numbers	40	96	140	26	52
	HI	**0.846**	0.783	0.767	0.751	0.760
	Unbiased *h*	**0.868**	0.791	0.773	0.781	0.775
D7S618	% Heterozygotes	74.5%	84.0%	68.8%	60.0%	50.0%
	Allele Numbers	102	100	128	30	56
	HI	**0.805**	0.774	0.693	0.753	0.641
	Unbiased *h*	**0.813**	0.782	0.698	0.779	0.653
D9S252	% Heterozygotes	83.8%	73.1%	80.0%	85.7%	70.8%
	Allele Numbers	74	104	130	28	48
	HI	0.769	0.777	0.769	**0.786**	0.764
	Unbiased *h*	0.801	0.785	0.775	**0.815**	0.780

Bold type indicates highest HI at that locus

8.g.i. Unbiased gene diversity, the sampling variance and standard error

The average population gene diversity ranged from 79.1% in the non-admixed Islander to 83.5% in the U.K. Leicestershire population (Table 8.5). The sampling variances were smallest in the U.K. Leicestershire population (0.0001) and highest in the non-admixed Islander population (0.0006). The standard errors were also smallest in the U.K. Leicestershire population and largest in the non-admixed Islander population.

Following Nei's (1987) test of significance, a comparison of the U.K. Leicestershire gene diversity to the Polynesian populations was carried out. Using the standard value of 2.26 as the level of significance, all the gene diversity comparisons were below this level, thus not significantly different (Table 8.6). However, the range of significance values did increase from 0.69 (comparing U.K. Leicestershire diversity to the admixed Islander diversity) to 1.66 (between the Leicestershire and non-admixed Islander diversities) as one may expect given the individual population gene diversities.

TABLE 8.5: UNBIASED GENE DIVERSITY ESTIMATES ACROSS ALL LOCI CALCULATED FOR EACH POPULATION

Population	Av. Gene diversity	Sampling Variance	Standard Error
U.K.	0.835	0.0001	0.0100
AI	0.823	0.0002	0.0141
AM	0.811	0.0003	0.0173
NAM	0.795	0.0005	0.0224
NAI	0.791	0.0006	0.0245

Using Nei's 8.7 equation Pg 179 and unbiased estimate of Gene diversity (HI).

TABLE 8.6: NEI'S TEST OF SIGNIFICANCE FOR GENE DIVERSITY BETWEEN TWO POPULATIONS (SEE NEI 1987 FOR EQUATION)

Population versus U.K.	Test of significance (t)
AI	0.69
AM	1.20
NAM	1.63
NAI	1.66

8.h. Variance Measures

These measures followed Neis' equations of 'variance of gene diversity' or *interlocus* variance equation 8.8, and *intralocus* variance ($V_s(h)$) associated with the sampling of individuals at each locus (see Nei 1987 equation 8.14).

The intralocus variance was higher than the interlocus variance except in the non-admixed Polynesian population (Table 8.7). The interlocus variance associated with the sampling of loci from the genome (Nei 1987), was highest in the non-admixed Islander (6.3%) and lowest in the U.K. Leicestershire population (0.06%). The intralocus variance associated with the sampling of individuals at each locus (Nei 1987), was observed to be highest in the admixed Islander population (2.7%) and lowest in the non-admixed Maori population (0.6%), with decreasing variances in the order non-admixed Islander (1.7%), admixed Maori (0.8%) and U.K. Leicestershire (0.7%).

TABLE 8.7: INTERLOCUS AND INTRALOCUS VARIANCE OF EACH POPULATION ACROSS ALL LOCI

	Interlocus Variance	Intralocus variance
Caucasian	0.0006	0.0074
AI	0.0019	0.0265
AM	0.0029	0.0076
NAM	0.0052	0.0061
NAI	0.0633	0.0168

8.i. F_{ST} Measures

FST (Theta-P) or the coancestry coefficient allows identification of variation between individuals within populations and between populations. If populations are affected by drift, then once equilibrium has taken place the alleles become fixed and FST=1 (Goldstein and Schlotterer 1999).

Variation within populations

All FST values were below 0.03 (Table 8.8), indicating little genetic differentiation or divergence (Hartl 1981). Negative values indicated that the true value was very small and that a negative bias had lowered the value, or, the parameter was negative.

The FST values ranged from –0.008 (NAI – D5S592) to 0.026 (AM – D7S618). The D12S297 locus had the highest mean FST value while three loci (D9S252, D7S1485 and D5S592) had mean negative values. Negative values for all the populations at D9S252 indicated that there was more variation within populations than between them.

TABLE 8.8: FST VALUES PER POPULATION PER LOCUS

FST	D10s520	D12s297	D1s407	D4s2285	D9s252	D3s1514	D2s262	D7s618	D7s1485	D5s592
U.K.	0.005	0.005	0.000	-0.005	-0.004	0.013	0.001	0.003	-0.006	-0.005
AM	0.016	0.027	0.016	0.011	-0.003	0.010	0.022	0.026	0.000	0.001
NAM	0.022	0.010	0.018	0.010	-0.002	0.014	0.014	0.012	0.000	-0.008
NAI	0.002	0.020	0.022	0.011	-0.002	0.004	0.014	0.015	-0.005	-0.001
AI	0.016	0.022	0.016	0.009	-0.005	0.012	0.018	0.021	0.000	-0.001
Mean	0.009	0.021	0.018	0.008	-0.004	0.008	0.015	0.018	-0.005	-0.001
Std. dev.	0.015	0.016	0.015	0.012	0.002	0.007	0.014	0.016	0.005	0.006

Inbreeding values – FIS

Inbreeding is relatively common in small populations where the opportunity to mate with an unrelated partner is reduced. One would expect that the inbreeding value for a diverse population as the U.K. Leicestershire population, to be low compared to non-admixed Islanders. This was indeed suggested by the results of the present study. The average inbreeding value of the U.K. Leicestershire population was –0.0293 compared to 0.0343 of the non-admixed Islanders (see Table 8.9). Interestingly, the non-admixed Maori population had a similar inbreeding value (-0.0248) to the U.K. Leicestershire population, suggesting that this population was of a sufficient size to support unrelated matings. The admixed Islanders were observed to have a lower inbreeding coefficient (0.005) than the admixed Maoris, however, were more inbred than the non-admixed Maori population.

TABLE 8.9: FIS VALUES – WITHIN POPULATION INBREEDING VALUES PER POPULATION PER LOCUS

Locus	Allele	U.K.	NAM	AM	NAI	AI	Mean
D10s520	All	0.015	0.018	-0.053	0.035	0.116	0.026
D12s297	All	-0.057	-0.063	0.115	0.089	-0.030	0.011
D1s407	All	0.106	0.153	0.192	-0.057	-0.036	0.072
D4s2285	All	0.045	-0.113	-0.077	0.015	-0.011	-0.028
D9s252	All	-0.076	-0.033	0.066	0.094	-0.054	-0.001
D3s1514	All	-0.026	0.000	0.007	0.041	-0.110	-0.180
D2s262	All	0.174	0.006	0.154	-0.031	0.225	0.106
D7s618	All	0.094	0.036	-0.076	0.231	0.251	0.107
D7s1485	All	-0.050	-0.092	-0.001	-0.057	-0.175	-0.075
D5s592	All	-0.129	-0.138	-0.176	0.004	-0.147	-0.117
Across all loci	---	0.0088	-0.0249	0.0142	0.0347	0.0052	

8.j. Genetic Distances and Phenograms

Nei's 1978 distance measure and coancestry distance were calculated using the computer program 'GDA' (see table 8.10). The coancestry distance measure was also calculated using the Y-chromosome lineage data (see Y chromosome results chapter).

The distance measures were calculated using all the populations at all 10 autosomal marker loci. Both distance measures produced a similar 'pattern' of results, with the closest genetic relationship observed to be between the admixed Maori and admixed Islanders and the most distant between the U.K. Leicestershire and non-admixed Islander populations. Negative distance values were observed between the admixed Islander and the other Polynesian populations.

TABLE 8.10: COANCESTRY DISTANCE MATRIX BELOW DIAGONAL AND NEI'S 1978 DISTANCE MATRIX ABOVE DIAGONAL.

	U.K.	AM	NAM	NAI	AI
U.K.	0	0.051	0.095	0.123	0.027
AM	0.012	0	0.004	0.018	-0.026
NAM	0.023	0.001	0	0.015	-0.024
NAI	0.029	0.005	0.004	0	-0.043
AI	0.006	-0.006	-0.005	-0.010	0

Unweighted Pair Group Method (UPGMA) and Neighbor Joining phenograms were constructed from the distance matrices (see figures 8.11 to 8.14). Node values were generated that gave an indication of the genetic 'distance' between populations. The phenograms were identical for both distance measures, although the 'node' distances varied. The structure of the UPGMA phenogram was distinctly different in comparison to the Neighbor Joining phenogram (refer to figures 8.11 to 8.14).

8.j.i. UPGMA Phenogram construction of the distance measures

The UPGMA phenogram clustered the Polynesian populations separately to the U.K. Leicestershire population. Furthermore, the non-admixed Islander population separated from the admixed Islander and admixed Maori and non-admixed Maori populations. The non-admixed Maori population was closely separated from the two admixed populations that clustered together. Thus, the closest genetic similarity was observed between the admixed populations and the second closest population was the non-admixed Maori population. Interestingly, the non-admixed Maori population held a closer genetic relationship to the admixed populations than to the non-admixed Islander population.

8.j.ii. Neighbour- Joining Phenogram construction of the distance measures

The Neighbor Joining phenogram expressed an early separation of the non-admixed Islanders from the remaining populations (see figures 8.11 and 8.13). The closest population to the non-admixed Islanders was the non-admixed Maori that branched from the admixed populations and the U.K. Leicestershire population. Interestingly, the admixed Maori branched from the admixed Islander and U.K. Leicestershire populations suggesting a slightly closer genetic relationship to its non-admixed parent population.

In this phenogram, the U.K. Leicestershire population was closest to the admixed Islander population, although the genetic separation between these two populations was the largest of all.

FIGURE 8.11: NEIGHBOUR JOINING PHENOGRAM OF THE COANCESTRY DISTANCE MEASURE

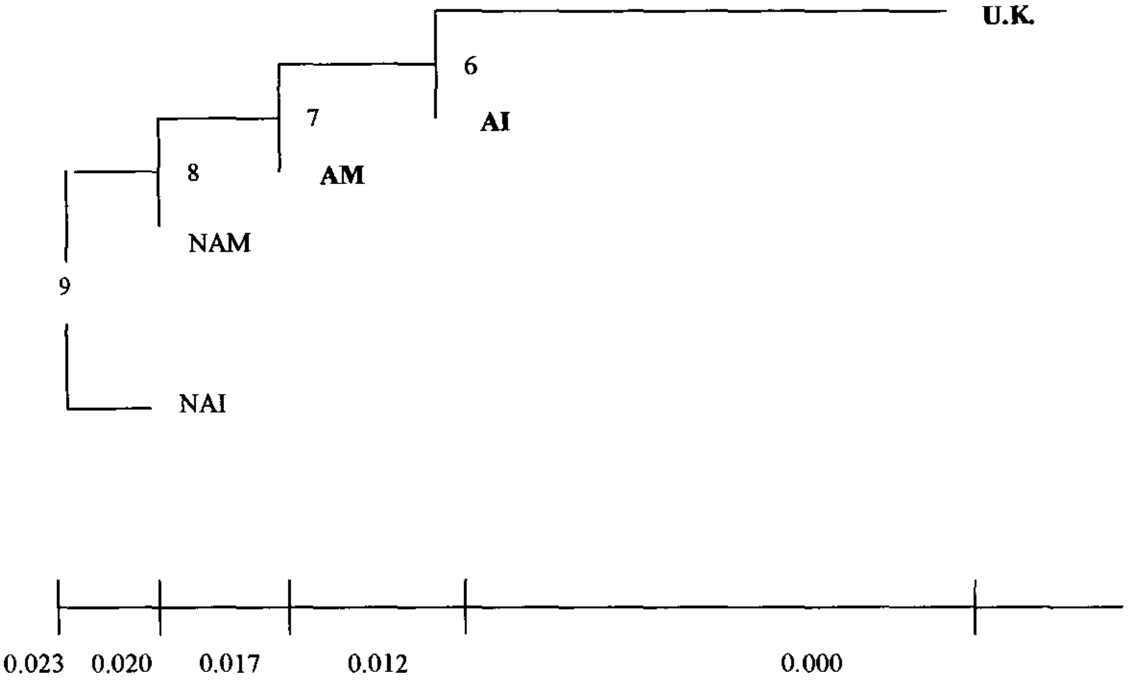


FIGURE 8.12: A UPGMA PHENOGRAM OF THE COANCESTRY DISTANCE MEASURE

Node 6 created (level = 0.0000)
Node 7 created (level = 0.0003)
Node 8 created (level = 0.0016)
Node 9 created (level = 0.0112)

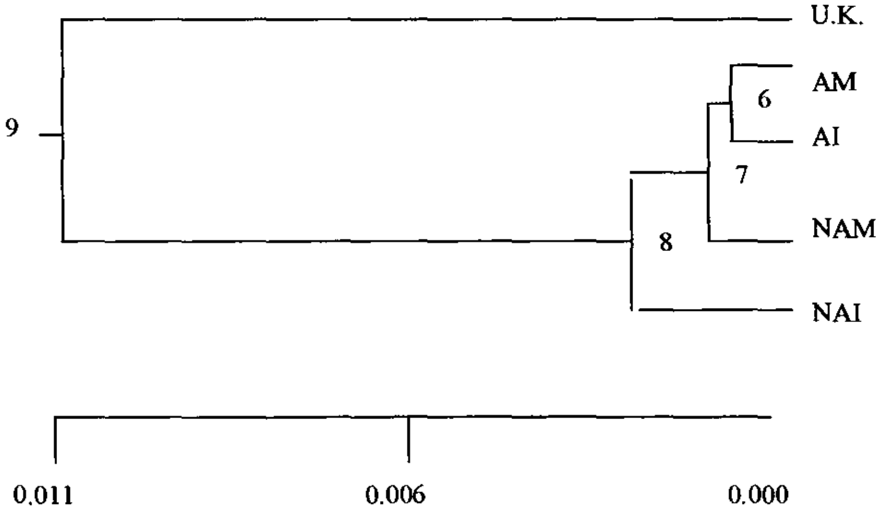


FIGURE 8.13: A UPGMA PHENOGRAM OF NEI'S 1978 DISTANCE MEASURE

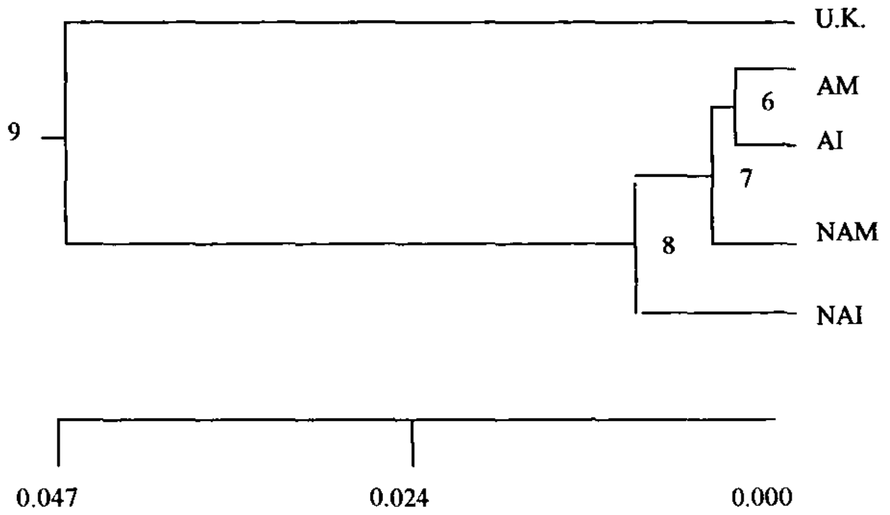
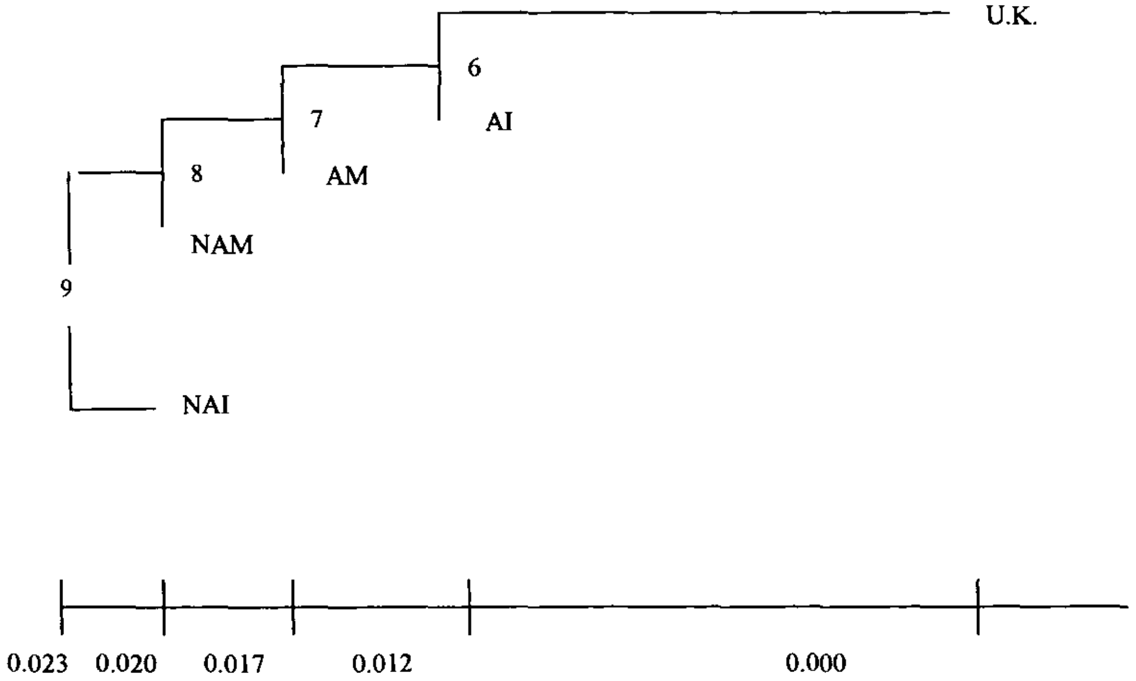


FIGURE 8.14: A NEIGHBOR-JOINING TREE OF NEI'S 1978 DISTANCE MEASURE

Node 6 created (Mij = -0.0806) Node 7 created (Mij = -0.0519)
Node 8 created (Mij = -0.0553) Node 9 created (Mij = 0.0143)



8.k. Forensic Statistical Considerations

The important question to be answered is; how informative are the ten tetranucleotide autosomal loci for use in Forensic investigations?

To answer this question a number of statistical analyses have been developed (Brenner and Morris 1990). There was a bias with respect to the analyses, as the sample sizes were not taken into consideration. Therefore, the informativeness of the loci were studied within populations, opposed to between populations at specific loci.

The *Powerstat* macro designed by *Promega* was used to calculate the matching probability, power of discrimination, polymorphic information content (PIC), power of exclusion and typical paternity index (details given in statistical methodology chapter). The heterozygosity index was also incorporated into the macro using the standard formula.

The best loci for use in forensic investigations are those that have low matching probabilities and corresponding high power of discrimination and high polymorphic information content (Lareu et al. 1998). These statistics use the allele frequency information. The 'power of exclusion' in paternity calculations, although incorporating allele frequency information, followed a different methodology dependent on the proportion of homozygotes to heterozygotes.

8.k.i.U.K. Leicestershire population

The probability of finding a matching DNA band ranged from 1 in 6 (D9S252) to 1 in 17 (D3S1514) and the polymorphic information content ranged from 0.69 (D9S252) to 0.83 (D3S1514) (Table 8.11).

The power of exclusion ranged from 39.2% (D1S407) to 100% (D5S592), and percentage heterozygotes 67.6% (D1S407) to 100% (D5S592). The order of the loci from highest percentage of heterozygotes to lowest was observed to be;

D5S592, D9S252, D3S1514, D4S2285, D12S297, D7S618, D10S520, D7S1485, D2S262 and D1S407.

8.k.ii. Admixed Maori Population

The probability of finding a matching DNA band ranged from 1 in 8 (D1S407) to 1 in 27 (D12S297) and polymorphic information content ranged from 70% (D1S407) to 89% (D12S297) (Table 8.12). The order of loci from most informative to least informative was observed to be;

D12S297, D4S2285, D3S1514, D5S592, D2S262, D7S1485, D9S252, D10S520, D7S618 and D1S407.

The power of exclusion ranged from 28.3% (D1S407) to 91.5% (D5S592) and percentage of heterozygotes 59.4% (D1S407) to 95.8% (D5S592). The order of the loci from highest percentage of heterozygotes to lowest was observed to be;

D1S407, D2S262, D9S252, D7S1485, D12S297, D3S1514, D10S520, D7S618, D4S2285 and D5S592.

8.k.iii. Non-Admixed Maori

The probability of a matching DNA profile band ranged from 1 in 6.5 (D1S407) to 1 in 23.1 (D12S297) and polymorphic information content ranged from 65% (D7S618) to 87% (D12S297). The order of loci from most informative to least informative was observed to be;

D12S297, D3S1514, D4S2285, D10S520, D5S592, D7S1485, D2S262, D9S252, D7S618 and D1S407 (Table 8.13).

The power of exclusion ranged from 36.7% (D1S407) to 89.1% (D12S297) and percentage of heterozygotes 65.9% (D1S407) to 94.6% (D12S297). The order of loci from highest percentage of heterozygotes to lowest was observed to be;

D12S297, D4S2285, D5S592, D3S1514, D7S1485, D9S252, D10S520, D2S262, D7S618 and D1S407.

8.k.iv. Admixed Islanders

The probability of a matching DNA profile band ranged from 1 in 6.3 (D1S407 and D7S1485) to 1 in 13.2 (D12S297) and polymorphic information content ranged from 69% (D1S407) to 85% (D12S297). The order of loci from most informative to least informative was observed to be;

D12S297, D2S262, D10S520, D5S592, D3S1514, D4S2284, D9S252, D7S618, D7S1485 and D1S407 (Table 8.14).

The power of exclusion ranged from 29.1% (D7S618) to 86.4% (D12S297) and percentage of heterozygotes 60% (D7S618) to 93.3% (D12S297). The order of loci from highest percentage of heterozygotes to lowest was observed to be;

D12S297, D5S592 and D7S1485 (both with 92.3% heterozygosity), D3S1514, D9S252, D4S2285, D1S407, D10S520, D2S262 and D7S618.

8.k.v. Non-Admixed Islanders

The probability of a matching DNA profile band ranged from 1 in 5.4 (D7S618) to 1 in 13.9 (D12S297) and polymorphic information content ranged from 59% (D7S618) to 85% (D12S297) (Table 8.15). The order of the loci from most informative to least informative was observed to be;

D12S297, D5S592, D4S2285, D10S520, D9S252, D2S262, D3S1514, D7S1485, D1S407 and D7S618.

The power of exclusion ranged from 18.8% (D7S618) to 66.2% (D4S2285) and percentage of heterozygotes 50% (D7s618) to 83.3% (D4S2285). The order of loci from highest percentage of heterozygotes to lowest was observed to be;

D4S2285, D2S262, D5S592, D7S1485, D12S297, D1S407, D10S520, D3S1514, D9S252, and D7S618.

8.l. In Summary

The polymorphic information content was highest in all the Polynesian populations at locus D12S297 and lowest in the AI, NAM and AM populations at locus D1S407.

On average, homozygosity was highest in the NAI (25%) in comparison to all the other populations. This increased level of homozygosity may be an indication of how inbred the population sample was and was further exemplified by the inbreeding statistic *fis*, which was observed to be highest in the non-admixed Islander population. The admixed and non-admixed Maori populations had similar average homozygosities of 20.5% and 19% respectively, indicating these were more outbred and diverse, which was also in agreement with *fis* estimates.

The average power of exclusion using these 10 loci was highest in the Leicestershire Caucasian population (67%) and lowest in the NAI population (53%) as one would expect given the previous results.

TABLE 8.11: COMMONLY USED FORENSIC STATISTICS IN THE U.K. LEICESTERSHIRE POPULATION

Forensic calculations	10S520	12S297	1S407	2S262	3S1514	4S2285	5S592	7S1485	7S618	9S252	Average	
Matching Probability	0.064	0.063	0.092	0.069	0.053	0.070	0.072	0.095	0.077	0.121	0.078	6.05151E-12
Expressed as 1 in ...	15.7	15.8	10.9	10.6	18.9	14.4	13.9	10.5	12.9	8.3		1.65E+11
Power of Discrimination	0.936	0.937	0.908	0.906	0.947	0.930	0.928	0.905	0.923	0.879	0.920	
PIC	0.80	0.80	0.780	0.73	0.84	0.83	0.83	0.83	0.78	0.73	0.795	
Paternity												
Power of Exclusion	0.630	0.757	0.358	0.424	0.773	0.630	0.919	0.695	0.501	0.671	0.636	0.999989
Typical Paternity Index	2.72	4.20	1.44	1.65	4.50	2.72	12.5	3.33	1.96	3.08	3.81	
Allele Frequencies												
Homozygotes	0.184	0.119	0.348	0.303	0.111	0.184	0.040	0.150	0.255	0.162	0.186	
Heterozygotes	0.816	0.881	0.652	0.658	0.889	0.816	0.960	0.850	0.745	0.838	0.814	
Total Alleles	98	84	92	66	108	98	100	40	102	74	86.2	

TABLE 8.12: COMMONLY USED FORENSIC STATISTICS IN THE ADMIXED MAORI POPULATION

Forensic calculations	10S520	12S297	1S407	2S262	3S1514	4S2285	5S592	7S1485	7S618	9S252	Average	Combined
Matching Probability	0.100	0.037	0.119	0.094	0.065	0.060	0.089	0.095	0.111	0.098	0.087	1.485E-11
Expressed as 1 in	10.000	26.800	8.400	10.600	15.400	16.600	11.200	10.600	9.000	10.200	12.880	6.73E+10
Power of Discrimination	0.900	0.963	0.881	0.906	0.935	0.940	0.911	0.905	0.889	0.902	0.913	
PIC	0.750	0.890	0.700	0.730	0.800	0.830	0.780	0.750	0.740	0.740	0.771	
Paternity												
Power of Exclusion	0.655	0.619	0.283	0.366	0.646	0.846	0.915	0.584	0.675	0.477	0.607	0.99998
Typical Paternity Index	2.940	2.640	1.230	1.460	2.860	6.630	12.000	2.400	3.130	1.860	3.715	
Allele Frequencies												
Homozygotes	0.170	0.189	0.406	0.342	0.175	0.075	0.042	0.208	0.160	0.269	0.204	
Heterozygotes	0.830	0.811	0.594	0.658	0.825	0.925	0.958	0.792	0.840	0.731	0.796	
Total Alleles	94	74	64	76	80	106	96	96	100	104	89	

TABLE 8.13: COMMONLY USED FORENSIC STATISTICS IN THE NON-ADMIXED MAORI POPULATION

Forensic Calculations	10S520	12S297	1S407	2S262	3S1514	4S2285	5S592	7S1485	7S618	9S252	Average	Combined
Matching Probability	0.081	0.043	0.154	0.120	0.056	0.079	0.103	0.106	0.147	0.128	0.102	5.8498E-11
Expressed as 1 in	12.300	23.100	6.500	8.300	17.900	12.700	9.700	9.500	6.800	7.800	11.460	1.71E+10
Power of Discrimination	0.919	0.957	0.846	0.880	0.944	0.921	0.897	0.894	0.853	0.872	0.898	
PIC	0.780	0.870	0.670	0.700	0.830	0.790	0.750	0.730	0.650	0.730	0.750	
Paternity												
Power of Exclusion	0.593	0.891	0.367	0.510	0.704	0.802	0.777	0.681	0.409	0.599	0.633	0.999986
Typical Paternity Index	2.460	9.330	1.460	2.000	3.440	5.170	4.580	3.180	1.600	2.500	3.572	
Allele Frequencies												
Homozygotes	0.203	0.054	0.341	0.250	0.145	0.097	0.109	0.157	0.313	0.200	0.187	
Heterozygotes	0.797	0.946	0.659	0.750	0.855	0.903	0.891	0.843	0.688	0.800	0.813	
Total Alleles	128	112	82	112	110	124	110	140	128	130	118	

TABLE 8.14: COMMONLY USED FORENSIC STATISTICS IN THE ADMIXED ISLANDER POPULATION

Forensic calculations	10S520	12S297	1S407	2S262	3S1514	4S2285	5S592	7S1485	7S618	9S252	Average	Combined
Matching Probability	0.124	0.076	0.160	0.124	0.125	0.107	0.124	0.160	0.120	0.173	0.129	1.03E-09
Expressed as 1 in	8.000	13.200	6.300	8.000	8.000	9.300	8.000	6.300	8.300	5.800	8.120	9.71E+08
Power of Discrimination	0.876	0.924	0.840	0.876	0.875	0.893	0.876	0.840	0.880	0.827	0.871	
PIC	0.780	0.850	0.690	0.810	0.760	0.750	0.760	0.710	0.720	0.750	0.758	
Paternity												
Power of Exclusion	0.543	0.864	0.599	0.416	0.830	0.633	0.843	0.843	0.291	0.709	0.657	0.999995
Typical Paternity Index	2.170	7.500	2.500	1.630	6.000	2.750	6.500	6.500	1.250	3.500	4.030	
Allele Frequencies												
Homozygotes	0.231	0.067	0.200	0.308	0.083	0.182	0.077	0.077	0.400	0.143	0.177	
Heterozygotes	0.769	0.933	0.800	0.692	0.917	0.818	0.923	0.923	0.600	0.857	0.823	
Total Alleles	26	30	20	26	24	22	26	26	30	28	26	

TABLE 8.15: COMMONLY USED FORENSIC STATISTICS IN THE NON-ADMIXED ISLANDER POPULATION

Forensic calculations	10S520	12S297	1S407	2S262	3S1514	4S2285	5S592	7S1485	7S618	9S252	Average	Combined
Matching Probability	0.112	0.072	0.164	0.116	0.117	0.087	0.079	0.145	0.186	0.115	0.119	3.826E-10
Expressed as 1 in	9.00	13.90	6.10	8.60	8.50	11.50	12.60	6.90	5.40	8.70		2.61E+9
Power of Discrimination	0.888	0.928	0.836	0.884	0.883	0.913	0.921	0.855	0.814	0.885	0.881	
PIC	0.730	0.850	0.690	0.740	0.700	0.810	0.780	0.720	0.590	0.730	0.734	
Paternity												
Power of Exclusion	0.491	0.599	0.599	0.633	0.463	0.662	0.617	0.613	0.188	0.441	0.531	0.999633
Typical Paternity Index	1.920	2.500	2.500	2.750	1.800	3.000	2.630	2.600	1.000	1.710	2.241	
Allele Frequencies												
Homozygotes	0.261	0.200	0.200	0.182	0.278	0.167	0.190	0.192	0.500	0.292	0.246	
Heterozygotes	0.739	0.800	0.800	0.818	0.722	0.833	0.810	0.808	0.500	0.708	0.754	
Total Alleles	46	50	30	44	36	48	42	52	56	48	45	

Chapter 9

Discussion: How informative were the 10 autosomal tetranucleotide STR loci systems?

The aims of the autosomal short tandem repeat loci study were met. The ten autosomal tetranucleotide STR loci were optimized incorporating a suitable method of sizing the PCR product. The autosomal hypothesis that the autosomal marker systems will be able to distinguish between the U.K. Leicestershire and the Polynesian populations was proven. Furthermore the hypothesis that a greater polymorphic diversity would be observed among the U.K. Leicestershire population than the Polynesian Islander populations was also proven.

This discussion has been structured elaborating the aims and hypothesis by answering specific questions based on population and forensic genetic issues.
How informative were the ten autosomal microsatellite loci for use in population and forensic genetic investigations?
This important question was chosen for its obvious ambiguity, to provoke a discussion encompassing many different practical and statistical perspectives. Therefore, this discussion was structured answering the aforementioned question by incorporating a set of simpler, more poignant inquiries.

9.a. Sample sizes: Just enough or as many as possible?

At the forefront of any piece of research is the choice of which analytical systems to use, and which samples to incorporate. This process of decision making can often be quickly decided by a count of the available samples. Ideally a database of every humans DNA would provide the means to test hypotheses of population dynamics, disease and so forth. Notwithstanding the ethical considerations associated with it, the global organisation and co-operation that would be required makes it an impractical consideration.

Studying populations through polymorphic marker systems is problematic, as the chances of observing all possible genotypes including rare alleles at short tandem repeat loci is very small. In fact, with a sample size of 1 million individuals, the probability of observing all genotypes at the HUMHPRTB tetranucleotide STR locus (Edwards et al. 1991) is 0.5 (50%) (Chakraborty 1992). This estimation was given on the observation that 17 alleles were present in a sample size of 40 individuals (Edwards et al. 1991). If a sample size of 50 individuals was used, the probability was estimated at 2.06 X 10^{-37} at the HUMHPRTB locus. Similarly, the probability of observing alleles at an alternative locus (with 10 alleles present in a sample size of 50 individuals) was estimated at 4.83 X 10^{-7} (Chakraborty 1992). In this present study, a sample size of about 50 individuals were representative for the U.K. Leicestershire caucasian population, admixed and non-admixed New Zealand Maoris. Across all ten loci, between 6-15 alleles were observed. Thus following Chakraborty's (1992) guidelines, the probability of observing all possible genotypes ranged between 0.025 – 2.06 X 10^{-37} respectively. This simplistic measure indicated that sample size requirements were dictated by the number of possible alleles. This in turn was affected by the mutation rate of the specific locus and the population structure and substructure (Brinkmann et al. 1998). Chakraborty's (1992) statistic does not control for a level of inbreeding or substructure and so may bias the more genetically diverse population, regardless of sample size.

However, in support of the numbers of samples used in the present study, 50 individuals (100 alleles) or less have been successfully incorporated into previous polymorphic STR studies (Hammond et al. 1994, Morell et al. 1995, Evett et al. 1996 and Pritchard et al. 1999).

9.b. The information content of allele frequency distributions

A number of details can be drawn from allele frequency distributions within and between populations. These include; differences in modal allele frequencies between populations, the numbers of different alleles forming a normal, bimodal or skewed distribution and the possibility of null alleles. Although basic, these statistics form one of the first steps in analysing the potential of a particular microsatellite locus for use in population and forensic genetic analyses.

The loci with normally distributed allele frequencies had a greater number of similarities than differences between populations. The significance of this finding was two-fold. Firstly, the 10 autosomal markers can be used effectively as markers for forensic purposes within U.K. Leicestershire and Polynesian populations, as high heterozygosities and no alleles specific to either population were observed. However, secondly, because one can not discriminate between the two populations at the allele distribution level, population genetic studies to discriminate between these populations would not be possible.

The numbers of different alleles across the loci varied with respect to the sample population. At seven of the ten autosomal loci the greatest number of different alleles were observed in the U.K. Leicestershire and admixed Maori populations (see autosomal results section). Lower numbers of alleles in the non-admixed Maori and Polynesian Islander populations may be consistent with the reported founder effects during the colonisation of Polynesia (Hurles et al. 1999; Hagelberg et al. 1999). Furthermore, since this observation is not consistent across all loci it was possible that a selective sweep had occurred reducing allele numbers, hence genetic diversity (Jorde et al. 1995). However, the lower numbers of different alleles observed in the U.K. Leicestershire population in comparison to the non-admixed Maori population at the D4S2285 locus (see results), was most likely to be a direct result of varying sample size. Indeed, over three times the number of samples were typed in the non-admixed Maori population in comparison to the U.K. Leicestershire population. This biased the non-admixed Maori population in favour of observing rare and 'infrequent' alleles.

Interesting contrasts existed between the allele frequency distributions between loci. All loci and populations exhibited a normal distribution except at D12S297 and D4S2285 (see autosomal results section). The D12S297 and D4S2285 loci both had a high frequency allele, which was five to seven repeat motifs smaller than the normal frequency distribution of the remaining alleles. Various theories exist to explain these findings. Firstly, the high frequency of one allele not falling within the normal frequency distribution may be the result of poor allele resolution during analysis (see Deka et al. 1995). However, if this was true one may expect to observe the alleles with one or two repeat motifs different to the anomalous allele at a higher frequency and an abnormally high proportion of homozygosity. This was not an observation supported by this study. Furthermore, the amplified D12S297 and D4S2285 loci produced alleles ranging 205 to 305 base pairs, which is the optimal range for resolution using Elchroms spreadex gel system as used in this study (see methodology section). Therefore, it was highly unlikely that incorrect allele calling within the optimal resolution range of the gel caused the anomalous results. Also to this point, the anomalous alleles were the smallest or second smallest recorded alleles within the system and so would have migrated through the gel faster achieving better separation than the higher molecular weight alleles.

Secondly, similar allele frequency distributions have been observed at the D12S297 locus within other populations (Jorde et al. 1997, Lum et al. 1998). In particular, and unsurprisingly a Samoan population (Lum et al. 1998) held close frequency distribution similarities to the non-admixed Islander and Maori population of this study (see figure 9.1). The corroboration of results from this research and Lum et al.'s (1998), clearly shows not only that the method of analysis chosen for this study is reliable, but also that this locus has distinguishing features, consistent across genetically diverse populations.

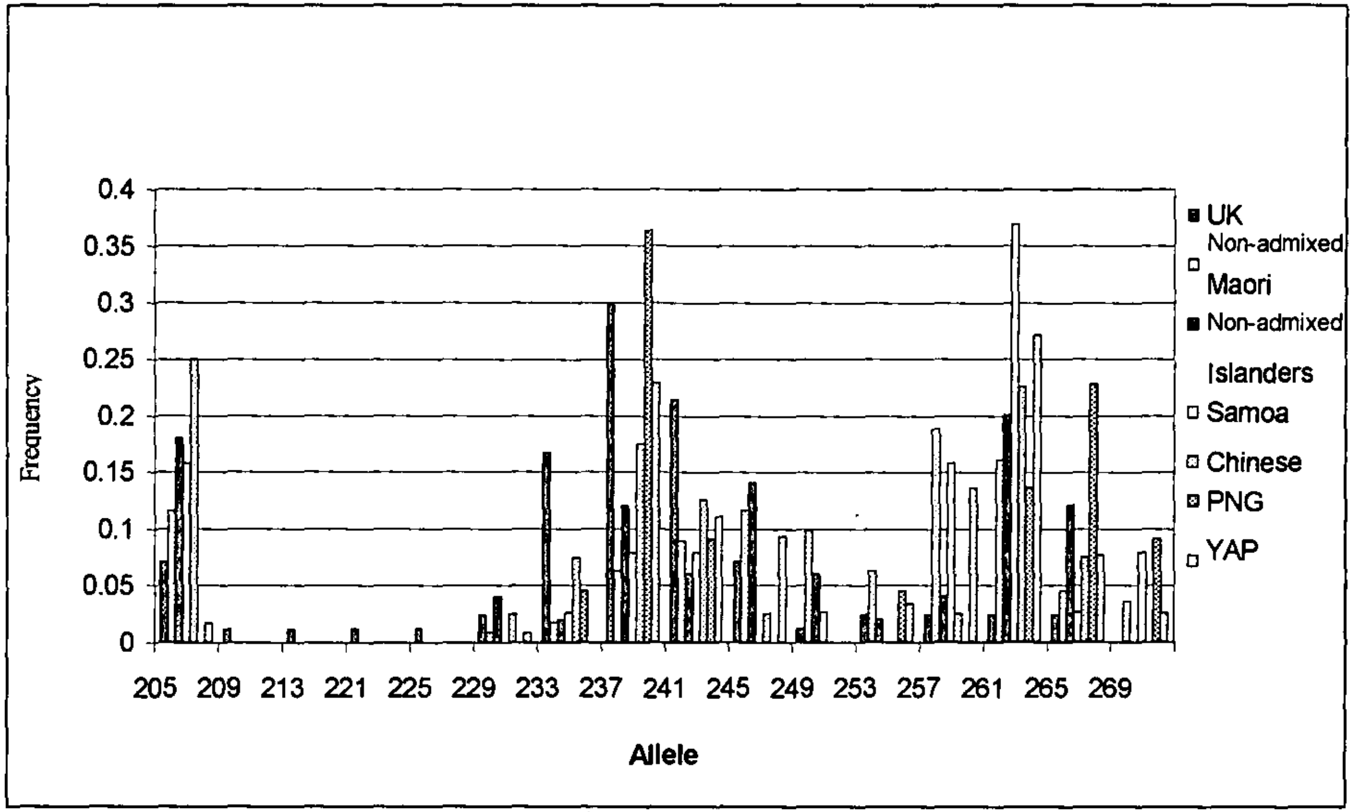


FIGURE 9.1: ALLELE FREQUENCY DATA AT THE D12S297 LOCUS (INCORPORATING DATA FROM THIS STUDY AND FOUR GENETICALLY DIVERSE POPULATIONS TAKEN FROM LUM ET AL. (1998)). POPULATIONS FROM LUM ET AL. (1998): SAMOA, CHINESE, PNG (PAPUA NEW GUINEA) YAP (ISLANDERS)

Thirdly, the frequency distributions as observed at the D12S297 and D4S2285 loci may be described as bimodal. Bimodal distributions of alleles may indicate the presence of subgroups of alleles within the sample population, with differences in sequence and arrangement of repeats (Brinkmann et al. 1998).

Examination of sequenced D12S297 alleles of length 205 and 265 base pairs indicated differences in sequence and arrangement of repeats as described by Brinkmann et al. (1998).

Half way through the D12S207 sequence (at about 100 base pairs) the 205 base pair allele had a more complex base sequence than the longer allele. Observed within the sequenced shorter allele was a section of base sequence 'TTTAAATTGTGTA', which was inserted within the repeat motif region. This section of DNA sequence was not observed in the longer allele that had consistent arrays of repeat motifs. Thus, the lack of sequence homology between the shorter allele and the longer allele may restrict the potential for DNA slippage during replication, more-so, than between longer alleles with a greater sequence homology. The base pair sequence at the D12S297 locus is in agreement with Brinkmann et al.'s (1998) findings, that interrupted repeat motifs mutate less often and that homogenous repeats encourages mutations. Furthermore, polymerase slippage during replication occurs more frequently in longer arrays (Schlotterer and Tautz 1992), and the mutations tend toward a positive asymmetry (Goldstein and Pollock 1997). Thus, it was a reasonable assumption that the longer alleles at the D12S297 and D4S2285 loci were formed by positive asymmetric mutations, distancing themselves from the shorter allele.

9.c. Relationships between allele sequence and number of alleles and sizes

The relationships between the allele sequences, the number of tandemly repeated motifs and number of alleles were examined. Brinkmann et al. (1998) purported that the sequences of alleles with irregular units interspersed within the tandemly repeat motifs, mutated less often, than the 'homogenous' repeat structure, where slippage during replication would be easier. However, this study did not reflect Brinkmann et al.'s (1998) observation. In fact greater numbers of different alleles were observed at loci, which held complex sequence structures. For example, the greatest number of different alleles was at D12S297, with twelve alleles, which as previously stated, had one of the most complex allele sequences. In contrast, the simplest allele sequence structure was at the D5S592 locus, where eight alleles were isolated. Thus, one may argue that if two or more tandemly repeated regions within a locus exist, the 'chances' or 'possibilities' of mutation during replication are increased and not decreased. However, a direct count of the numbers of alleles at a frequency above 10%, within a population revealed the opposite. Loci with a simple tandem repeat structure had on average greater numbers of alleles with frequencies above 10% of the population. Thus, one could conclude that greater numbers of rare or less frequent alleles were observed at loci with complex sequence alleles. Therefore, this study corroborates Brinkmann et al's (1998) theory, if alleles with frequencies above a 10% threshold are included and rare or infrequent alleles are discounted.

9.d. Descriptive answers

The informativeness of the polymorphic marker loci, was further investigated through the use of descriptive statistics. These included the heterozygosity index, unbiased gene diversity estimates and their standard errors (Nei 1987), interlocus and intralocus variances (Nei 1987), polymorphic information contents, powers of discrimination and exclusion and paternity indices. The former statistical calculations were previously used to describe microsatellite data (Nei 1987) for use in population and forensic genetic analyses (Hammond et al. 1994). The latter analyses were previously used with forensic calculations, in particular the powers of discrimination and exclusion and paternity indices (Pastore et al. 1996, Gill and Evett 1995).

9.e. Diversity issues

The unbiased gene diversity differs from the gene diversity by incorporating sample size into the equation (Nei 1987). This study observed a range of unbiased gene diversities across loci and populations. The D12S297 locus had on average the highest unbiased gene diversity ranging from 83% (U.K. Leicestershire population) to 96% (non-admixed Maori population). This was not surprising given that this locus had the greatest number of different alleles. Conversely, D7S618 had on average the lowest unbiased gene diversity ranging from 65% (non-admixed Islander population) to 81% (U.K. Leicestershire population). The ranges of gene diversities observed in this study were similar to those of other tetranucleotide marker systems (Edwards et al. 1992, Hammond et al. 1994, Jorde et al. 1997).

The DNA samples provided by Dr Chambers, Wellington University, New Zealand, were also incorporated into the Forensic Science Service (FSS) database. The FSS analysed the samples using six of their multiplex tetranucleotide markers (Overall personal communication). The unbiased gene diversity for each population at each locus is given in table 9.1. This clearly expressed consistencies of gene diversities, not only between the FSS markers but also the ten autosomal markers used in this study (see chapter 8: results section).

Population	D18	D21	HUM THO1	D8	FGA	VWA	Sample Number
Polynesian	0.874	0.813	0.752	0.814	0.806	0.800	35
W Samoan	0.885	0.795	0.772	0.811	0.780	0.820	19
Maori	0.793	0.808	0.653	0.816	0.799	0.794	41
Cook Islanders	0.754	0.791	0.740	0.796	0.742	0.825	18
Mixed Maori	0.819	0.781	0.732	0.818	0.826	0.809	27

TABLE 9.1: UNBIASED GENE DIVERSITIES AT SIX TETRANUCLEOTIDE STR LOCI

As part of a large study, Lum et al. (1998) investigated the D1S407 and D12S297 loci within the Samoan and Papua New Guinean (PNG) populations. Gene diversities of 67% and 58% at D1S407 were observed in the Samoan and PNG populations respectively (Lum et al. 1998). Similarly, at D12S297 the gene diversities were 79% (Samoan) and 78% (PNG) (Lum et al. 1998). The gene diversity of the Samoan population at D1S407 was slightly lower than the observed value of 73%, of the non-admixed Islander population in this study. However, this was anticipated as the non-admixed Islander population incorporated Tongan, Tokelau and Rarotongan samples, although 61% of the Islander population were Samoan in origin. The gene diversity of the Samoan population at D12S297 in Lum et al.'s (1998) study was comparable to the observations of this study.

Average gene diversities and their sampling variance and standard errors across loci were calculated for each population in this study (see results table 8.5). The standard errors of the heterozygosity values were calculated using equation 8.7 in Nei (1987). These errors were observed to be within previous standard error estimates (Jorde et al. 1997, Nei 1987). The sampling variance across loci was smallest in the U.K. Leicestershire population (0.0001) and highest in the non-admixed Islander population (0.0006). The admixed Polynesian populations had lower sampling variances than the non-admixed populations. The significance of this may be that the sampling variances and standard errors of the non-admixed populations were higher than the U.K. Leicestershire and admixed Polynesian populations due to the differences in the population structure. The founding populations in Polynesia have experienced genetic bottlenecks and founder effects during the last 1000 – 5000 years before present (Irwin 1989, Hagelberg et al. 1999, Hurles et al. 1998). In comparison, the more diverse U.K. Leicestershire population may be a direct reflection of the varied and long-standing population history and prehistory as a region of northern Europe (refer to chapter 3).

In retrospect, the gene diversity statistics were a good indication that the autosomal marker loci used in this study, effectively differentiated between the U.K. Leicestershire, admixed and non-admixed Polynesian populations. This was in agreement with the known population structures and interactions between them (refer to Irwin 1989, Hurles et al. 1998, Lum et al. 1998, Jorde et al. 1997 and Hammond et al. 1994).

9.f. Variances

Interlocus and intralocus variances were also calculated. The interlocus variance was described as 'the variance associated with the sampling of loci from the genome' (Nei 1987). In accordance with the gene diversity and standard error, the interlocus variance increased from 0.0006 in the U.K. Leicestershire population to 0.0633 in the non-admixed Islander population (see results table 8.7). To reduce this variance greater numbers of loci need to be analysed (Nei 1987). Thus, from this basic statistic, one can observe that the ten autosomal loci were more informative in the U.K. Leicestershire and admixed Polynesian populations than the non-admixed Polynesian populations. The autosomal loci may have been less informative within the non-admixed populations, either because of the low number of chromosomes analysed or, due to the effects of a population bottleneck occurring during colonisation.

Comparing the admixed Maori and non-admixed Maori populations, this study observed a slight increase in sampling variance, whilst the average gene diversity decreased 2%. Similar comparisons between the U.K. Leicestershire population and non-admixed Maori revealed a slight increase in sampling variance. Whilst the average gene diversity decreased 4%, from 83.5% within the U.K. Leicestershire population to 79.5% in the non-admixed Maori population (see table 8.5: autosomal results section).

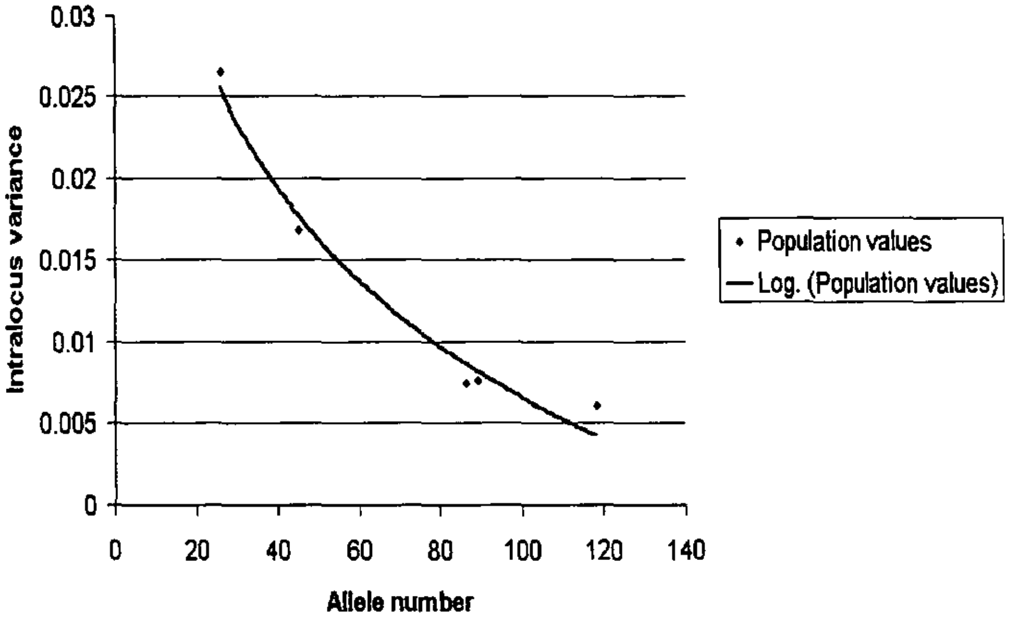
The intralocus variance was described as 'the sampling of individuals at each locus' (Nei 1987). This variance directly reflected the total number of samples per population. To reduce variance, a greater number of samples must be analysed (Nei 1987). The greatest variance (0.0265) was observed in the admixed Islander population with an average allelic sample number of 26 and the smallest variance (0.0061) was observed in the non-admixed Maori with an average allelic sample number of 118.

Examination of the relationship between allele number and intralocus variance within the autosomal marker systems used in this study (figure 9.2), indicated a sharp decrease in variance between 20 to 80 alleles. Thereafter, the variance began to taper, and at 120 alleles was below 0.005.

9.g. Sample sizes: Effects on diversity and variances

Expectedly, warnings were made against the use of extremely small sample numbers. Whereby, small sample numbers increased the standard error of average gene diversity and (as observed in this study) intralocus variance (Nei 1987). Jorde et al. (1997) incorporated 60 microsatellite loci across three defined ethnic populations with sample sizes between 60 – 120. The standard errors of heterozygosity were all 0.02, which were the same or greater than this present study. Nei (1987) recommended at least 20 individuals to be examined across about 20 loci. Although this study has examined 10 autosomal loci, the numbers of individuals ranged from 13 to 60. The lower sample numbers were observed in the admixed and non-admixed Islander populations. Although the statistics generated from the non-admixed Islander populations had not deviated from any 'expectations', further interpretations should be met with the understanding that they were derived from a 'small' sample population.

Figure 9.2: Relationship between allele number and intralocus variance



The polymorphic information content (PIC) for each population at every locus was calculated (see results section). In general, just as the gene diversities varied according to the population and locus, so did the PIC. As anticipated from the gene diversity estimates, the descending order of averaged PIC across loci was; U.K. Leicestershire (0.795), admixed Maori (0.771), admixed Islander (0.758), non-admixed Maori (0.750) and non-admixed Islander (0.734).

9.h. Matching probability

The individual matching probabilities were calculated following the formula of Jones (1972), and the combined matching probability of the ten STR loci was calculated using the product rule, following the methodology of Kimpton et al. (1993). The combined matching probabilities were: 6.1 X 10^{-12} in the U.K. Leicestershire population, 1.5 x 10^{-11} in the admixed Maori population,
5.8 X 10^{-12} in the non-admixed Maori population, 1.0 X 10^{-9} in the admixed Islander population and 3.8 X 10^{-10} in the non-admixed Islander population. These values were comparable to the loci used by the U.K. Forensic Science Service, whereby a multiplex incorporating ten loci had a combined matching probability of 3.1 X 10^{-11} in a U.K. Caucasian population. Similarly, Urquhart et al. (1995) combined heptaplex (seven locus marker system) and quadraplex (four locus marker system) systems producing a combined matching probability of 2.9 X 10^{-12} in a caucasian U.K. Leicestershire population. A heptaplex system was purported to be sufficiently discriminatory for most forensic applications and a ten or eleven locus system would only be used when an extremely high discriminating power was required (Urquhart et al. 1995). Evett et al. (1996) further observed that the power of discrimination of even a four locus STR profile, would be sufficient to ensure the profile was unlikely to be contained in a forensic database.

To place a significance of the discriminatory potential of the ten autosomal loci in this study, a matching probability of 6.1 X 10^{-12} estimated in the U.K. Leicestershire population could be expressed as 1 profile in 0.16 billion. A population census in the year 1999, estimated an English population size of approximately 48 million. Thus, the discriminatory potential of this system, exceeded the population size. Similarly, the matching probability of

1 X 10^{-9} estimated in the non-admixed Maori population could be expressed as 1 profile in 1000 million. A population census of New Zealand in the year 1999, estimated the Maori population size to be 355 thousand. Thus, once again, the discriminatory potential of this system exceeded the population size.

Therefore, when these ten loci are analysed together, they form a highly discriminatory system exceeding the existing 'potential' of the Forensic Science Service system currently employed.

9.i. Power of exclusion

This is the ability of a locus or set of combined loci to distinguish samples from different individuals (NRC 1996). As anticipated the powers of exclusion varied according to the population and locus studied, with the highest combined power of exclusion observed in the U.K. Leicestershire population (0.999989) (see results section). However, on average the ten autosomal STR systems used in this study exhibited greater powers of exclusion than those used by the Forensic Science Service (Pastore et al. 1996). This was reflected by the combined power of exclusion for their 11 STR loci (0.9993) (Pastore et al. 1996).

Thus one could conclude that the 10 autosomal STR systems used in this present study, were more informative than other STR systems routinely used in forensic casework (see Pastore et al. 1996).

9.j. Paternity Indices

Comparisons made between the paternity indices of this study and those of Pastore et al. (1996) indicated the greater discriminatory potential of the systems used in the present study (ranging from 1.44 at D1S407 to 4.5 at D3S1514 among the U.K. Leicestershire population) over those chosen by the Forensic Science Service (ranging from 1.0 at Hum RENA4 to 3.6 at HumARA among an Italian Caucasian population). The greatest difference was observed between the U.K. Leicestershire population and Pastore et al.'s (1996) 'white' Caucasian population. Large differences in paternity indices were also observed between the non-admixed Islander and non-admixed Maori populations in comparison to the genetically diverse populations of Pastore et al.'s (1996) study.

The differences not only observed between paternity indices but also Powers of exclusion and discrimination may stem from mutational differences among the polymorphic loci.

It was clear that the loci incorporated in the present study were highly informative, even in the less genetically diverse populations of the Maori and Polynesian Islanders. In comparison to the present study, the sample sizes used by Pastore et al. (1996) were larger, thus, it was highly unlikely that Pastore et al. (1996) had not at least detected the most common if not rarer alleles. Therefore the bias of the loci in the present study being more informative, was not a result of direct differences of sample sizes. Furthermore, Pastore et al.'s (1996) findings were consistent previous studies, incorporating the same marker systems (Urquhart et al. 1995, Evett et al 1996).

9.k. Hardy-Weinberg proportions

The Hardy-Weinberg law describes the relationship between individuals of a sample population by virtue of their observed gene frequencies. It is a test of the assumption that the sample population mates at random, and the allele frequencies are at a proportionate equilibrium that remains constant from generation to generation (Sensabaugh 1982).

Two independent tests (Exact test and Chi-square measure) of Hardy-Weinberg equilibrium were computed using the GDA software (Lewis and Zaykin 1999). The exact test or probability test was a preferred measure as small sample sizes or low frequency alleles were expected (Weir 1996). The Chi-square measure followed the standard formula, although through random re-shufflings of the data a probability ('p') value was generated (Weir 1996).

The use of more than one test to measure Hardy-Weinberg equilibrium (HWE) has been implemented by other workers (Evett et al. 1996, Hammond et al. 1994, Edwards et al. 1992). The different tests hold their own strengths and weaknesses. Thus, by incorporating more than one test, the opportunity to detect deviations from HWE was increased (Hammond et al. 1994). Furthermore, Evett et al. (1996) observed that if there was no consistency between tests that a significant departure from HWE existed at a specific locus, then the significance level occurred by chance, possibly due to a sampling phenomenom.

This study reported 5 out of 50 significant departures from HWE using the exact test and 5 out of 50 significant departures using the Chi-square measure (see chapter 8: results section). Of the five departures, three were observed to be consistent between tests. These were at D4S2285 within the U.K. Leicestershire population and D1S407 within the non-admixed Maori population and D12S297 within the non-admixed Islander population. The significant departure from HWE within the non-admixed Islander population may be attributable to a very low sample number at the D12S297 locus, thus biasing the results (see results section). Departures within Caucasian populations have been observed previously (Evett et al. 1996). This study observed a significant deviation at the

D4S2285. Closer examination of the allele frequency distribution at this locus revealed this to have a bimodal distribution of alleles (for a detailed report on the allele frequency distribution, see the section on allele frequency distributions this chapter). Nine alleles were observed at this locus within the U.K. Leicestershire population, although almost 25% of the data was described by one allele. However, this was no different to the other populations, where no significant deviations from HWE were observed. Furthermore, there was no significant deviation from HWE at any of the other loci within the U.K. Leicestershire population. Therefore, following Evett et al.'s (1996) philosphophy, the deviation from HWE at D4S2285 was possibly a chance event, caused by a sampling phenomenon.

The significant departure from HWE at D4S2285 within the non-admixed Maori population may have been the result of two interacting factors. Firstly, six different alleles were observed, whereby, 90% of the data was described by three alleles. Hence, the remaining 10%, was distributed between three alleles. Therefore, the sharp clines in allele frequencies may have caused the anomalous result. Secondly, a null allele was observed at this locus (see results section), possibly not detected due to its very low frequency or not correctly identified within the sample population (Edwards et al. 1992).

Interestingly, with regard to the D1S407 locus, a deviation from HWE was observed within the non-admixed Maori population which may have been partly attributable to its possible linkage to a functional gene. Within the D1S407 locus, a gene was isolated, which was thought to function as a tumour suppressor gene, for a rare prostate cancer (Gibbs et al. 1999). Often, linked genes exhibit a loss of heterozygosity (Xin et al. 1999). Thus, there may have been a selective advantage for particular sized alleles and or sequences. Re-examination of this locus did not reveal any obvious anomalies that would indicate linkage to any gene. Although <50% of the population data was described by two alleles, a similar result was observed at other loci (see chapter 8: results section). Furthermore, a random sample of individuals was used in this study, half of which were males. It was highly unlikely, that an appreciable fraction of non-admixed Maori males would express a loss of heterozygosity at the D1S407 locus that would cause a significant departure from HWE.

Overall, there was no consistency between significant deviations from HWE at loci or across populations. Since no significant deviations were present across all loci or the majority of loci within a particular population, there was no indication of sub-structuring or levels of inbreeding (Nei 1987).

9.1. The F statistic measures

F_{ST} has been described as 'a genetic distance measure, usually used to estimate migration or separation time in geographically structured populations that focuses on changes in gene frequency caused by 'genetic drift' (Goldstein and Schlotterer 1999).

In this study differences of F_{ST} values were observed across populations and loci. The observed variations may have arisen as a result of differences in selection or mutation rates (Balding et al. 1996). The U.K. Leicestershire population in this study had F_{ST} values in the range –0.6% (D7S1485) to 1.3% (D3S1514) across the ten STR loci. Similar F_{ST} values between 0.25% to 1% within U.K. Caucasian populations were supported by other workers incorporating tetranucleotide STR loci (Balding et al. 1996).

A higher range of F_{ST} values were observed within the non-admixed Maori population (-0.2% at D9s252 to 2.2% at D10s520) of this study. This finding was consistent with that of a small isolated population (Balding et al. 1996) and so is in agreement with the known prehistory of the founding populations in New Zealand (Cowan 1930). It should be observed that F_{ST} can be a measure of 'inbreeding' through the interpretation that two random alleles were identical by descent (Hartl 1981). Therefore, the larger F_{ST} values, as observed with the non-admixed Maori population, may indicate that although random mating was occurring, the effective population size was so small that non-random mating was interpreted from the results.

Interestingly, negative F_{ST} values were observed across all populations at the D9S252 locus. This may be the result of a high mutation rate (Balding et al. 1996) at the D9S252 locus. The allele frequency distributions at this locus indicated stronger similarities than differences between the populations. The distribution was skewed with 3-4 different alleles isolated that were larger than the modal allele and 1-2 different alleles, which were smaller than the modal allele. Thus, the frequency distribution at the D9S252 locus was consistent with the theory that alleles mutate with an upward length bias (Rubinsztein et al. 1995, Amos and Rubinsztein 1996, Amos et al. 1996).

In summary, the F_{ST} values of the present study fell within the ranges of other F_{ST} values calculated using Caucasian populations (Gill and Evett 1996, Balding et al. 1996). This indicated that these loci were accurately reflecting the diverse population structures as one would anticipate given the known prehistory of the populations (Cowan 1930, Renfrew 1974).

9.m. FIS Values

The FIS measured the reduction in heterozygosity of an individual, through non-random mating within its subpopulation (Hartl 1981). This study calculated FIS values per population per locus (see results chapter). Interestingly, the lowest average within population inbreeding value was observed within the non-admixed Maori population. Thus, the high F_{ST} value observed within the non-admixed Maori population was possibly not caused by inbreeding, but the result of a small sized population (see Hartl 1981). In general the FIS values were larger than their respective F_{ST} values. This 'trend' has previously been observed among tetranucleotide polymorphic marker systems (Gill and Evett 1995).

9.n. Genetic distance measures:

Do these loci accurately reconstruct the genetic relationships between the populations, given the known history of the populations?

Two measures of genetic distance were used to describe the data; Nei's 1978 distance measure and the Coancestry distance measure (pairwise Fst). These measures have been successfully used with STR data in previous population studies (PerezLezaun et al. 1997a, PerezLezaun et al. 1997b).

In the present study, the distance measures were calculated using all the data (see autosomal results section). The distance measures were incorporated to test the ability of the ten autosomal systems to define a clear divergence between the U.K. Leicestershire and Polynesian populations. Furthermore, the inclusion of admixed Maori and Polynesian populations provided the opportunity to assess the potential of the STR systems to distinguish between closely related populations. The distance measures were not used to form assumptions of genetic evolutionary relationships. Instead, the loci were investigated to assess how well they could reproduce the relationship between the populations. Of particular interest was the relationship between admixed Polynesian populations to the other populations.

The coancestry distance was analysed as it only measured genetic divergence due to drift, with no assumptions about the ancestral population. Coancestry measures have been used as short-term genetic distances between populations (Slatkin 1995).

Nei's 1978 distance measure was built upon Nei's (1972) standard distance measure, although incorporated a bias correction (Weir 1996). Nei's distance measure was chosen to examine the divergence between the U.K. Leicestershire population and the Polynesian populations, since this measure was purported to be appropriate for long-term evolution when populations diverged because of drift and mutation (Weir 1996, Pp197). Thus, the difference between the two distance measures resided in the fact that the coancestry distance accounted for divergence due to drift and Nei's 1978 distance measured divergence due to drift and mutation.

Both distance measures produced similar relationships between populations (see chapter 8). The coancestry distance measure was previously recognised as a suitable measure for use with tetranucleotide markers (PerezLezaun et al. 1997a). Although it did not consider different mutational relationships among alleles, it correctly reflected the known archaeological and anthropological history of the populations, more so, than mutation-based distances (PerezLezaun et al. 1997a). Furthermore, genetic drift (PerezLezaun et al. 1997a) was considered the main factor generating the present distributions of microsatellite alleles as the Polynesian populations in particular, may not have had enough time (in generations) to accumulate significant divergences due to mutations.

In this study, negative distance values were observed between the admixed Islander population and the other Polynesian populations. This indicated a very close genetic relationship between the populations (Weir 1996). However, this must be viewed rather sceptically due to the low allele numbers of the admixed Islander population, which would have biased the result.

Through the genetic distance measures, the autosomal marker systems could clearly discern between the admixed and non-admixed populations. The distance measure also, indicated that closer genetic relationships existed between the admixed populations (of mainly European admixture) and the U.K. Leicestershire population. European admixture within the Polynesian gene pool of the present study has also been isolated with varying frequencies in previous studies (Clark et al. 1995, Hurles et al. 1998, Hagelberg et al. 1999).

European admixture can be traced back as far as the middle of the twentieth century, beginning with the voyages made by Magellan and Captain James Cook (Beaglehole 1934). Through subsequent and periodic execution of the native Polynesians and Maoris' (Cowan 1930), it was estimated that the Maori gene pool contained only 55% of 'native' DNA (Hurles et al. 1998). In total, 42% of the Maori samples in this present study originated from persons who regarded themselves as admixed and 91% of these admixed people had European genetic connections. Upon re-examination of the distance measurements, the admixed Maori population value (for either distance measure) was approximately half the non-admixed Maori population value, thus midpoint between the Maori and U.K.

Leicestershire populations. This may have been a reflection of the extent of European admixture within the Maori gene pool attributed by either parent.

9.n.i. Phenogram constructions of the distance measures

The unweighted pair group method (UPGMA) and neighbor joining methods of phenogram construction were chosen to describe the distance matrices. The UPGMA and neighbor joining methods describing distance matrices of STR data have been successfully incorporated into other studies (see Jorde et al. 1995, Chu et al. 1998, Hagelberg et al. 1999, Ruiz-Linares et al. 1999).

The UPGMA tree construction of the coancestry distance matrix clearly separated the U.K. Leicestershire population from the Polynesian populations. The admixed populations were grouped together, indicating their close genetic similarity. The distance or 'node' created was at a level of 0.000, also reflecting the close genetic relationship between the admixed populations. This was not surprising given the extent of European admixture within each group. The non-admixed Maori population was next to be grouped to the admixed populations, by virtue of the way in which the UPGMA clusters populations with the smallest distance (Nei 1987). The non-admixed Islanders then clustered to the other Maori and Polynesian populations, although the distance or 'node' created was at a level of 0.0016, indicating less genetic similarity to the admixed populations than the non-admixed Maori population (node level 0.0003). The reasons for this observation may be two fold. Firstly, the non-admixed Maori population may have a percentage of undetected admixture, thus biasing the population to cluster more closely to the admixed populations. Secondly, the non-admixed Polynesian Islander population contained small sample numbers in comparison to the non-admixed Maori and this may have biased the genetic distance and subsequent UPGMA construction.

Finally, the U.K Leicestershire population clustered to the other populations expressing the deepest separation. The obvious lengthy separation of the U.K. Leicestershire population indicated a very great genetic difference between these populations. This also indicated, as one would have expected a long divergence time since a common ancestry. This was an interesting observation, since there was little genetic similarity between the admixed populations (with known European admixture) to the U.K. Leicestershire population, especially when other measurements, for example, allele frequency distributions and gene diversities intimated genetic similarities. However, one must not forget how the UPGMA method constructs the trees by clustering the closest genetic distances first. An amalgamation of the clustered groups forms a unit to which the remaining populations were compared. The next closest distance to the amalgamated group then joins the tree at a specific level (Nei 1987, Weir 1996). Thus, the U.K. Leicestershire population would have been compared to the amalgamated Polynesian Islander and Maori populations, which as a collective had clear genetic distinctions to the U.K. Leicestershire population.

The neighbor joining phenogram of the coancestry distance measure gave a very different result. This method identified closest pairs or neighbours (populations) in a manner to minimise the total length of a tree. The sum of the lengths of the branches joining the populations equated to the distance between them, if the data was additive (assuming constant evolutionary rate) then the tree would be the minimum evolution tree (Weir 1996). The data in this study was possibly not 'additive' as the evolutionary rates of the Polynesian Islander and Maori populations may differ in comparison to the U.K. Leicestershire population.

Interpretation of the unrooted neighbor joining phenogram, clustered the Polynesian Islander and Maori populations closer to each other than to the Caucasian population. The admixed Islander population paired, although at great distance, to the U.K. Leicestershire population. This association was possibly due to the European admixture present in the admixed Islander population. However, this may be a tenuous link given the small sample size of this population and the aforementioned biases associated with it. Similarly, to the UPGMA tree, the remaining populations clustered in the order of admixed Maori, non-admixed Maori and lastly non-admixed Islander (see figure 8.12: autosomal results section).

The UPGMA and Neighbor joining methods of tree construction used to describe Nei's (1978) distance matrix, produced identical tree topologies to the coancestry distance matrix. The only observable difference was the length of the overall tree, analogous to an evolutionary time measure. Both the UPGMA and NJ trees were over four times longer describing Nei's (1978) distance in comparison to the coancestry distance measure. The difference in divergence times may be a reflection of the underlying assumptions of the distance measures used. Nei's (1978) distance was appropriate for long-term evolution assuming populations diverged because of drift and mutation (Weir 1996). Thus, applying Nei's (1978) distance measure to the Polynesian populations (and in particular where this study incorporated admixed populations), the divergence times may have been overstated, as an assumption that mutations occurred over a long-term evolutionary period (Weir 1996). This assumption would be inappropriate given the known relatively recent colonisation of Polynesian populations in the Pacific (see chapter 4).

Further distance measures (RST and $\delta\mu^2$) were investigated and UPGMA and Neighbor joining trees constructed. However, none of these measures produced a sensible tree construction that corroborated any findings from this present study, or from other studies incorporating European admixture (Hurles et al. 1998, Hagelberg et al. 1999). The RST and $\delta\mu^2$ distances did not show a consistency of tree construction between them or in comparison to either the coancestry or Nei's (1978) distance measures. The lack of consistency was surprising, since the RST and $\delta\mu^2$ measures follow the stepwise mutation model, which has been purported to be of more use with STR analyses (Goldstein et al. 1995, Goldstein and Pollock 1997). However, the lack of sensible results using the aforementioned distance measures has been observed elsewhere (Perez-Lezaun et al. 1997a). The aforementioned issues may be less surprising, given an important point raised by Perez-Lezaun et al. (1997a). These workers concluded that the lack of consistency between the RST and $\delta\mu^2$ measures suggested that genetic drift was the main cause, generating the present allele frequency distributions. Furthermore, mutation must have been less important because of the time constraint imposed on the populations (Perez-Lezaun et al. 1997a). Applying this theory to the present study it was obvious the same dynamics of genetic drift versus mutation were present. The archaeological (Ward 1972), anthropological (Terrell 1986) and genetic evidence (for a detailed synopsis, see Polynesian prehistory introduction), all tended towards a recent colonisation of the Pacific, whereby genetic drift was a strong influence on the allele frequency distributions today. Thus, distance measures that primarily measure for genetic divergence due to drift, such as the coancestry measure, would indeed produce a more sensible tree diagram than those primarily concerned with mutational models.

In summary, the autosomal loci were able to accurately describe the relationships between the populations, using the Coancestry distance and Nei's (1978) distance measures. Serious consideration was given to the appropriateness of the distance measures used. Distance measures incorporating the stepwise mutation model were observed to be suitable for constructing distance matrices using STR data (Goldstein et al. 1995, Takezaki and Nei 1996 and see chapter 7: statistical introduction). However, in this study, distances based on mutational models did not give sensible results, in comparison to the archaeological (Ward 1972) and anthropological (Terrell 1986) records of Polynesia. Thus, the distance measures assuming divergence due to drift (Weir 1996, Nei 1987), were observed to be more appropriate.

9.o. In summary

Overall, this study has provided the first informative and thorough examination of ten autosomal polymorphic marker loci, in U.K. Leicestershire and Polynesian populations. No previous studies have incorporated all ten systems within such genetically diverse populations.

There was no strong indication that these loci deviated from HWE, or that with the known population prehistory, they do not behave irregularly with respect to describing the structure and variation of the genetically diverse populations. Moreover, the powers of discrimination matched if not improved upon those presently used by the Forensic science service.

FST measures reflected the ability of the autosomal markers to accurately describe the population structures given the known prehistory of each population. The genetic distance measures used to define the relationships between the populations indicated that the autosomal loci in this study were able to distinguish between closely related populations.

Further research could be applied to this study to develop multiplex PCR reactions to reduce analysis time. Also, a more in depth study comparing specific allele sequences to biases in mutation length and frequency may provide additional information on loci deviating from a normal distribution. For example, as observed at the D12S297 locus.

Chapter 10
Y Chromosome Results

10.a Y Chromosome Results: Haplotype and Haplogroup analyses

Previous studies of populations in the south Pacific have analysed male lineages among Polyneisan and Melanesians (Hurles et al. 1998, Forster et al. 1998) and have observed close genetic similarities between the groups.
This present study analysed male lineages at a number of levels, within and between the New Zealand admixed Maori (Y chromosomes of non- native Maori descendency) non-admixed Maori, admixed Polynesian Islanders (Y chromosomes of non-native Polynesian descendency), non-admixed Polynesian Islanders and the U.K. Leicestershire Caucasian populations.

The Y chromosome statistical analyses were chosen so that as much information as possible, with respect to the inter- and intrapopulation similarities and differences, could be extracted. These analyses included;

- Comparisons of allele frequency distributions among populations
- Haplogroups of the unique event polymorphisms (UEPs) were constructed following the work of Jobling and Tyler-Smith (1995) also incorporated by Hurles et al. (1998), so that one could make comparisons between studies.
- Associations between specific UEPs and specific microsatellite alleles were examined. In particular, associations between the UEP M9 and the microsatellite loci DYS390 and DYS388.
- Examination of unbiased gene diversities between populations.
- Differences between haplotypes within haplogroups among populations. In particular, identifying modal and shared haplotypes among populations. This included 'Network' analyses of microsatellites identifying clustered groups of haplotypes.

- Variance analyses within and between populations to understand the apportionment of variance attributable to differences within and among populations.
- Genetic structure and migration analyses provided a means to statistically evaluate the genetic similarity or dis-similarity between two specific populations.
- Finally, the common ancestry of the New Zealand Maori 'ancestral' haplotype was evaluated to ascertain an approximate dating of the 'ancestral' modal haplotype cluster.

10.b. The Y Chromosome

A total of 25 U.K. Leicestershire Caucasian, 48 Maori (inclusive of admixture) and 18 Polynesian Islander DNA samples (inclusive of admixture) were assayed using six microsatellite and eleven UEP (or biallelic) polymorphisms, following the methodology of Thomas et al. (1999). Eighty-eight chromosomes gave full microsatellite and UEP haplotypic data.

A database of raw data containing all the results was constructed (see appendix), although for clarity the data has been summarised (see table 10.1). To make analyses easier, the microsatellite haplotypes were coded alphabetically one letter per locus, corresponding to the number of repeats. Thus the six locus haplotype with repeat numbers: 15, 12, 24, 10, 13, 14 of DYS19, DYS388, DYS390, DYS391, DYS392, DYS393 respectively, had the microsatellite code; KHTFIJ. In this system A=5 repeats, B=6, C=7 and so on to Z=30. The UEP coding system followed that of Jobling and Tyler-Smith (1995) and Hurles et al. (1998). The allelic states of each haplogroup were derived based on the collective works of Jobling and Tyler-Smith (1995) and Hurles et al. (1998). The groups were in the order of SRY10,831 (equivalent to SRY1532 in Hurles et al. 1998), M13, YAP, SRY4064, SY81, M9, M20, TAT, 92R7, M17, and SRY+465: haplogroup 1 chromosomes had the compound haplotype **GG**NGA**G**AT**T**G^{+}C, haplogroup 2 had **GG**NGACATCG^{+}C, haplogroup 3 had **AG**NGA**G**AT**TG**$^{-}$C, haplogroup 21 had **GGPA**ACATCG^{+}C and haplogroup 26 had **GG**NGA**G**ATCG^{+}C. This nomenclature indicated either the ancestral or derived forms of the polymorphism. The derived forms have been indicated by **bold typeface**.

The UEP and microsatellite codes were included in the database for every sample (see appendix).

TABLE 10.1: SAMPLE DATA, CLASSIFIED BY HAPLOGROUP AND MICROSATELLITE HAPLOTYPE.

(Population group listed as; L = U.K. Leicestershire, NAM = non-admixed Maori, AM = admixed Maori, NAP = non-admixed Polynesian Islander, AP = admixed Polynesian Islander. Microsatellite and UEP codes as described in previous page.)

Haplogroup	Microsatellite Code	AM	AI	L	NAM	NAI	TOTAL
1	IHRFKI				1		1
	IHTFII			1			1
	JHSFJH				1		1
	JHSGII	1	3	3			7
	JHSGIJ	1					1
	JHSGJI			1			1
	JHSHII	1					1
	JHTFII	2		2			4
	JHTGII	4		2			6
	JHTGIJ			1			1
	JHTHII			1			1
	JKTGII	1					1
	KHTFIJ	1					1
2	JJRFGI			1			1
	JJSFGI	2					2
	JJSFGJ			1			1
	JKSFGI				1		1
	JLRFGI	1					1
	KIRFGJ				1		1
	KJSFGI	1		1			2
	KJTFGI			1			1
	KKPFGJ					1	1
	KKPFHJ				1		1
	KKPGHJ	1					1
	KLRFGI				1		1
	LKOFHJ					1	1
	LKPFHJ				17	7	24
	LKPGHJ				1		1
	LKPGII				1		1
	LKQFHJ				1		1
	MITGHI	1					1
21	IHTFGI			3			3
	JHTFGI			1			1

Continued over page..

*From previous page..*Table 10.1 continued:

Haplogroup	Microsatellite Code	AM	AI	L	NAM	NAI	TOTAL
26	IHTGGI	1					1
	JFTFJH	1					1
	JHSFII	1		1			2
	JHTGJI			1			1
	JHUGJI				1		1
	KITFII	1	2				3
	KJTFII					1	1
	LJTEKI					1	1
	LJTELI					2	2
3	KHUGGI			1			1
	TOTAL	21	5	22	27	13	88

10.c. Allele frequency distributions

Allele frequency distributions among U.K Leicestershire, New Zealand Maori and Polynesian Islander populations at six polymorphic microsatellite and eleven biallelic marker loci.

In general, the allele frequencies of the admixed Polynesian populations and the U.K. Leicestershire population were similar. All loci had specific alleles which were observed among, at least, 50% of the chromosomes within any given population.

DYS19

Five different alleles among the populations were observed at this locus (see figure 10.1). Alleles with 13 and 17 repeat motifs were observed at low frequencies (<20% of the population), with over 80% of the total data described between alleles 14, 15 and 16. Allele 14 was observed to be most frequent in the U.K. Leicestershire, admixed Islander and admixed Maori populations with respective frequencies of 70%, 60% and 70%. Allele 16 was observed in 71% of non-admixed Islander and non-admixed Maori populations. The 15 repeat allele was observed in all the populations at frequencies below 30%.

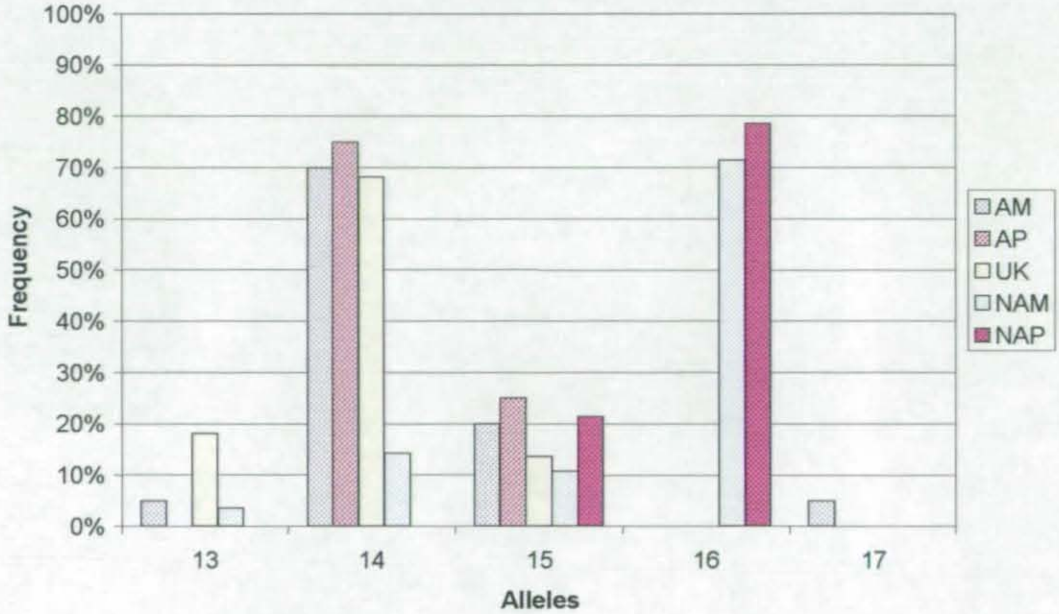


FIGURE 10.1: ALLELE FREQUENCY DATA AT THE DYS19 LOCUS. FREQUENCIES GIVEN AS PERCENTAGES OF THE TOTAL POPULATION DATA.

DYS392

Six different sized alleles were observed at this locus (see figure 10.2). Alleles 14 and 15 described less than 20% of the sample population data, and the 16 repeat allele was only isolated among the non-admixed Islander chromosomes. The U.K. Leicestershire and admixed Islander populations' modal allele had 13 motif repeats and was observed within 52% and 80% of the respective populations. The non-admixed Islanders, and non-admixed Maori modal allele had 12 motif repeats, observed within 71% and 50% of the respective populations. The 11 repeat allele was present in all the populations (with the exception of the admixed Islanders), below a frequency of 35%.

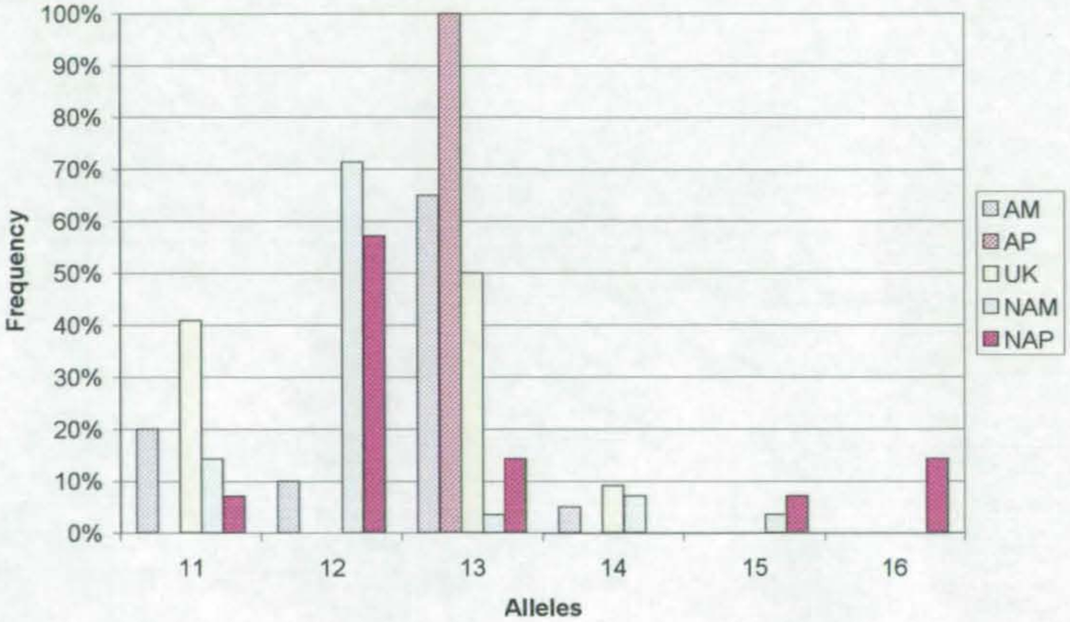


FIGURE 10.2: ALLELE FREQUENCY DATA AT THE DYS392 LOCUS. FREQUENCIES GIVEN AS PERCENTAGES OF THE TOTAL POPULATION DATA.

DYS391

Four different sized alleles were isolated among the five sample populations (see figure 10.3).

The 12 repeat allele was only present among the U.K. Leicestershire and admixed Maori chromosomes with frequencies below 5%. The 10 repeat allele was the modal allele in the U.K. Leicestershire, non-admixed Islander and non-admixed Maori chromosomes with respective frequencies of 51%, 89% and 78%. The 11 repeat allele was most frequent in the admixed Islander (60%) and admixed Maori (50%) populations. The 9 repeat allele was only observed among 4% of the U.K. Leicestershire chromosomes and 22% of the non-admixed Islander chromosomes.

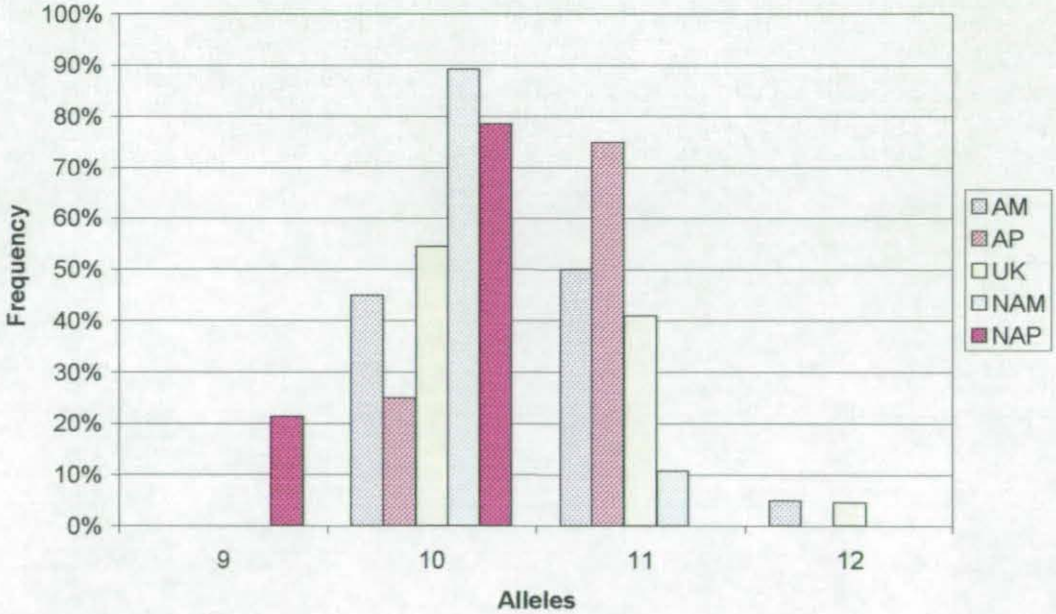


FIGURE 10.3: ALLELE FREQUENCY DATA AT THE DYS391 LOCUS. FREQUENCIES GIVEN AS PERCENTAGES OF THE TOTAL POPULATION DATA.

DYS388

Six different sized alleles were observed among the five defined populations (see figure 10.4). A clear difference in modal allele frequencies existed between the U.K Leicestershire, admixed Polynesian populations and the non-admixed Maori and non-admixed Islander populations. The former populations were observed to have the 12 repeat allele at a frequency of between 60-80%. In comparison, the 15 repeat allele was most common in the non-admixed Maori and non-admixed Islanders found within 57-78% of the total population. Alleles with 13 and 14 repeats were observed between populations, below a frequency of 30%. Interestingly, alleles 10 and 16 were only isolated at low frequencies (below10%) within the admixed Maori and non-admixed Maori chromosomes.

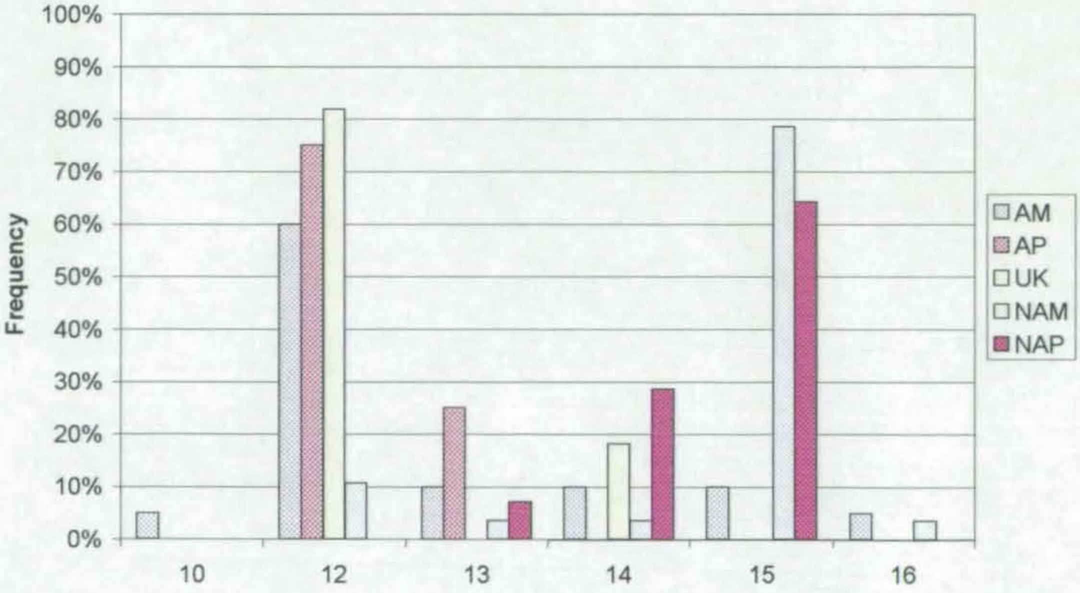


FIGURE 10.4: ALLELE FREQUENCY DATA AT THE DYS388 LOCUS. FREQUENCIES GIVEN AS PERCENTAGES OF THE TOTAL POPULATION DATA.

DYS390

This locus exhibited the most variation of modal allele frequencies between populations (see figure 10.5).

Seven different alleles at varying frequencies between populations were observed at this locus. The 20 repeat allele was most frequent in the non-admixed Maori and non-admixed Islander populations with respective frequencies of 72% and 50%. Interestingly, this allele was not observed in the U.K. Leicestershire sample population. The 23 repeat allele was most common among the admixed Islander chromosomes, with an observed frequency of 60%. The 24 repeat allele was the modal allele in the admixed Maori and U.K. Leicestershire populations with respective frequencies of 60% and 52%. Alleles 19, 21, 22 and 25 were all observed below 11% across all the populations.

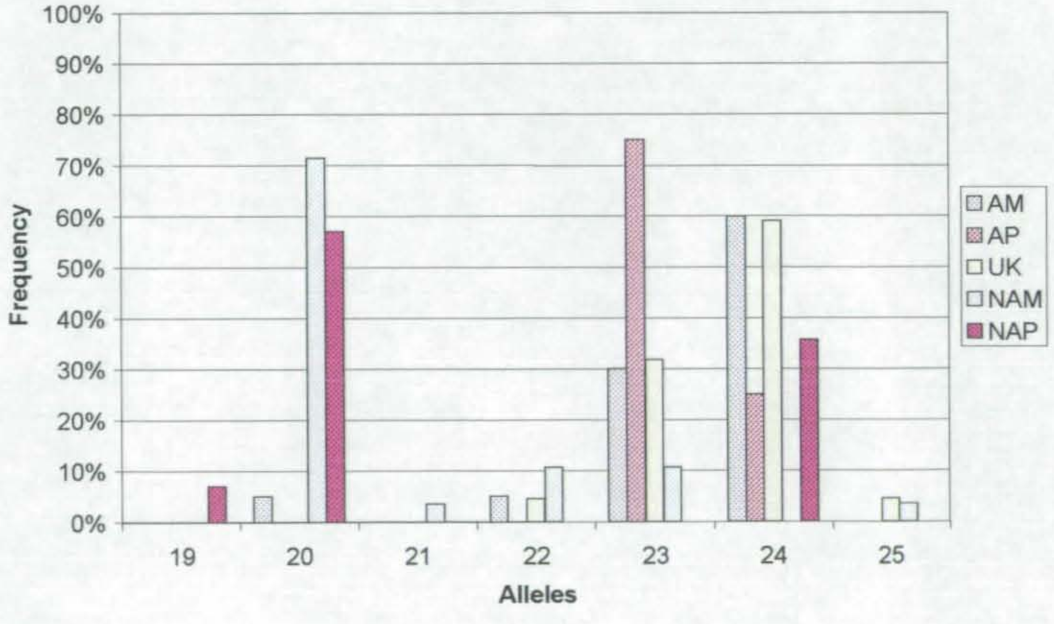


FIGURE 10.5: ALLELE FREQUENCY DATA AT THE DYS390 LOCUS. FREQUENCIES GIVEN AS PERCENTAGES OF THE TOTAL POPULATION DATA.

DYS393

Over 90% of the population data was described by two alleles. A clear distinction was observed between the modal frequencies of the admixed and the U.K. Leicestershire populations and the non-admixed populations (see figure 10.6). The 13 repeat allele was most frequent in the admixed Maori (80%), admixed Islander (80%) and U.K. Leicestershire (88%) populations and the non-admixed Maori and Islander populations with frequencies below 45%. Conversely, the 14 repeat allele was most frequent in the non-admixed Maori and Islander populations with frequencies of 75% and 58% and the admixed and U.K. Leicestershire populations with frequencies below 20%. The 12 repeat allele was only observed in the non-admixed Maori and admixed Maori populations with frequencies below 10% of the sample data.

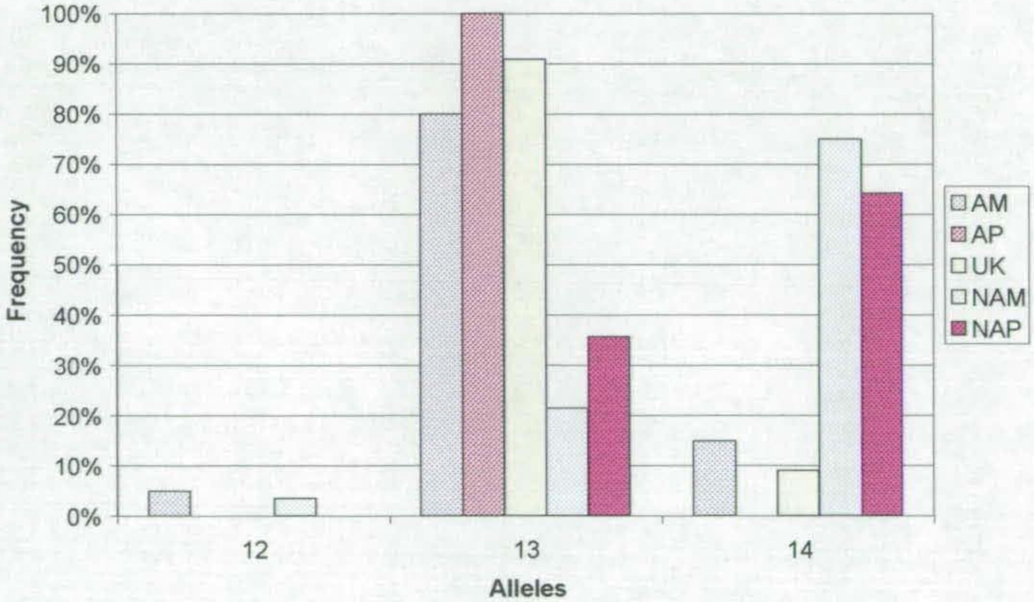


FIGURE 10.6: ALLELE FREQUENCY DATA AT THE DYS393 LOCUS. FREQUENCIES GIVEN AS PERCENTAGES OF THE TOTAL POPULATION DATA.

10.d. Unique Event Polymorphisms (UEPs)

In total eleven UEPs (or diallelic markers / single nucleotide polymorphisms) were included in this study allowing clear differences or similarities between populations and groups within the populations.

One insertion deletion and 10 base substitutions distinguished the Y-chromosome UEP groups.

Five different UEP haplogroups were observed across the complete database. The allelic states of each haplogroup followed the collective works of Jobling and Tyler-Smith (1995) and Hurles et al. (1998), and have been described at the beginning of this chapter.

The U.K. Leicestershire population contained all five UEP groups with varying frequencies. The modal UEP group for each population differed, with 'haplogroup1' most frequent in the admixed Maori (55%), admixed Islander (75%) and U.K. Leicestershire Caucasian (50%) populations and 'haplogroup 2' most frequent in the non-admixed Maori (89.3%) and non-admixed Islander populations (64.3%) (see table 10.2 for haplogroup frequencies).

TABLE 10.2: HAPLOGROUP FREQUENCIES WITHIN DEFINED POPULATIONS. ABBREVIATED POPULATION NAMES CORRESPOND TO AM = ADMIXED MAORI, AI = ADMIXED POLYNESIAN ISLANDER, U.K. = U.K. LEICESTERSHIRE, NAM = NON-ADMIXED MAORI, NAI = NON-ADMIXED POLYNESIAN ISLANDER.

	Population				
	Percentage frequency				
Haplogroup	AM	AI	U.K.	NAM	NAI
hg1	55.0	75.0	50.0	7.1	0.0
hg2	25.0	0.0	18.2	89.3	64.3
hg21	0.0	0.0	18.2	0.0	0.0
hg26	20.0	25.0	9.1	3.6	35.7
hg3	0.0	0.0	4.5	0.0	0.0
Grand Total	100%	100%	100%	100%	100%
Total Population Number	20	4	22	28	14

10.e. Associations between Unique Event Polymorphisms (UEPs) and Microsatellites

Distinct trends were observed throughout the whole data set with respect to UEP markers and specific microsatellites. The M9 UEP marker was characterised by a base substitution from the ancestral 'C' to the derived 'G' form. The C allele of the M9 locus was more closely associated to the smaller repeat numbers of the DYS390 microsatellite and conversely the G allele to the larger allele repeats. In particular, alleles between 19-21 repeats at the DYS390 locus, were only observed in samples with the 'C' or ancestral type at the M9 locus (see table 10.3). Similarly, the C allele at the M9 locus was observed to be associated with higher repeat numbers at the DYS19 locus and conversely the G allele was more frequent with the lower repeat numbers (see table 10.4). Similar trends were observed between the M9 marker and the loci; DYS388 and DYS391 (see tables 10.3-10.5) also the 92R7 UEP and microsatellite locus DYS393 (see table 10.5).

TABLE 10.3: ASSOCIATIONS BETWEEN THE M9 MARKER AND THE STR LOCI: DYS390 AND DYS393. INTEGERS REFER TO THE NUMBER OF CHROMOSOMES WITH THE TYPE.

	M9			M9	
DYS390 Alleles	C	G	DYS393 Alleles	C	G
19	1	0	12	0	2
20	28	0	13	16	36
21	1	0	14	31	3
22	4	1			
23	7	14			
24	6	24			
25	0	2			
Grand Total	47	41	Grand Total	47	41

Table 10.4: Associations between the M9 marker and the STR loci: DYS392 and DYS388. Integers refer to the number of chromosomes with the type.

	M9			M9	
DYS392 Alleles	C	G	DYS388 Alleles	C	G
11	15	2	10	0	1
12	29	0	12	4	32
13	3	30	13	3	3
14	0	5	14	7	4
15	0	2	15	31	1
16	0	2	16	2	0
Grand Total	47	41	Grand Total	47	41

Table 10.5: Associations between the M9 marker and the STR locus: DYS391 and the 92R7 marker and the STR locus: DYS391. Integers refer to the number of chromosomes with the type.

	M9			92R7	
DYS391 Alleles	C	G	DYS393 Alleles	C	T
9	0	3	12	1	
10	42	15	13	29	24
11	5	21	14	31	3
12	0	2			
Grand Total	47	41	Grand Total	61	27

10.f. Similarities and differences between the populations

Examination of the database (see table 10.1) uncovered numerous male lineage similarities and differences both within and between the populations.

Two unbiased gene diversity estimates were calculated for each population (table 10.6). The first estimate was only based on defined UEP haplogroup frequencies and the second included both UEP and defined microsatellite haplotype frequencies. These calculations were performed across all loci per population.

Unbiased gene diversity estimates based on UEP frequencies ranged from 0.232 (s.d ± 0.089) in the non-admixed Maori population to 0.721 (s.d ± 0.065) in the U.K. Leicestershire population. The aforementioned UEP estimates were, as anticipated, lower than the combined microsatellite and UEP data. However, both unbiased gene diversity estimates of the admixed Islander chromosomes were the same (0.600 ± 0.175). This was a direct reflection of the small number of chromosomes isolated (n=5), where only two haplogroups were observed with a single microsatellite type in either group.

Combining microsatellite haplotype and UEP haplogroup data, a similar trend of gene diversities was observed, with the greatest gene diversity in the U.K. Caucasian population (0.965 ± 0.024) and the least within the admixed Islander population (0.600 ± 0.175).

TABLE 10.6: GENE DIVERSITIES AND THEIR SAMPLING VARIANCES ACROSS ALL LOCI PER POPULATION.

Population	UEP		STR+UEP	
	Unbiased Gene Diversity	Sampling variance (+/-)	Unbiased Gene Diversity	Sampling variance (+/-)
U.K. Leicestershire (22)	0.721	0.065	0.965	0.024
Admixed Maori (21)	0.651	0.079	0.961	0.030
Admixed Islanders (5)	0.600	0.175	0.600	0.175
Non-admixed Maori (27)	0.232	0.089	0.613	0.111
Non-admixed Islanders (13)	0.536	0.082	0.717	0.128

10.g. Microsatellite haplotypes with respect to the different haplogroups

Haplogroup 1

Twelve different microsatellite haplotypes were observed within UEP haplogroup 1. Two non-admixed Maori chromosomes, 55% of admixed Maori chromosomes, 50% U.K. Leicestershire Caucasian samples and 7% of the 'native' Maori chromosomes, were isolated within this group. Three microsatellite haplotypes with codes: JHSGII, JHTFII and JHTGII were shared between the admixed Maori and U.K. Leicestershire populations (see table 10.1).

Haplogroup 2

A total of 18 different microsatellite haplotypes were observed across the data set. In comparison to haplogroup 1, relatively few U.K. Leicestershire (18%) and admixed Maori (25%) chromosomes were isolated. However, 89% of the non-admixed Maori and 64% of the non-admixed Polynesian Islander chromosomes were observed within this haplogroup. A dominant or modal microsatellite with code: LKPFHJ was isolated in 17 (68%) of 25 Maori chromosomes and 7 (78%) of 9 Polynesian Islander chromosomes and was not present in the U.K. Leicestershire sample set. The modal haplotype was characterised by a small DYS390 allele (20 repeats) and large DYS19 allele (16 repeats), in comparison to the average allele sizes of the DYS390 (24 repeats) and DYS19 (14/15 repeats) loci isolated in the U.K. Leicestershire chromosomes. This combination was not observed in any sample other than the 'Polynesian' samples.

Another microsatellite haplotype (KJSFGI) was shared between an admixed Maori and U.K. Leicestershire sample, and another (JJSFGI) between an admixed Maori and non-admixed Maori sample.

Haplogroup 21

Only four of twenty-two U.K. Leicestershire chromosomes were observed in this haplogroup, of which three were of the microsatellite type: IHTFGI and the fourth differing by a single step mutation at the DYS19 locus. This haplogroup has been observed primarily within European and African populations (Karafet et al. 1999), although has also been isolated within Japanese chromosomes (Hammer 1994, Hammer and Horai 1995, Ruiz-Linares et al. 1996).

Haplogroup 26

A total of nine different microsatellite haplotypes with varying frequencies were observed across all populations. The microsatellite haplotype with code: JHSFII was shared between an admixed Maori and U.K. Leicestershire chromosome and the microsatellite haplotype, KITFII shared between one of five admixed Maori and two of six admixed Islander chromosomes.

Haplogroup 3

Only one (U.K. Leicestershire) haplogroup 3 chromosome (microsatellite code: KHUGGI) was isolated in this present study. The microsatellite haplotype had the largest DYS390 allele (25 repeats) isolated across the data set. Previously, chromosomes analogous to haplogroup 3 have been found in both Europe and Asia (Hammer et al. 1998).

10.h. Network Analyses using 'NETWORK V2.0b'

Network analyses were carried out using the computer program Network V2.0b obtained from Dr Arne Roehl (E-mail: arne.roehl@de.pwcglobal.com).

Microsatellite networks of the UEP haplogroups and UEP haplogroups per population were constructed to determine the relationship between groups and populations.

Maori and Polynesian Islander UEP Networks

The New Zealand Maori and Polynesian Islanders both had the same UEP groups although the frequency of each group varied between populations (see figure 10.7). The haplogroups (shown in black boxes in figure 10.7) differed by single base substitutions at specific loci. For the purposes of showing the number of different haplogroups in the Maori and Islander populations the admixed samples were included in these analyses. Only five admixed Islander chromosomes were included in the present study, as this was all that was available at the time of DNA collection.

Haplogroup 2 was most common among the New Zealand Maori and Polynesian Islander populations with respective frequencies of 64.6% and 47.4%. The network diagram of these groups expressed a proportionately larger number of haplogroup 1 individuals within the Maori population in comparison to the Islanders. This was because the circle sizes corresponded to the actual number of chromosomes and not the relative percentage frequencies within the population. Thus within haplogroup 1 the bias of larger population numbers within the Maori population appeared to be more frequent, than the Islander population. However, the percentage frequency of haplogroup 1 relative to the population sample number, was 25% for the Maori and 21% for the Islanders, hence reducing the observable 'diagramatic' network difference.

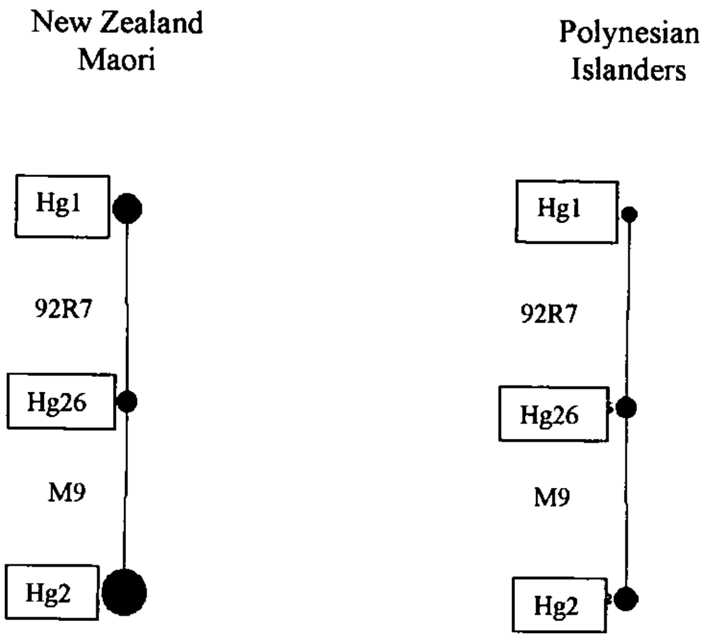


FIGURE 10.7: UEP HAPLOGROUP DISTRIBUTIONS: UNROOTED TREE FOLLOWING THE WORK OF JOBLING AND TYLER-SMITH (1995) AND HURLES ET AL. (1998). THE SIZES OF THE CIRCLES ARE PROPORTIONAL TO THE NUMBERS OF CHROMOSOMES. (N.B THE CIRCLE SIZE OF HAPLOGROUP 1 IN THE POLYNESIAN ISLANDER DIAGRAM IS EQUIVALENT TO 4 CHROMOSOMES).

U.K. Leicestershire UEP Network analysis

Five different haplogroups were observed among the U.K. Leicestershire chromosomes. The relationship between each haplogroup was clearly defined and described simply using a network diagram (see figure 10.8). The frequencies of each haplogroup were as listed in table 10.2. In addition to the three haplogroups (haplogroups 1, 2 and 26) observed within the Islander and Maori populations' two further haplogroups (haplogroup 3and 21) were isolated. Haplogroup 3 extended from haplogroup 1 with point mutations at loci SRY 10,831 and M17. Similarly, haplogroup 21 extended from haplogroup 2 with a deletion event at the YAP locus forming the YAP positive haplotype and a point mutation at the SRY 4064 locus from the ancestral 'G' to the derived 'A'. These mutations were closely linked, thus the derived form of the SRY 4064 marker was consistently observed with the YAP positive haplotype.

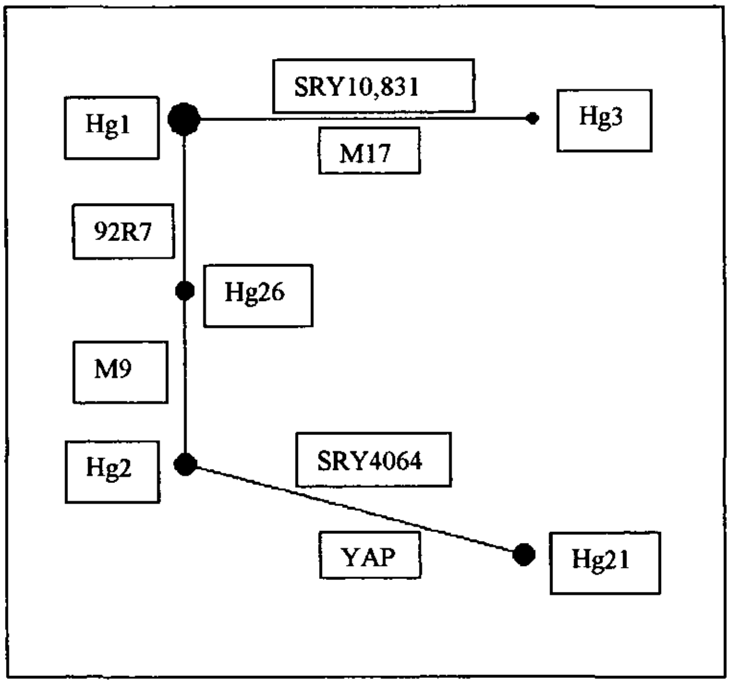


FIGURE 10.8: U.K. LEICESTERSHIRE HAPLOGROUP: UNROOTED TREES BASED ON THE WORK OF JOBLING AND TYLER-SMITH (1995), FOLLOWING THE GROUP STRUCTURE OF HURLES ET AL. (1998). THE HAPLOGROUPS ARE WRITTEN IN BOXES AND THE MUTATED LOCUS LINKING EACH GROUP HAS BEEN INCLUDED. (N.B HAPLOGROUP 3 REFERS TO 1 CHROMOSOME)

10.i. Microsatellite Networks

Microsatellite networks were constructed per population per haplogroup, toidentify relationships between specific haplotypes within the populations.

Previously, minimum spanning networks, such as those generated using the Network program (Schneider et al. 1999), provided 'pictorial' indication of whether the population was expanding, or if selection favored a particular haplotype (Jobling and Tyler-Smith 1995).

Microsatellite Network analyses of Haplogroup 1 U.K. Leicestershire Chromosomes

This network positioned the microsatellite haplotype 'JHTGII' centrally, with haplotypes varying by one or two repeats 'radiating' from it. The microsatellite haplotypes JHTFII and JHSGII had the same number of chromosomes as the 'central' haplotype (n=2) and from these, two additional haplotypes were observed, that differed by a single mutation (see figure 10.9).

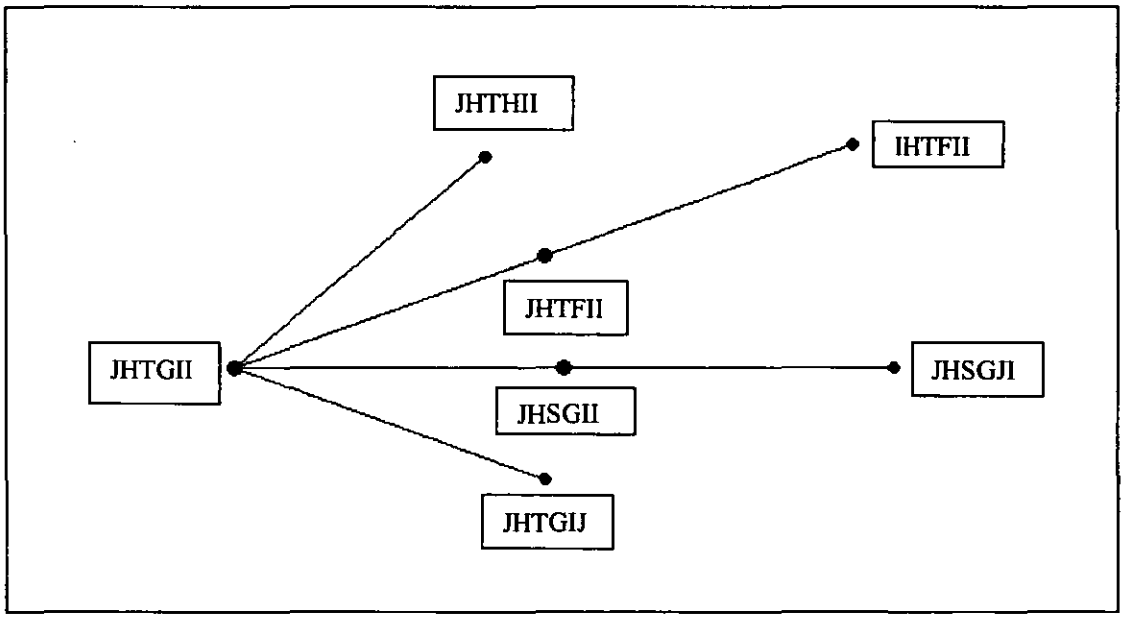


FIGURE 10.9: MICROSATELLITE NETWORK OF HAPLOGROUP 1 U.K. LEICESTERSHIRE CHROMOSOMES: UNROOTED TREE BASED ON THE WORK OF JOBLING AND TYLER-SMITH (1995), FOLLOWING THE GROUP STRUCTURE OF HURLES ET AL. (1998). THE MICROSATELLITE CODES CORRESPOND TO THOSE GIVEN IN TABLE 10.1.

Microsatellite Network analyses of Haplogroup 1 New Zealand Maori

This network formed a more complex microsatellite array than the haplogroup 1 U.K. Leicestershire microsatellite network (see figure 10.9). However, similarly to the U.K. Leicestershire network, the modal microsatellite haplotype was 'JHTGII', and the haplotypes with codes JHTFII and JHSGII stemming from it. Haplotypes differing by more than one or two repeats from the modal haplotype were observed, and the most distant haplotype (IHRFKI) was six mutational events apart from the modal haplotype.

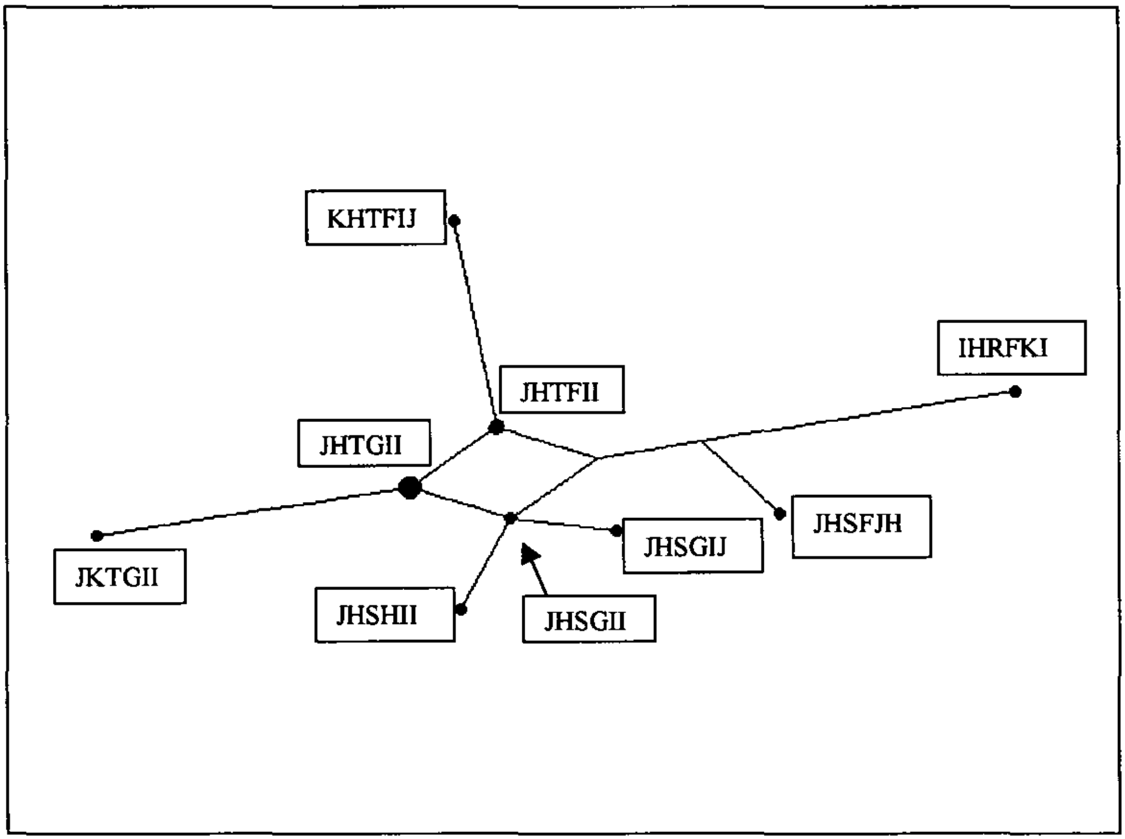


FIGURE 10.10: MICROSATELLITE NETWORK OF HG1 NEW ZEALAND MAORI CHROMOSOMES: UNROOTED TREE BASED ON THE WORK OF JOBLING AND TYLER-SMITH (1995), FOLLOWING THE GROUP STRUCTURE OF HURLES ET AL. (1998). THE MICROSATELLITE CODES CORRESPOND TO THOSE GIVEN IN TABLE 10.1. CIRCLE SIZES CORRESPOND TO CHROMOSOME NUMBERS, WHEREBY JKTGII = 1 CHROMOSOME.

Microsatellite Network of Haplogroup 2: U.K. Leicestershire Chromosomes

This formed an interesting network, as no central or modal haplotype was observed (see figure 10.11).

The haplotypes clearly differed by one repeat at a specific locus, to the adjacent haplotype, thus rather than constructing a radiative dispersion as seen in figure 10.11, one could construct sequential 'linear' network in the order;

JJRFGI – JJSFGI – KJSFGI – KJTFGI.

However, this network was constructed from 4 chromosomes therefore may exhibit greater complexity with greater population numbers.

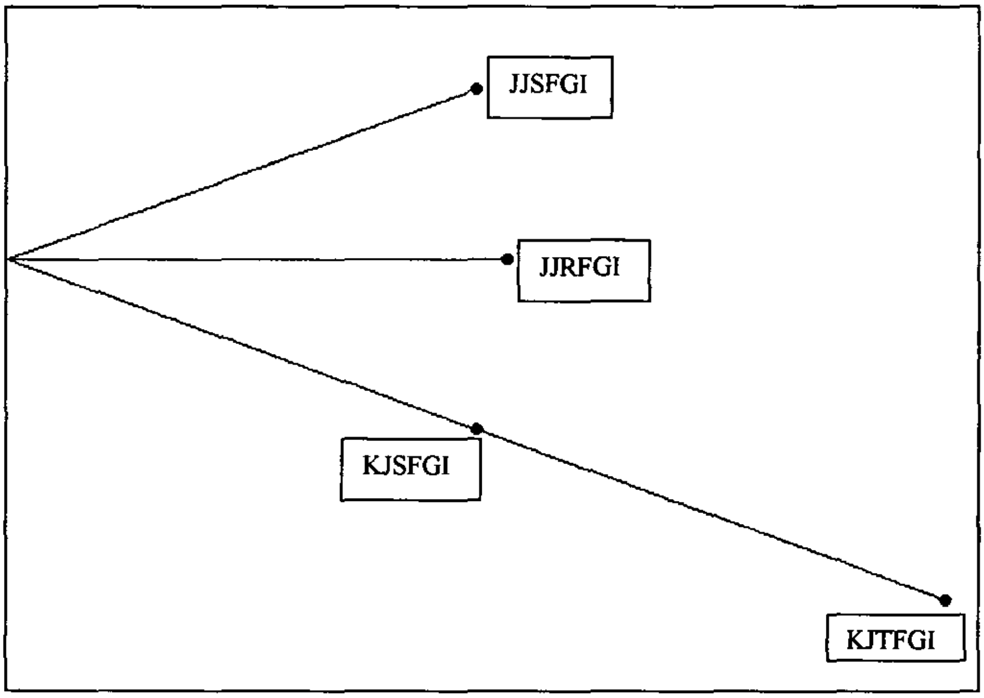


FIGURE 10.11: MICROSATELLITE NETWORK OF HAPLOGROUP 2 U.K. LEICESTERSHIRE CHROMOSOMES: UNROOTED TREE BASED ON THE WORK OF JOBLING AND TYLER-SMITH (1995), FOLLOWING THE GROUP STRUCTURE OF HURLES ET AL. (1998). THE MICROSATELLITE CODES CORRESPOND TO THOSE GIVEN IN TABLE 10.1. CIRCLE SIZES CORRESPOND TO CHROMOSOME NUMBERS, WHEREBY KJTFGI = 1 CHROMOSOME.

Microsatellite Network for Haplogroup 2: Maori Chromosomes

This was the most complex microsatellite network of all. The LKPFHJ microsatellite haplotype (see appendix for detailed allele information) was clearly the modal haplotype within the network. Radiating from the 'modal' haplotype, were three haplotypes differing by a single mutation and two differing by two mutations. These haplotypes formed a cluster to which another joined, although this was distinctly separated from the modal haplotype (LKPFHJ) and its derivatives.

The second cluster attached to the first with the closest haplotype distanced by five single-step mutations. The second grouping contained predominantly admixed samples of known European origin (see appendix for admixture). Further comparison of this cluster to other European chromosomes indicated all haplotypes within the second cluster were of European origin (Thomas personal communication). Similarly, all of the first cluster, including the modal haplotype, were not observed among the European population used for comparison (Thomas personal communication).

Interestingly, within the present study, haplotype JJSFGI was observed in two samples, the first was of known northern european admixture and the other purported to be 'pure' Maori. An anomalous haplotype was observed in this haplogroup recorded by just one chromosome. This haplotype (MITGHI) had the highest recorded number of repeats at the DYS19 locus and was observed in a sample of known admixture.

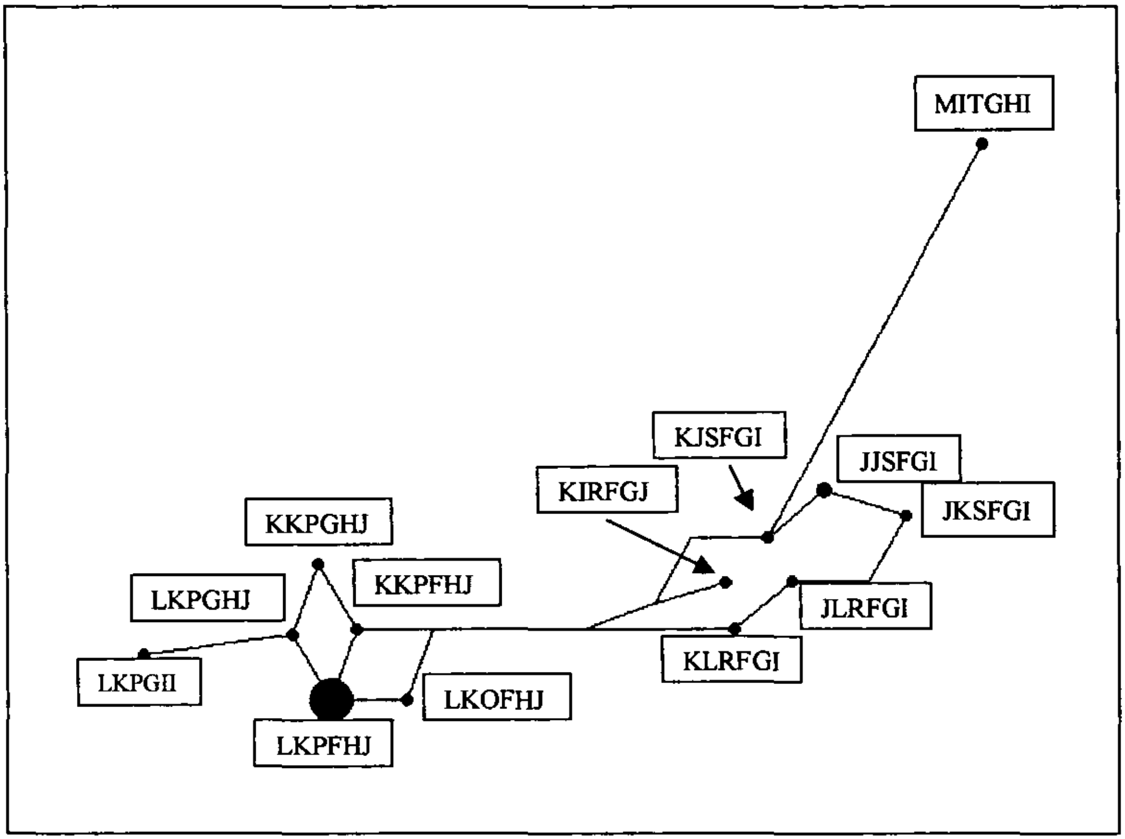


FIGURE 10.2: MICROSATELLITE NETWORK OF HAPLOGROUP 2 MAORI CHROMOSOMES: UNROOTED TREE BASED ON THE WORK OF JOBLING AND TYLER-SMITH (1995), FOLLOWING THE GROUP STRUCTURE OF HURLES ET AL. (1998). THE MICROSATELLITE CODES CORRESPOND TO THOSE GIVEN IN TABLE 10.1. CIRCLE SIZES CORRESPOND TO CHROMOSOME NUMBERS, WHEREBY MITGHI = 1 CHROMOSOME.

Microsatellite Network of Haplogroup 2: Polynesian Islander chromosomes

Similarly to the Maori network, the modal microsatellite haplotype was LKPFHJ (see figure 10.13). Two chromosome haplotypes were separated from the modal haplotype by a single repeat difference. However, this network may become more involved if larger chromosome numbers were analysed.

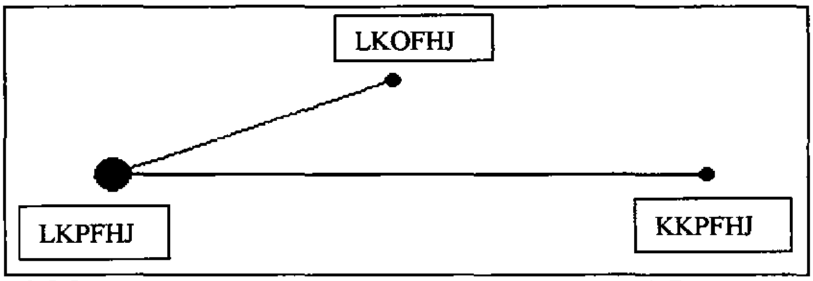


FIGURE 10.13 MICROSATELLITE NETWORK OF HAPLOGROUP 2 POLYNESIAN ISLANDER CHROMOSOMES: UNROOTED TREE BASED ON THE WORK OF JOBLING AND TYLER-SMITH (1995), FOLLOWING THE GROUP STRUCTURE OF HURLES ET AL. (1998). THE MICROSATELLITE CODES CORRESPOND TO THOSE GIVEN IN TABLE 10.1. CIRCLE SIZES CORRESPOND TO CHROMOSOME NUMBERS, WHEREBY KKPFHJ = 1 CHROMOSOME.

Microsatellite Network of Haplogroup 26: Maori chromosomes

This star shaped network had no central or modal haplotype. All the chromosomes except one with microsatellite haplotype code JHUGJI, originated from admixed DNA samples. The JHSFII haplotype was also isolated within the U.K. Leicestershire chromosomes. The chromosomes with microsatellite haplotype codes IHTGGI and JFTFJH had purported European/Chinese admixture. In particular, the sample with the JFTFJH code had the smallest recorded DYS388 repeat number of the whole database.

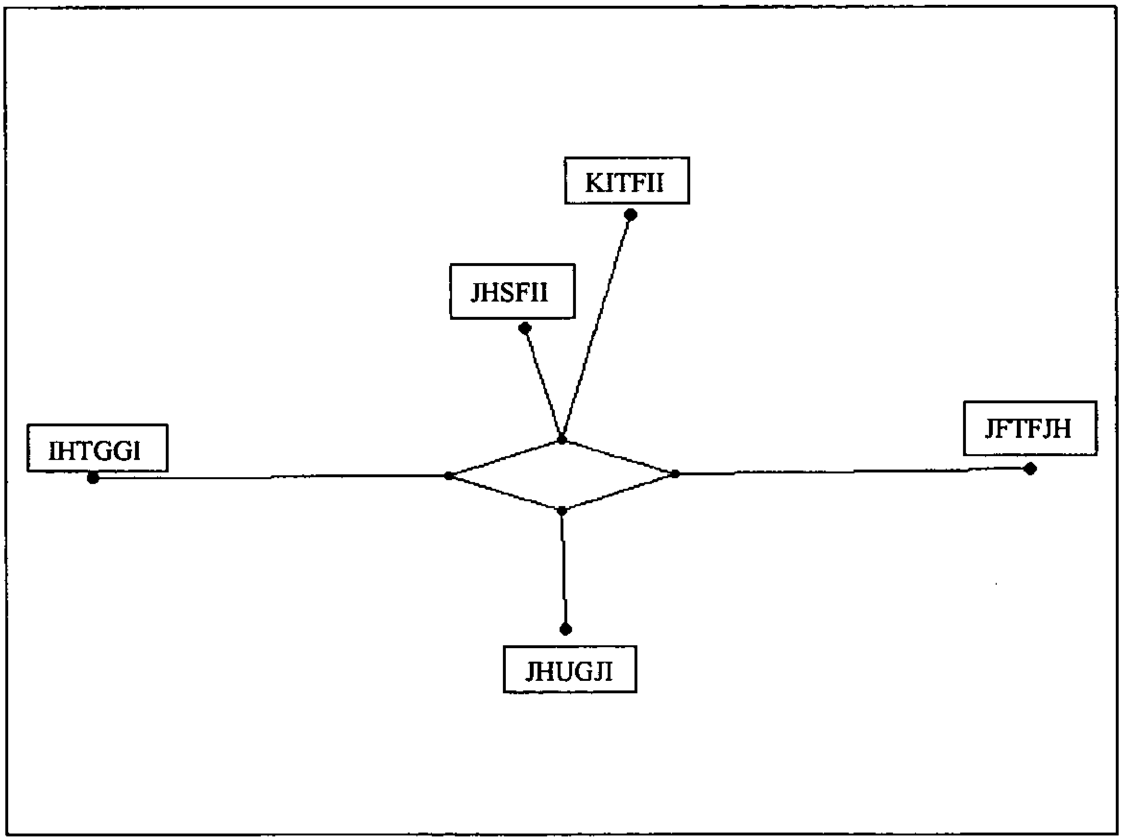


FIGURE 10.14: MICROSATELLITE NETWORK OF HAPLOGROUUP 26 MAORI CHROMOSOMES: UNROOTED TREE BASED ON THE WORK OF JOBLING AND TYLER-SMITH (1995), FOLLOWING THE GROUP STRUCTURE OF HURLES ET AL. (1998). THE MICROSATELLITE CODES CORRESPOND TO THOSE GIVEN IN TABLE 10.1. CIRCLE SIZES CORRESPOND TO CHROMOSOME NUMBERS, WHEREBY KITFII = 1 CHROMOSOME. THE UNLABELLED POINTS ON THIS NETWORK CORRESPOND TO HAPLOTYPES THAT SHOULD JOIN ONE TO ANOTHER, ALTHOUGH WERE NOT OBSERVED IN THIS STUDY.

Microsatellite Network of Haplogroup 26 U.K. Leicestershire Caucasians

Only two chromosomes were isolated forming haplogroup 26. These samples differed by a single mutation. In general, these microsatellite haplotypes had smaller repeat numbers at loci; DYS19, DYS388 and DYS390, than the equivalent haplogroup among the Polynesian Islander chromosomes.

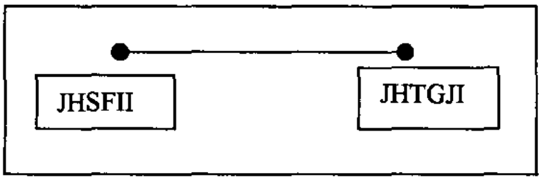


FIGURE 10.15: MICROSATELLITE NETWORK OF HAPLOGROUP 26 U.K. LEICESTERSHIRE CHROMOSOMES: THE MICROSATELLITE CODES CORRESPOND TO THOSE GIVEN IN TABLE 10.1. CIRCLE SIZES CORRESPOND TO CHROMOSOME NUMBERS, WHEREBY JHSFII = 1 CHROMOSOME.

Microsatellite Network for Haplogroup 26 Polynesian Islander Chromosomes

A total of 6 Islander chromosomes were isolated, originating from individuals of no known admixed ancestry with no purported admixture were observed within this group (see figure 10.16).

The microsatellite haplotypes of the Tokelau chromosomes were distinguished by a single repeat difference at the DYS392 locus (see appendix for details) and clustered separately to the Tongan and Tahitian chromosomes by a minimum of four repeat motifs across three or more loci.

Similarly to the U.K. Leicestershire network, no modal microsatellite haplotype was observed. However, one should note that haplogroup 26 was isolated in only 6 Polynesian Islanders and 3 U.K. Leicestershire, thus such small sample sizes may possibly under represent the microsatellite diversity within this haplogroup.

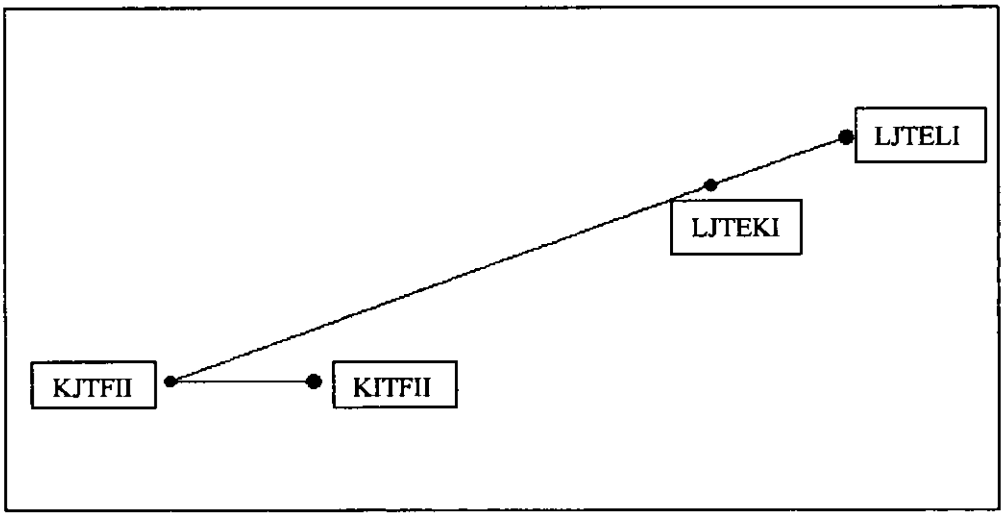


FIGURE 10.16: MICROSATELLITE NETWORK OF HAPLOGROUP 26 POLYNESIAN ISLANDER CHROMOSOMES: UNROOTED TREE BASED ON THE WORK OF JOBLING AND TYLER-SMITH (1995), FOLLOWING THE GROUP STRUCTURE OF HURLES ET AL. (1998). THE MICROSATELLITE CODES CORRESPOND TO THOSE GIVEN IN TABLE 10.1. CIRCLE SIZES CORRESPOND TO CHROMOSOME NUMBERS, WHEREBY KJTFII = 1 CHROMOSOME.

Microsatellite Network Haplogroup 21: U.K. Leicestershire Chromosomes

This haplogroup was only isolated in the U.K. Leicestershire chromosomes of this study, included just 4 chromosomes (see figure 10.17). Of these, two different haplotypes were isolated with two chromosomes of either haplotype. Similarly, to other haplogroups with few isolated chromosomes, further analyses may increase the numbers of different haplotypes.

JHTFGI
IHTFGI

FIGURE 10.16 MICROSATELLITE NETWORK OF HAPLOGROUP 21 U.K. LEICESTERSHIRE CHROMOSOMES: UNROOTED TREE BASED ON THE WORK OF JOBLING AND TYLER-SMITH (1995), FOLLOWING THE GROUP STRUCTURE OF HURLES ET AL. (1998). THE MICROSATELLITE CODES CORRESPOND TO THOSE GIVEN IN TABLE 10.1. CIRCLE SIZES CORRESPOND TO CHROMOSOME NUMBERS, WHEREBY JHTFGI = 1 CHROMOSOME.

10.j. Variance Analyses

Analyses of Molecular Variance (AMOVA)

The most outstanding observation was that in every instance there was far greater variation within than among populations. The least variation among populations was observed within a clustered group formed by U.K. Leicestershire and admixed Maori samples (see table 10.7 for details) indicating the genetic similarity between these populations. This was anticipated, as the admixture within the Maori samples was mainly, of European origin (refer to the raw data in the appendix).

A similarly low variation among populations was observed in a group formed from 'pure' Maori and Polynesian Islander Y chromosomes, thus expressing the close genetic relationship between these populations.

Interestingly, when grouping the non-admixed and admixed Maori data, greater variation among populations was observed (42%) than grouping all the sample data (31%). However, the inclusion of all the data may have 'absorbed' the differences between the admixed and non-admixed Maori chromosomes highlighted by their separate analysis.

Table 10.7: Combinations of populations have been grouped to test the variance within and between them. Calculations carried out using the AMOVA test in Arlequin (Schneider et al. 1997).

Source of Variation	STR + UEP 1 Group All Populations	STR + UEP 2 Groups 1 group= Caucasian Admixed populations 1 group= Pure Maori Pure Islander	STR + UEP 1 Group Pure Maori Pure Islander	STR + UEP 1 Group Caucasian Admixed Maori	STR + UEP 1 Group Pure Maori Admixed Maori
Among Populations	31.80	40.00	2.87	-0.21	42.1
Among Populations within groups		0.73			
Within Populations	68.2	58.86	97.13	100.21	57.9
FST Values	0.318	0.411	0.0287	-0.002	0.421

The Fst values were calculated as part of the AMOVA analyses (Schneider et al. 1997; see Table 10.7).

Following Hartls' (1981 Chapter 2 Pp79) qualitative guide to the interpretation of FST statistics; a range of 0 - 0.05 may be considered as an indication of little genetic differentiation, 0.05 – 0.15 an indication of moderate differentiation and above 0.25 as an indication of very great differentiation.

Following the aforementioned Fst guidelines very little differentiation was observed between the non-admixed Maori and non-admixed Islander chromosomes. Similarly, very little differentiation was observed between the U.K. Leicestershire and admixed Maori chromosomes. The greatest genetic differentiation (FST = 0.421) was observed between the non-admixed and admixed Maori chromosomes.

10.k. Genetic Structure Analyses

Coancestry coefficients were calculated, incorporating the natural logarithm of a pairwise difference between two populations as function of their variance and frequency (see table 10.8). The statistic ranged 0 – 1, where 0 suggested little or no difference and 1 complete genetic differentiation. As expected from the AMOVA statistics, a genetically close coancestry was observed between the U.K. Leicestershire and admixed Maori chromosomes. The most distant coancestry was between the admixed Maori and pure Islander chromosomes, which was marginally greater than the coancestry distance between the U.K. Leicestershire and pure Islander populations.

10.l. Migration

An identical pattern of results to the coancestry coefficient was observed with the matrix of genetic migration per generation (see upper matrix table 10.8). In agreement with the coancestry coefficient data, a very close genetic relationship was observed between the U.K. Leicestershire and admixed Maori population, by virtue of a purported 'infinite' migration per generation. Similarly to the coancestry coefficient matrix, marginally greater migration per generation was observed between the U.K. Leicestershire and non-admixed Islanders than the admixed Maori and non-admixed Islanders (see table 10.8 for details).

Table 10.8: The lower triangular matrix refers to the coancestry coefficients and the upper triangular matrix refers to the genetic migration per generation between two populations'.

Population	U.K.	NAM	NAI	AI	AM
U.K.	-	0.679	0.899	12.97	Inf
NAM	0.552	-	16.89	0.376	0.687
NAI	0.442	0.029	-	0.482	0.866
AI	0.038	0.845	0.711		Inf
AM	0.000	0.546	0.435	0.0000	-

10.m. Common Ancestry Date

The common ancestry date (t) in generations for the 'ancestral' Maori cluster of closely related Y-chromosomes, centred around the modal haplotype (LKPFHJ), was assessed by calculating the average squared difference (ASD) in allele size between all chromosomes included the cluster and the ancestral haplotype (assumed to be the same as the modal haplotype), averaged over all loci.

This had the expectation μt where μ is the microsatellite mutation rate (Goldstein et al. 1995, Slatkin 1995). A point estimate of μ=4/3155=0.00127 was used. This was based on data from 3 published studies (Heyer et al. 1997, Kayser et al. 1997, Bianchi et al. 1998), but restricted to the same microsatellite loci as those used here (giving 4 observed single-step mutations in 3155 separate meiotic events) (Thomas personal communication).

A narrow confidence interval (CI) on t was calculated according to the method of Thomas et al. (1998) using 100,000 replications. This interval assumed μ was known without error and reflected the sampling variance of mutations of a star genealogy only.

The Polynesian-specific cluster found in the Maori sample contained a modal haplotype surrounded by singletons connected by one, two and three mutational steps. All other haplogroup 2 chromosomes isolated among the Maori / Polynesian chromosomes were at least 5 mutation steps removed from this cluster and are likely to be of European origin (Thomas personal communication). This was suggestive of a recent common origin for the Polynesian-specific cluster of chromosomes, where the modal haplotype was representative of the ancestral form. Assuming this, a mutation rate of 0.00127 and a generation time of 25 (30) years, the coalescence time for this cluster was estimated at 1193 (1432) years. The fact that the observed non-ancestral chromosomes are singletons was suggestive of an uncorrelated genealogy and allows estimates of the 95% CI to be calculated (Thomas et al. 1998, Thomas personal communication). Ignoring uncertainty in the mutation rate the 95% CI fell between 448 and 2241 (538 and 2689) years before present.

Chapter 11

Discussion: Male lineages among the U.K. Caucasian, New Zealand Maori and Polynesian Islander Populations

The aims of the male lineage study were to present the first New Zealand Maori data, to evaluate the extent of European admixture and to compare the findings to previous research. These aims were all completely met, furthermore, the hypothesis that 'the New Zealand Maori males would be closely associated with the Polynesian Islander males and extensive European admixture would be present', was proven by the results of this study and comparisons to previous studies.

One of the oldest recorded male lineages describes 'Adams' descendents to Noah, in the Old Testament, Genesis, Chapter 5:

> '*when Adam had lived one hundred and thirty years, he became the father of a son in his likeness, according to his image, and named him Seth. Seth begat Enos, Enos begat Cainan, Cainan begat Mahalaleel, Mahalaleel begat Jared, Jared begat Enoch, Enoch begat Methuselah, Methuselah begat Lamech and Lamech begat Noah —he of the Arc and great flood*'

Although the Biblical lineage bared no reference to the transfer of genetic material, it acknowledged the inherent similarity between father and son.

Thousands of years later, geneticists have been tracing male lineages to elaborate upon the descendency of not only individuals (as described in Genesis of the Old Testament) but also populations and retracing ancestries of those inhabiting Earth today.

To ensure Y chromosome inheritance was devoid of recombination between the female sex chromosome, studies of male lineages have concentrated on the non-recombining portion of the Y chromosome (Karafet et al. 1999).

Archaeological (Ward 1972), anthropological (Terrell 1986) and genetic studies, including mtDNA (Hagelberg et al. 1999), VNTR (Hamilton et al. 1996) and Y chromosome studies among the Polynesian Islanders (Hurles et al. 1998), have all observed a recent colonisation of the Pacific. Significant European genetic admixture within Polynesia has also been observed in recent studies (Hurles et al. 1998, Hagelberg et al. 1999).

It was hypothesized that the New Zealand male Maoris would have close genetic affinities to the Polynesian Islanders and have a proportion of European admixture. This discussion compares and contrasts male lineages between the populations investigated in this study and where appropriate, previous research. There has been a bias towards discussing the New Zealand Maori lineages in more detail, as this is the first study incorporating Y chromosome analyses among the New Zealand Maori population.

11.a. Mutations: Single Nucleotide Polymorphisms' and their significance

The Y chromosome has been documented as having a very low mutation rate, with genetic variation characteristically hard to find (Seielstad et al. 1999). The lack of polymorphic diversity in comparison to autosomal polymorphic marker studies was described by three theories (Jobling and Tyler-Smith 1995);

i) Smaller Y chromosome: X chromosome population in the ratio 1:3

ii) Reduction of effective population size through fathers with many male offspring, and

iii) Advantageous mutations within the non-recombining portion of the chromosome may be accompanied by a hitchhiker low in variation (for a detailed synopsis on mutations see Y chromosome introduction).

The mutation rate of polymorphic marker loci has varied frequencies ranging 5.6 X 10^{-4} mutations per generation (Weber and Wong 1993) to 1.2 X10^{-3} mutations per generation (Heyer et al. 1997 and Bianchi et al. 1998). However, single nucleotide polymorphisms were purported to have the lowest mutation frequency, and have often been referred to as unique event polymorphisms (UEPs), since they occurred only once in history (Thomas et al. 1998). However, exceptions to the rule occur and the UEP marker SRY10,831 had reverted back from the derived to the ancestral state (Hammer et al. 1998, Santos et al. 1999 and Karafet et al. 1999). Although the reversion has not been observed in the present study.

This present study included 11 UEPs; YAP (Hammer 1994), sY81 (DYS271) (Seielstad et al. 1994), 92R7 (Mathias et al. 1994), SRY10,831 (Whitfield et al. 1995), SRY4064 (Whitfield et al. 1995), TAT (Zerjal et al. 1997), SRY465 (Thomas et al. 1998), M9, M13, M17 and M20 (Underhill et al. 1997). Combinations of these UEPs had been incorporated into previous studies including; analyses of Oceanic populations (Hurles et al. 1998, Hagelberg et al. 1999), the Lemba 'black jews' of southern Africa (Spurdle and Jenkins 1996), Finnish population origins (Kittles et al. 1996) and even World populations (Jobling et al. 1997). However, no study had previously investigated the male lineage in the New Zealand Maori population.

11.b. UEP Haplogroups: Within and between populations

This study observed 5 different UEP haplogroups within the UK Leicestershire population and 3 different haplogroups among the Polynesian populations (see Chapter 10). The haplogroups were coded following Jobling and Tyler-Smith (1995) and Hurles et al. (1998) nomenclature systems (see Chapter 10; results section), which provided the opportunity to compare haplogroups between the aforementioned studies to this thesis.

Statistical analyses were performed using the 'Arlequin' package (Schneider et al. 1997). Arlequin was used in previous studies and found to be appropriate for haploid / haplotype data (Bianchi et al. 1998, Kittles et al. 1998, Scozzari et al. 1999 and de Knijff et al. 1999).

UEP Haplogroup 1

This haplogroup referred to the derived forms of the loci, SRY10,831, M13, M9 and 92R7 and the ancestral forms of YAP, SRY4064, SY81, M20, TAT, M17 and SRY465 (refer to raw data appendix).

Haplogroup 1 was observed among <43% the U.K. Leicestershire chromosomes, <75% of the admixed Islander chromosomes, 55% of the admixed Maori chromosomes, <4% of the non-admixed Maori chromosomes and not at all among the non-admixed Polynesian Islander chromosomes (refer to table 10.2). Jobling et al. (1997) observed haplogroup 1 chromosomes in 47% of European. Thus, the frequency of haplogroup 1 chromosomes in Europeans (Jobling et al. 1997) was only 4% greater than the haplogroup 1 frequency observed in this study.

The overall haplogroup 1 frequency combining all the Polynesian populations was 17%. Hurles et al. (1998) studied Polynesian and Melanesian Y chromosomal lineages and found haplogroup 1 chromosomes among 27% of the Polynesian chromosomes but not at all within the Melanesian chromosomes. It was interesting to observe that Hurles et al.'s (1998) haplogroup 1 frequency in Polynesians was 10% greater than the combined Polynesian populations in the present study. It was highly likely that the present study underestimated the haplogroup 1 frequency across all the Polynesian populations, as a small Polynesian Islander sample size was analysed, hence biasing the frequency statistic.

Haplogroup 1 chromosomes have also been isolated among Northern European and Northern African Jewish populations, as part of a study examining the lineages of Cohanim Jewish priests (Thomas et al. 1998). Thus, the haplogroup 1 samples in the present study had strong affiliations to Europe and Asia. Therefore, it was unlikely that any common ancestral Maori or Polynesian Islander haplotype would be isolated within haplogroup 1 chromosomes.

Microsatellites of Haplogroup 1

The most striking observation were the numbers of identical haplotypes shared between all the populations (see table 10.1). In particular, 8 of 13 chromosomes (62%) within the Maori sample were also observed with identical haplotypes to chromosomes within the U.K. Leicestershire population. Only, three haplogroup 1 chromosomes were observed within the Polynesian Islander population. These were all the same haplotype (microsatellite code: JHSGII) which was also observed within the Maori and UK Caucasian populations. Identical haplotypes within haplogroups was suggestive of a common ancestry (Thomas et al. 1998, de Knijff et al. 1997), thus a European / Caucasian admixture within the New Zealand Maori and Polynesian Islander populations.

In comparison to previous studies, identical haplotypes between the Maori, Polynesian Islander and U.K. Leicestershire chromosomes were observed among Highland Papuan chromosomes (Forster et al. 1998), Cook Islander chromosomes (Hurles et al. 1998), Ashkenazic and Sephardic Israelites (Thomas et al. 2000) and Europeans (Iberian) (Hurles et al. 1999).

Forster et al. (1998) also observed identical haplotypes between the Papua New Guinea and European chromosomes, although that data was not shown. Further corroborating the observation of European / Caucasian admixture within the Maori and Polynesian Islander chromosomes of this study, de Knijff et al. (1997) purported that allele 14 at the DYS19 locus, was predominantly observed among 'Caucasians', primarily of European descent. In this study, allele 14 at the DYS19 locus was observed in all the aforementioned haplotypes that were shared between the Polynesian and U.K. Leicestershire chromosomes.

Microsatellite Network analysis of Haplogroup 1

The U.K. Leicestershire network included seven different haplotypes, all of which were isolated in either one or two chromosomes per haplotype (see figure 10.9). Thus, no modal haplotype indicative of a population specific haplotype was observed (see Thomas et al. 1998). The even distribution of branches of the network as observed in figure 10.9, indicated a stable constant-sized population with no genetic selection (Jobling and Tyler-Smith 1995).

Furthermore, the genetic diversity of the Caucasians was the highest of the populations studied. This directly reflected the numbers of different haplotypes observed but also was consistent with a constant size population with a correlated genealogy (Thomas et al. 1998).

The New Zealand Maori network was more complex than the U.K. Leicestershire network (see figure 10.10), with haplotypes linked by more than one mutation. Four haplotypes were shared between the U.K. Leicestershire and Maori networks. This was anticipated since the purported admixture within the Maori haplogroup 1 was European (refer to raw data in appendix for the origins of the admixed chromosomes). The lack of a 'star-like' network construction, including strong modal haplotype did not indicate any selection or population expansion within the haplogroup 1 New Zealand Maori lineage. The European admixture may have incorporated into the Maori gene pool either through a prehistoric European haplotype (Spurdle et al. 1994) or through contact during the past 300 years (Hurles et al. 1998, Hagelberg et al. 1999).

Comparisons between the U.K. Leicestershire and Maori networks expressed two very different arrays of haplotypes. The U.K. Leicestershire network expressed haplotypes that differed by single mutations following the stepwise mutation model. The Maori network was more complex and was missing haplotypes linking one to another, forming a more 'random assortment' of haplotypes. The random assortment of haplotypes could have been a result of 'random' Y chromosome European contact with the established New Zealand Maori population. For example, if a small random assortment of European male lineages were introduced to the New Zealand Maoris, the chromosomes may not be representative of all European haplotypes, thus observed as a random assortment. The European haplotypes observed in the Maori chromosomes, would only have been introduced to Polynesia during the past 300-400 years with the advent of the first European voyagers (Ward 1972, see also Polynesian prehistory introduction chapter 4). The combination of the low mutation rate of the Y chromosome (Weber and Wong 1993, Jobling and Tyler-Smith 1995 and Heyer et al. 1997) and lack of generation time, may explain why the random assortment of European Y chromosomes in admixed Maori haplotypes were not in mutation / drift equilibrium (see Hartl and Clark 1989).

UEP Haplogroup 2

This haplogroup referred to the derived forms of the loci, SRY10,831 and M13 and the ancestral forms of M9, 92R7, YAP, SRY4064, SY81, M20, TAT, M17 and SRY465.

Haplogroup 2 was observed among 18% of the U.K. Leicestershire chromosomes, 93% of the non-admixed Maori chromosomes, 62% of the non-admixed Islander chromosomes, 25% of the admixed Maori chromosomes and was not observed in any of the admixed Islander chromosomes (see Chapter 10). Haplogroup 2 was clearly the modal haplogroup for the non-admixed Maori and non-admixed Polynesian Islander populations. This haplogroup could not be taken as recent common ancestry of the Maori and Polynesian Islander populations on its own, as it was observed in the U.K. Leicestershire population too.

Microsatellite Polymorphisms of Haplogroup 2

Haplogroup 2 chromosomes have previously been observed in 5% of Melanesian and 58% of Polynesian Y chromosomes (Hurles et al. 1998). Hurles et al. (1998) observed that this haplogroup could not, on its own be taken as evidence of recent common ancestry as the haplogroup itself was commonly observed in African, Indian and European populations (Santos et al. 1999). Hurles et al. (1998) sequenced specific microsatellites with identical haplotypes only observed within Melanesia and Polynesia, and found they were identical. This, Hurles et al. (1998) purported, was suggestive of a common ancestry between Melanesians and Polynesians.

The microsatellite haplotypes observed within UEP haplogroup 2 of this study formed distinct clusters of chromosomes within and between populations. The greatest distinction was a modal haplotype (LKPFHJ) within the Maori and Polynesian Islander populations (for a detailed synopsis of the microsatellite coding system, refer to the Y chromosome results chapter). The 'LKPFHJ' haplotype was observed in 17 of 25 (68%) non-admixed Maori chromosomes and 5 of 8 (63%) non-admixed Polynesian Islander chromosomes and was not observed in the U.K. Leicestershire samples. The LKPFHJ haplotype contained a rare short DYS390 microsatellite allele (19-20 repeats). The short allele has almost exclusively been observed in Western Samoa at a frequency of 70%, Papua New Guinea at a frequency of 15% and Australia at a frequency of 60% (Forster et al. 1998). Furthermore, the modal haplotype observed by Forster et al. (1998) in a Western Samoan population was identical to the modal haplotype observed in the present study.

Sequencing of the shorter allele lengths at the DYS390 locus revealed a different arrangement of their tandem arrays in comparison to the longer repeat alleles (Forster et al. 1998). Typically, the longer alleles were characterized as having a stretch of CTAT repeats interrupted by eight CTGT repeats. The shorter alleles were found to have four of the eight CTGT repeats deleted from the allele sequence, also fewer CTAT repeats at the 3' section of the allele (Forster et al. 1998).

Typically, in the present study, the admixed and U.K. Leicestershire populations had longer DYS390 microsatellite alleles (22-24 repeats). Three different haplotypes were shared between the Maori and U.K. Leicestershire populations, all of which had the longer alleles at the microsatellite locus DYS390. A further haplotype in the Maori population and another in the Polynesian Islander population were shared with chromosomes isolated in the Highland Papuans of a previous study (Forster et al. 1998). Thus, possibly linking the Melanesian and Polynesian regions, which is in agreement with Hurles et al. (1998).

Microsatellite network of Haplogroup2

The U.K. Leicestershire network was similar to the construction of the haplogroup 1 network. No modal haplotype was observed and the four haplotypes were linked by single mutations across various loci.

The New Zealand Maori network was in complete contrast to the U.K. Leicestershire network. Two clearly separated clusters of haplotypes were

UEP Haplogroup 3

This haplogroup was essentially the same as haplogroup 1, although the derived forms at SRY10,831 (A to G substitution) and M17 were observed. No reversions to the ancestral type were observed in this study, as have been isolated elsewhere (Hammer et al. 1998, Santos et al. 1999, Karafet et al. 1999). Only one U.K. Leicestershire chromosome was observed within this haplogroup (see chapter 10; table10.1).

Chromosomes with the same haplogroup have been isolated within Europe, Asia and at low frequencies in Indonesia (Hurles et al. 1998). Karafet et al. (1999), observed the derived form of the SRY10,831 marker within Europe, in particular at a high frequency in a Russian chromosomes and also in northern, central and southern Asian Y chromosomes.

UEP Haplogroup 21

Similarly to haplogroup 3, these chromosomes were only observed in the U.K. Leicestershire population. Haplogroup 21 was characterised as having the derived forms of the UEP markers at SRY10,831, M13, YAP and SRY4064.

This haplogroup has been observed primarily within European and African populations (Karafet et al. 1999), although has also been isolated within Japanese chromosomes (Hammer 1994, Hammer and Horai 1995, Ruiz-Linares et al. 1996). The YAP positive type of haplogroup 21, has been observed at varying frequencies within European studies, ranging from 4% (Santos et al. 1999) to 7% (Ruiz-Linares et al. 1996). In comparison, this present study recorded a higher YAP positive frequency of 18% among the U.K. Leicestershire chromosomes. The higher Yap positive frequency among the U.K. Leicestershire samples in comparison to the 'European' frequency may indicate that sampling chromosomes from a specific region rather than over a wider area, (i.e. the U.K. in general and/or the continent) introduces a bias. With this in mind, the U.K. Leicestershire samples may not be representative of the European gene pool and may even support a regional variation within the U.K.

Microsatellite haplotypes of haplogroup 21

Two different haplotypes were observed, in haplogroup 21, with microsatellite codes IHTFGI and JHTFGI. These differed by one repeat (alleles 13 and 14) at the DYS19 locus. Hammer and Horai (1995) observed the same sized alleles (alleles 13 and 14) were most common in western European populations in comparison to the Japanese populations, where the longer alleles (15-16 repeats) were more frequent.

The U.K. Leicestershire samples were chosen to have at least two or three generations native to the region. The U.K. Leicestershire alleles at the DYS19 locus were of 13 and 14 repeats. Hence, in agreement with Hammer and Horai (1995) that alleles 13 and 14 were most common in western European populations.

UEP Haplogroup 26

This haplogroup was observed at varying frequencies across all the populations. The M9 single nucleotide polymorphism with the transition from the ancestral C allele to the derived G allele characterised haplogroup 26. The derived form of M9 defined a major lineage found in all regions except Africa (Underhill et al. 1997). This form of the polymorphism was found at high frequencies in Asia and Australia (Western Pacific) and was less common in Europe (Karafet et al. 1999). This observation was in agreement with the findings of this present study. Whereby the derived M9 type was observed in 2 of 21 U.K. Leicestershire chromosomes, 6 of 18 Polynesian Islander chromosomes and 5 of 48 New Zealand Maori chromosomes. Haplogroup 26 was previously isolated in Cook Islander and Papua New Guinea chromosomes. It was postulated that the common ancestry of these chromosomes probably derived from the Southeast Asian migration (Hurles et al. 1998), which also contributed the majority of Polynesian mtDNA (Murray-McIntosh et al. 1998).

Microsatellite haplotypes of haplogroup 26

Different haplotypes were observed between the populations in this study. However, one haplotype (microsatellite code: JHSFII) was shared between U.K. Leicestershire and a New Zealand Maori chromosomes. The Maori chromosome originated from an individual that was of admixed European descent (refer to raw data appendix).

Furthermore, the microsatellite haplotype with code: KITFII observed in Maori and Polynesian chromosomes, was also found in Papua New Guinea chromosomes within the same haplogroup and was believed to have originated in southeast Asia (Hurles et al. 1998).

Microsatellite Network of Haplogroup 26

The New Zealand Maori network had formed an interesting cluster of admixed and non-admixed samples (see chapter 10; figure 10.14). Similarly to haplogroup 1 the network construction did not express any modal motif, and haplotypes were linked by more than three mutation steps.

The U.K. Leicestershire network was simple, with only two chromosomes linked by three mutational steps, thus no modal haplotype could be inferred.

The Polynesian Islander network had formed a coherent structure of clearly linked haplotypes. Two small clusters were observed, the first included the KITFII haplotype isolated within two chromosomes and its single-step neighbor with the KJTFII haplotype. The second cluster included the LJTELI haplotype observed in two chromosomes and its single-step neighbor with the LJTELI haplotype. The first cluster was observed in the Tongan / Tahiti region and the second cluster the Tokelau / Samoa region. Interestingly, the separate clustering was possibly a chance event and analysis of more haplogroup 26 chromosomes may reduce the distinct clustering to a more homogenous group. If the clustering within the network was a reflection of male lineages between the Polynesian Islanders, it could be an indication of gene flow with respect to sequential Island hopping that followed a specific route dependent on the wind patterns in the Pacific. Referring to the weather patterns at the time of Polynesian settlement (see chapter 4; figure 4.1), trade winds blew from East to West below the south Pacific 'convergence zone'. Thus, voyages between Tahiti and Tonga would have been reasonably easy. Likewise, 'Easterly' winds blowing North to South towards the 'convergence zone', made voyages between Tokelau and Samoa fairly easy. However, voyages between Samoa and Tonga would have required sailing across the inhospitable convergence zone, where the weather was purported to have variable winds, calms, heavy showers and thunderstorms

(Irwin 1992). Thus, travelling North to South and *vice versa* crossing the convergence zone may not have been as easy and perhaps as frequent as voyages East to West once either side of the zone. Correlations between weather patterns and gene flow have not been studied previously, possibly because of vague knowledge of number of persons per canoe and numbers of canoes travelling at once (but see Murray-McIntosh et al. 1998).

Finally, the LJTELI and LJTEKI haplotypes of the Tokelau and Samoan chromosomes may be the first observation of what has previously been purported as a possible Melanesian only haplotype (Hurles et al. 1998). A previous study observed the microsatellite haplotype equivalent to KJTELI (this study) solely within a Melanesian cluster (Hurles et al. 1998). Distinctively the Tokelau and Samoan haplotypes of this study and the purported Melanesian haplotype (Hurles et al. 1998) had the smallest DYS391 allele and the largest DYS392 allele of all the chromosomes typed in both studies. However, the distinctive haplotype has not been observed in the Highland Papuans or Australians (Forster et al. 1998). Further studies are necessary to clarify the existence of this haplotype within Polynesia and Melanesia and the extent of the diversity of haplogroup 26 within these two regions of the World.

11.c. Commentary of the variance analyses

Analyses of molecular variance (AMOVAs) have been constructive in their use, describing the apportionment of variance within and between populations. Previously, AMOVAs were used to successfully describe male lineage relationships (Rower et al. 1996, Bianchi et al. 1998, Kittles et al. 1998, Scozzari et al. 1999, Santos et al. 1999, Bosch et al. 1999 and de Knijff et al. 1999).

This present study consistently observed the greatest variation within populations than between, regardless of the inclusion or exclusion of specific populations (see chapter 10; table 10.6). This was consistent with findings from other populations including; European Caucasians (de Knijff et al. 1999), African populations (Scozzari et al. 1999 and Bosch et al. 1999) and Finnish provinces (Kittles et al. 1998).

Five different group structures were tested omitting various populations to observe the effect on the variance statistic. The first group structure tested incorporated all the populations (all data). The variance within populations was

68.2% with ϕ_{ST} value of 31.8% indicative of a great differentiation among populations (Hartl 1981). Separating the populations to two groups the U.K Leicestershire and admixed populations were grouped together and the non-admixed Maori and Polynesian Islander populations were grouped together. In doing that, the among population variance increased from 31.8% to 40% among populations. Similarly, the variance within populations decreased from 68.2% to 58.6%, indicating that segregating the admixed populations from the non-admixed populations reduced the variance. The among population variance of 42.1% (ϕ_{ST}= 0.421) was observed when the non-admixed and admixed Maori populations were grouped. The significance of ϕ_{ST} was calculated for the aforementioned group structures and all were 0.000, thus highly significant (de Knijff et al. 1997). However, two further group structures were tested which did not show significant levels of between population genetic variability. The first of these groups clustered the non-admixed Maori and non-admixed Polynesian Islander populations (ϕ_{ST}= 0.029, significance 0.144). The second of these groups clustered the U.K. Leicestershire and admixed Maori populations (ϕ_{ST}= -0.006, significance 0.484). A negative among population value alone indicated the very close genetic male lineage between these populations, which was further qualified by the significance value. Low levels of variability in other studies indicated the Y chromosomes were genetically similar (de Knijff et al. 1997).

11.d. Genetic Structure Analyses

The coancestry coefficients were calculated using the Arlequin program (Schneider et al. 1997). In agreement with the variance analyses, a genetically close male coancestry was observed between the U.K. Leicestershire Caucasian and admixed Maori chromosomes also the admixed Islander and admixed Maori chromosomes with coancestry distances of zero (see results table 10.7). The pair-wise comparison between the non-admixed Maori and Polynesian Islanders also expressed a close genetic relationship with a coancestry distance of 0.029. This result was also anticipated given the extent of shared haplotypes between these populations. However, one should observe that there was a bias in sample size between the populations with a 2:1 ratio of Maori: Polynesian Islander chromosomes. Thus, the coancestry distance may alter indicating a closer or more distant association with an increased number of Polynesian Islander samples. However, the likelihood would be that the genetic relationship could be closer given the previous male lineage studies of Polynesian populations (Hurles et al. 1998, Forster et al. 1998).

11.e. Migration

Migration estimates gene flow between populations in terms of persons per generation (Santos et al. 1997). This statistic directly reflected the coancestry coefficient with an infinite amount of migration between the U.K. Leicestershire and admixed Maori population and the admixed Islander and admixed Maori population (see results table 10.7). Interestingly, the smallest migration (0.376) was not between a U.K. Leicestershire and Polynesian population, but the admixed Islander and non-admixed Maori populations. Once again, this could have reflected the very small sample size of the admixed Islander population (n=5) in comparison to the other populations, which had on average sample sizes three to four times greater.

11.f. Common Ancestry Date

The common ancestry date has been used previously to estimate a time at which lineages diverged. The modal haplotype or root of a haplotype tree within a population may be considered ancestral (Thomas et al. 1998, Hurles et al. 1999). The DYS388 locus was not included in this calculation, because of the observations of previous studies recognizing its lack of conformity to the stepwise mutation model (SMM) (Thomas et al. 1998). Hence, the DYS388 was inappropriate for the ASD statistic that assumed a SMM (Goldstein et al. 1995).

This study estimated the Maori ancestral haplotype to have a common ancestry 1193 years before present (95% CI 448 – 2241 with generation time of 25 years).

Archaeological findings of ancient 'Lapita' pottery have been isolated on Islands from Papua New Guinea to Tonga and Samoa, with radiocarbon dating estimating the pottery to be roughly 3,000 years old (Ward 1972). Linguistic evidence supported this date, with languages purported to have diversified between Islands, with the largest separation occurring with the settlement in eastern Polynesia (300 B.C) (Bellwood 1989).

Settlement of New Zealand was predicted to be dated 1000-2000 ybp, and was not a single colonisation, rather a wave of settlements throughout a few hundred years (Irwin 1992). This present study corroborated Irwins' (1992) prediction, with a calculated common ancestry of 1193 ybp (95% CI 448 – 2241) for the modal haplotype within the Maori population. Although, one must not forget the large confidence interval with the dating method (Thomas et al. 1998), clearly the modal haplotype was ancestral, not only within the Tonga / Samoa region, but also New Zealand.

Interestingly, this study found no evidence of repeat waves of colonisation of New Zealand (Irwin 1992), as only one 'ancestral' modal haplotype was observed. However, the findings of this present study were in agreement with Murray-McIntosh et al. (1998), whereby through mtDNA analyses these workers found little evidence to suggest an earlier permanent settlements. Hagelberg et al. (1994) also observed only one principal mtDNA lineage in Polynesia and purported the findings to be consistent with population bottlenecks during colonization.

11.g. Admixture: Observations and Reasonings

Theories of the sequence of human movement through the Pacific vary.

Diamond's (1988) 'express train to Polynesia' assumed a Lapita culture originated in Island Southeast Asia and spread west to Melanesia, with an Indonesian or southern Chinese origin.

Terrell (1986) however, purported that the origins of Polynesian culture developed between Island Southeast Asia, inland, and island Melanesia. The settlement expansion was thought to be fastest in Southeast Asia and slowest in inland Melanesia, marked by the Lapita archaeological remains (for a detailed synopsis see Polynesian Prehistory Introduction, chapter 4).

Interestingly, Cowan (1930) without linguistic, archaeological or biological evidence available today summarised the origins of the Polynesians. The Polynesians were thought to form part of an ancient Gangetic race present in India from remote antiquity, but were modified by the intrusion of Semites

(Hebrews), Tibetan and other races. Their movement from Asia to New Guinea and then the Western Pacific during the first millenium AD, took hundreds of years and mixing of tribes was doubtless (Cowan 1930).

A third theory also existed to explain the colonisation of the Pacific. This purported an ancient American contact prior the arrival of the first 'Polynesians' (Heyerdahl 1950). This theory was based on basic and scarce archaeological findings, including a 'typically American' Andean sweet potato (for further discussion on the theories of colonisation see chapter 4 and figure 4.1).

Whatever the argument for a prehistoric colonisation of the Pacific, at 3,000 ybp Polynesian ancestry can be placed in the Tonga-Samoa region (Terrell 1986, Hill and Serjeantson 1989).

However, and very importantly, none of these theories had included the extent of European admixture that has diluted the clear origins of the 'natives' in the Pacific. Through mtDNA and Y chromosome analyses not only has female and male lineages and gene flow been assessed but also the possible origins of the lineages.

This present study observed shared haplotypes between distinct populations, for example: within UEP haplogroup 1, the microsatellite haplotype with code: JHTGII was observed in U.K. Leicestershire Caucasian and Maori chromosomes (this study), Cook Islanders (Hurles et al. 1998), Norwegian, Basque and Bearnais (Hurles et al. 1999), Highland Papuans (Forster et al. 1998) and Ashkenazi and Sephardic Israelites (Thomas et al. 2000). However, shared haplotypes between these populations were only observed at low frequencies. Thus, since haplogroup 1 had mainly European origins (de Knijff et al. 1997) the incidence of this group among the Polynesian and Melanesian populations was indicative of European chromosomes entering the 'native' gene pool as admixture. However, if the shared haplotypes were modal between the respective populations then one could assume a more probable common 'ancient' ancestry (see Thomas et al. 1998).

The first recorded European contact with Polynesian populations occurred during the sixteenth century by the voyager Ferdinand Magellan (Web site www.mariner.org/age/magellan.html). Voyagers traversing the Pacific held tempestuous relationships with the native Polynesians (Beaglehole 1934). Magellan was purported to have killed many 'savages' possibly destroying established villages (Beaglehole 1934). This would have reduced the gene pool, possibly removing male lineages, which of course could not be replaced. Furthermore, diseases were introduced to the Pacific which drastically reduced the population from 3.5 million to 2 million over a period of 300 – 400 years (Ward 1972). No doubt, small communities including all family members living in close quarters would have suffered from disease, which would have also destroyed lineages.

Those that sailed to the Pacific not only left disease and destruction but also European male lineages, which have persisted to this day.

In this study, 48 different Maori chromosomes were analysed. Twenty-five of the forty-eight chromosomes were of non-admixed ancestry, which was recorded in accordance by the individuals own admission. However, of the twenty-five chromosomes, three haplotypes were shared with known admixed chromosomes and a further three haplotypes with U.K. Leicestershire chromosomes. Therefore, six of twenty-five (24%) purported non-admixed samples contained admixture. It was therefore conceivable that population studies of the Maori population, in particular autosomal analyses, will be biased towards clustering the Maori closer to the Europeans, than other Polynesian or Melanesian populations. In fact, this had already been documented. A previous study did not show a close relationship between the Maoris and Samoans (Spurdle et al. 1994). This was explained as elevated levels of European gene flow into the Maori population (Spurdle et al. 1994).

In addition to finding admixture within the 'non-admixed' chromosomes, seven non-admixed male lineages (known by their inclusion in the haplogroup 2 modal cluster, 'native' to the Polynesian and Melanesian region) were observed within purported 'admixed' chromosomes. These seven individuals were probably admixed, although the admixture may have been in the female lineage. Thus, the male lineage as analysed in the Y chromosomal study would be reported as native (non-admixed). Autosomal studies would not be affected in this instance as the sample has been reported as admixed and indeed would be if the maternal lineage was not native Maori in origin.

Overall, this study observed just 19 of 48 chromosomes (40%) native to Polynesia. The majority of the male lineages express European admixture, which unfortunately may be the result of recent contact during the past 400 years.

11.h. Summary

This was the first study to analyse male lineages in the New Zealand Maori population. The hypothesis that close genetic relationships between the Maori and other Polynesian Islanders with an appreciable amount of European admixture was in total agreement with the findings of this present study. Haplotypes were not only shared between the populations in this study but also between Melanesian and Asian male lineages. There was no evidence of shared haplotypes between 'native' American Indians and either the Maori or other Polynesian Islanders. Thus, this study could not support the findings of Cowan (1930) who purported ancient contact between the Americas and Polynesian populations in the Pacific. In agreement with Hagelberg et al. (1999) this study found only one highly frequent haplotype within in the Maori chromosomes known to be present throughout Polynesia. This was indicative of one colonization event of New Zealand rather than a series of colonisations, or repeat colonisations with males with extremely similar haplotypes. However, the 'noise' of European admixture today together with severe bottlenecks through conflict and disease, may have obscured any other ancestral lineages. Perhaps, DNA analyses of ancient skeletal/tooth material may in part answer further questions of lineages. Attempts so far have been with mtDNA analyses and have shown shared haplotypes between ancient remains and the lineages in 'todays' gene pool (Hagelberg et al. 1999). Perhaps further studies incorporating ancient DNA will uncover more lineages lost by the 'noise' of admixture of the last millenium.

Chapter 12

Comparing and contrasting issues: The Autosome and Y-chromosome united

The autosomal marker systems and male lineage analyses have been discussed separately. This chapter compares and contrasts the main issues of the separate discussions. The hypothesis that greater similarities than differences among populations at the autosomal STR level would exist, in comparison to male lineages was clearly proven. Additionally when the male lineage data was analysed it became apparent that a percentage of the male 'non-admixed' DNA samples used for autosomal analyses contained genetic admixture. Although this could be detected through the use of Y-chromosome analyses, individual admixed samples could not be detected using autosomal marker systems. This discussion elaborates upon the aforementioned key issues and summarises the significant findings of this study.

Obvious differences between the autosomal and Y-chromosomal analytical systems arise from issues including differences in mutational processes, leading to distinctions between genetic diversity. Also the differences of genetic distances between the two systems can be compared, whereby the relationships among the populations can be examined in terms of purely male interaction and through autosomal analyses, both male and female interaction. Hence, one may expect genetic distance differences comparing the two systems.

12.a. Mutational events compared between autosomal and Y-chromosome loci.

Microsatellites within the autosome contain one set of haploid genetic information from the male and another set from the female parent, by virtue of the process in which inheritance takes place at fertilisation (Edwards et al. 1995). The Y chromosome contains haploid genetic information that passes from father to son usually as an exact copy (Jobling and Tyler-Smith 1995).

However, rare mutations do occur altering the sequence or length. The Y chromosome mutational processes include

- Rare base substitutions known as 'single nucleotide polymorphisms'

- Replication slippage increasing repeat numbers or a deletion decreasing the repeat numbers (Thomas et al. 1998).

Similarly, autosomal mutational processes resulting in either a lengthening or shortening of the overall fragment length, may be the result of

- Integral numbers of unit slippage at replication or
- Unequal recombination between the tandemly repeated sequences, or both (Craig et al. 1988).

The distinguishing factor between the mutational processes of the two systems resides in the absence of recombination on the Y chromosome other than at the pseudoautosomal region (Jobling et al. 1997). Unlike the autosomes the remaining non-recombining portion of the Y chromosome has all its genes in linkage disequilibrium. Hence Y chromosomes are transmitted as an unaltered unit generation to generation (Bianchi et al. 1998).

Differences in mutational processes were reflected by mutation rates. Autosomal mutation rates were estimated up to 1.5 X 10^{-2} (Jin et al. 1996) and Y chromosome mutation rates of about 2.1 X 10^{-3} (Heyer et al. 1997).

Low mutation frequencies as observed within the non-recombining portion of Y chromosomes were accompanied by a reduction of polymorphic diversity (Jobling and Tyler-Smith 1995). In comparison, regular recombination occurs between homologous autosomal chromosomes allowing greater potential for mutation and sequence errors (Levinson and Gutman 1987). In agreement, this study observed a lower diversity among Y chromosome polymorphisms in comparison to the autosomal diversity. For example, the average gene diversity of the Y chromosome haplogroups among the U.K. Leicestershire chromosomes was 67.3% and the corresponding autosomal STR average gene diversity was 83%.

Other than differences in diversity between the autosomal and Y chromosome loci, contrariety in male and female gene flow throughout Polynesia has been observed (Hurles et al. 1998, Murray-McIntosh et al. 1999).

This study reports extensive European male lineages within the Maori and Polynesian Islander populations. Comparing known European haplotypes to non-admixed Maori samples, 10.4% (5 of 48 chromosomes) of Maori chromosomes which were purported to be 'native' Maori in origin, were found to share identical haplotypes to European chromosomes. Isolating this European admixture within a 'supposed' purely native sample population was quite informative for the following reasons.

An admixed male classified as non-admixed was not a problem when examining male lineages within Polynesia, as distinct haplogroup and haplotype structures distinguished major population groups. However, incorrect classification of admixed/non-admixed samples had strong implications of biasing autosomal studies. Autosomal STR analyses incorporate sample populations of known ethnic origin. At point of sample collection records made including details of sex, and ethnic origin, often including parental and grandparental ethnic origin are used to cluster similar individuals into a specific group (Mastana et al. 1998, Chambers, personal communication). Incorporating samples of admixed origin into an autosomal STR study may go undetected at the individual level if incorrect information is given at point of sample collection. Quite often and as observed in this study, greater similarities than differences exist between populations at individual loci (Deka et al. 1995, Jorde et al. 1997). However, admixed samples absorbed as part of a sample population could bias statistical analyses and 'cloud' the correct population structure. Hurles et al. (1998) also acknowledged this problem and remarked the 'noise' of recent European admixture among Polynesians could detract from the true population structure.

This autosomal study contained the 'noise' of admixture within the non-admixed Maori population, qualified by the Y chromosome study. The Y chromosome study therefore provided a suitable warning that admixture was present within purported 'native' samples.

The knowledge that low levels of admixture was present in supposed 'native' samples aided the interpretation of the neighbour joining (NJ) phenogram construction of the autosomal distance measures. The NJ phenogram clustered the non-admixed Maori population closer to the admixed populations (refer to figure 8.11). Similarly, to the Neighbour Joining phenogram, the UPGMA tree construction also clustered the non-admixed Maori population closer to the admixed populations (refer to figure 8.12). Initially this was explained in terms of too few non-admixed Polynesian Islander samples analysed, biasing the genetic distance measures. However, after analysing the male lineages it was possible that because a small number of non-admixed Polynesian Islander samples were analysed, the chances of incorrectly absorbing admixed samples in the non-admixed group were reduced. In fact, within the Y-chromosome study, no shared haplotypes were observed between the non-admixed Islanders and the U.K. Leicestershire chromosomes. Thus, when examining population relationships using autosomal loci if a greater proportion of undetected admixture existed within the non-admixed Maori population in comparison to the non-admixed Islander population, it would artificially increase the genetic distance between these two populations. This theory, in fact, describes a previous observation made by Spurdle et al. (1994). Whereby, closer genetic affinities were observed between European and Maori populations than Maori and Samoan populations, even though all the populations were supposedly native.

Further comparisons could be made between the coancestry genetic distance measures of the autosomal data and the Y chromosome data[1]. The closest genetic relationship was observed between the U.K. Leicestershire and admixed Maori chromosomes in the Y chromosome study and between the admixed Maori and non-admixed Maori of the autosomal study. The differences of 'opinion' between the distance measures can be explained.

Firstly, as previously reported admixed Maori samples had been unknowingly included in the non-admixed population. Hence, the observed close genetic relationship between the admixed and non-admixed Maori in the autosomal study. However, the inclusion of known admixture was from the male lineage and further undetected admixture contributed by admixed females may have been present in the autosomal study.

One must further observe that although samples were grouped according to whether admixture was present or absent, the degree of admixture could also influence the similarity between admixed and non-admixed groups. Thus, the numbers of 'non-native Maori' individuals within any given family tree would influence the percentage admixture among each generation. For example a Maori

[1] One should note that the admixed Islander sample data was not included in the Y chromosome coancestry genetic distance analysis as too few samples were analysed. Therefore, no comparisons between autosomal and Y chromosome data involving the admixed Islander sample population have been made.

individual with 12.5% European admixture signifies that one of their great grandparents was of European descent. Whilst, an individual with 50% European admixture signifies one of either parent was of European descent.

In contrast to the autosomal genetic distance measure, the closest pair of populations in the Y chromosome study were the U.K. Leicestershire and admixed Maori. The difference was due to the examination of just the male lineage. It was easier to distinguish admixed chromosomes from non-admixed chromosomes. Thus, grouping the Maori population into admixed and non-admixed by virtue of either their European or Polynesian haplotypes. The admixed population with predominantly European origins held greater genetic similarity to the U.K. Leicestershire population than the non-admixed Maori. This was not surprising since more shared haplotypes were observed between the admixed Maori and U.K. Leicestershire populations than the non-admixed Maori.

Finally, one last aspect of admixture should be noted. Analyses of Y chromosomes can infer male lineages, similarly analyses of mtDNA can infer female lineages. Analysing both together one may assume that all male and female gene flow can be traced. For the most part this is true. However, the following example may show special cases of admixture, which possibly would go unnoticed. This further corroborates observations and conclusions drawn from the phenogram constructions of the populations using the autosomal data.

SEX:		
ORIGIN:	NATIVE MAORI	EUROPEAN
CONTRIBUTION TO THE 1ST GENERATION:	AUTOSOMAL XX mtDNA	AUTOSOMAL XY
POSSIBLE OFFSPRING 1ST GENERATION:	Native mtDNA Admixed autosomes 1 Native X, 1 European X	Native mtDNA Admixed autosomes 1 Native X 1 European Y

Then suppose the 'admixed' daughter paired with a native Maori male, then:

SEX:		
ORIGIN:	50% Maori 50% European	NATIVE
CONTRIBUTION TO THE 2^{ND} GENERATION:	Native mtDNA Admixed autosomes 1 Native X, 1 European X	Native autosomes 1 Native X 1 Native Y
POSSIBLE OFFSPRING 2^{ND} GENERATION:	Native mtDNA Admixed autosomes 1 Native or European X 1 Native X	Native mtDNA Admixed autosomes 1 Native or European X 1 Native Y

At the first generation, admixed males would be easier to detect than females and at the second generation, admixture within either sex would be difficult to detect. Admixture of the kind described by the second generation would be of little consequence if one only traced specific lineages and their movement between populations. However, the extent of foreign admixture as described above can not be accounted for through mtDNA or Y chromosome analyses.

Population genetics may not be concerned with the extent of admixture other than controlling the background 'noise' it creates (see Hurles et al. 1998), but it may be more significant in Forensic analyses (Shriver et al. 1997). Weir (1996 Pp253) commented that disequilibrium caused by an initial admixture would decline over time because of recombination between the loci, although this would be slowed by a continual genetic contact with founding populations. However, studies have been carried out using autosomal marker systems to estimate 'ethnic-affiliations' within populations (Shriver et al. 1997). It was believed that it would be feasible to estimate individual admixture so that interethnic individuals (first or second generation hybrids of one or more populations), could be classified appropriately (Shriver et al. 1997). This was deemed important for forensic investigations involving human remains, whereby the ethnic origin was difficult to assume through the classification of skeletal material (Shriver et al. 1997). This present study did not observe any autosomal loci with population specific alleles and so could not infer the ethnic origin of individuals.

12.b. In summary

This complete study provided novel population and forensic information with respect to 10 autosomal tetranucleotide short tandem repeat systems and Y chromosome haplogroup and haplotype analyses.

Comparing and contrasting the two analytical systems gave greater insight to the genetic dynamics within the populations, especially with respect to admixture issues, which may otherwise have been overlooked.

The significant findings of this study can be summarised thus;

Y chromosome

- A modal haplotype was detected in the Maori and other Polynesian Islander populations which has only been detected in the Polynesian / Melanesian regions of the World. The modal haplotype was considered ancestral and native to those regions of the World.
- Common ancestry dates of the modal haplotype of the Maori chromosomes were in agreement with anthropological (Terrell 1986) and archaeological (Ward 1972) dates of colonization of New Zealand.
- Extensive European admixture was observed which was consistent with previous Y chromosomal (Hurles et al. 1998) and autosomal (Hagelberg et al. 1999) studies.

Autosomal findings

- All loci were highly polymorphic, highly informative thus of good potential for use in both Population genetic and Forensic analyses.
- Gene diversities across loci per population were within the expected range for STR loci and cxpressed greatest diversity in the U.K. Leicestershire population and least diversity in the non-admixed populations.
- Genetic distance measures accurately described the relationships between the populations, thus the 10 STR systems were so sensitive that even closely related populations were distinguished.
- Interindividual ethnic origins could not be assumed using these autosomal systems.

Overall, this thesis provided an opportunity to examine the New Zealand Maori in close detail. This research has not only provided novel population genetic data on 10 autosomal STR systems, but also novel male lineage data on the New Zealand Maori population. Where few comparisons could be made between this and previous studies using the same autosomal systems, comparisons between studies incorporating the same Y chromosome UEPs and STRs were highly informative.

Appendix

Methodological Appendix

A.1 DNA Quantification

DNA quantification of the New Zealand Maori and Polynesian Islander samples was estimated to allow known concentrations of DNA to be administered to the PCR reaction. Table 1 lists all the samples and their concentrations in µg/ 50µl. Samples were identified by their number as used by Wellington University (whom provided the DNA samples).

Table 1: Concentration of Polynesian DNA samples, quantified using ethidium bromide stained agarose gels by visual comparison to known DNA concentration standards.

Sample	Conc µg/	*Sample*	Conc µg/	*Sample*	Conc µg/	*Sample*	Conc µg/
	50µl		50µl		50µl		50µl
59	3	*116*	1	*456*	10	*513*	10
60	5	*401*	15	*457*	2.5	*514*	10
61	10	*404*	10	*458*	3	*515*	7.5
63	5	*406*	2.5	*459*	10	*516*	7.5
64	7.5	*409*	1	*460*	1	*517*	2.5
65	7.5	*410*	2.5	*462*	2.5	*518*	3
66	10	*411*	1	*463*	7.5	*520*	5
67	10	*412*	2.5	*464*	4	*521*	5
72	3	*413*	2.5	*465*	4	*523*	5
73	3	*414*	10	*466*	4	*524*	2.5
74	2	*417*	10	*467*	-	*525*	2.5
75	2	*418*	10	*468*	4	*527*	5
76	5	*419*	10	*469*	3	*528*	2.5
77	10	*420*	10	*470*	7.5	*529*	5
78	2	*421*	10	*471*	7.5	*530*	2.5
79	5	*422*	10	*472*	2.5	*531*	2.5
82	5	*423*	10	*473*	1	*532*	4
83	10	*424*	10	*474*	1	*533*	4
84	3	*425*	10	*475*	1	*534*	10
86	10	*426*	10	*476*	3	*600*	7.5
88	2.5	*427*	10	*477*	32.5	*601*	3
90	2.5	*428*	10	*478*	1	*602*	-
91	2.5	*429*	10	*479*	10	*603*	5

Sample	Conc µg/ 50µl	Sample	Conc µg/ 50µl	Sample	Conc µg/ 50µl	Sample	Conc µg/ 50µl
92	2.5	432	10	480	5	604	2.5
93	1	433	10	481	5	605	7.5
94	-	434	10	482	10	606	10
95	5	435	10	484	3	607	2.5
97	2.5	436	10	485	5	608	3
98	5	437	10	487	10	609	3
100	3	439	2.5	488	7.5	610	7.5
101	3	440	5	489	10	611	5
102	3	441	2.5	500	10	612	5
103	3	442	5	501	10	613	8
104	3	443	2.5	502	10	614	7.5
106	3	444	10	503	10	615	10
107	3	446	10	504	10	616	2
108	3	448	3	505	2.5	617	2.5
109	3	449	10	506	10	618	7.5
110	3	450	10	507	10	621	10
111	3	451	10	508	2	622	2.5
112	3	452	7.5	509	5	623	5
113	3	453	10	510	7.5	624	10
114	3	454	10	511	10	625	10
115	3	455	10	512	2.5	-	-

N.B Sample numbers 100-115 were highly degraded. These were all Polynesian Islander samples.

Sample numbers 94, 467 and 602 did not appear to show any recovered DNA at all.

A.2 Tris-EDTA buffer

The buffer was made using reagents and concentrations as listed in table 2. The final solution was adjusted using NaOH to provide pH 8.0.

TABLE 2:TRIS – EDTA PH 8.0 (TE) BUFFER FOR USE AS THE REHYDRATE MEDIUM FOR DNA STORAGE

Components	Concentration
Tris-HCL	10mM
EDTA pH7.4	1mM

A.3 Running TAE Buffer for use with Precast Spreadex Gels

The composition of the precast gels gives sharp bands with 30mM TAE (Tris-acetate EDTA) as the running buffer (table 3).

TABLE 3: A 40X STOCK SOLUTION OF TAE BUFFER

Components	Amount for 1 litre (40 X)
Tris(hydroxymethyl) aminomethane	145.37g
$Na_2EDTA . 2H_2O$	11.16g
Acetic Acid (glacial)	34.4ml

A.4 Running TBE Buffer for Agarose electrophoresis

This buffer was made following the reagent/concentrations as listed below (table 4):

TABLE 4: A 10X STOCK SOLUTION OF TBE BUFFER

Components	Amount for 1 litre (10 X)
Tris-HCl (pH 8.0)	0.89M
EDTA	20mM
Boric Acid	0.89M

A.5 Allelic Ladders

The CTT and FFv allelic ladders used by the Foresnsic Science Service encompassed the complete range of sizes of the ten systems used in the present study. Thus provided a means withi which to compare unkown sized bands to known sized bands. Table 5 lists the ladders and the sizes of the alleles.

TABLE 5: THE ALLELES AND BASE PAIR SIZES OF THE MULTIPLEX CTT AND FFV ALLELIC LADDERS

CTT								
Locus	Allele	Base Pair	Locus	Allele	Base pair	Locus	Allele	Base Pair
CSF1PO	15	327	TPOX	13	252	THO1	11	203
	14	323		12	248		10	199
	13	319		11	244		9	195
	12	315		10	240		8	191
	11	311		9	236		7	187
	10	307		8	232		6	183
	9	303		7	228		5	179
	8	299		6	224			
	7	295						
FFv								
Locus	Allele	Base Pair	Locus	Allele	Base Pair	Locus	Allele	Base Pair
F13AO1	16	331	Fes/Fps	14	250	vWA	20	163
	15	327		13	246		19	159
	14	323		12	242		18	155
	13	319		11	238		17	151
	12	315		10	234		16	147
	11	311		9	230		15	143
	9	307		8	226		14	139
	8	299		7	222		13	135
	7	295						
	6	291						
	5	287						
	4	283						

A.6 Preparation of samples for sequencing

In order to provide a pure amplified sample for sequencing purposes, all residual primers, dNTP's and enzyme were removed. A commercially available kit was used for this purpose. The method of purification is given below:

A.6.a Method for Qiagens PCR Purification Kit

This was used to prepare samples for DNA sequencing.

Following Qiagens protocol:

Preliminary stages

- Ethanol (96 – 100%) was added to Buffer PE before use
- All centrifugation was carried out at 10-13,000 rpm

1. 5 volumes of Buffer PB was added to 1 volume of the PCR product and mixed.
2. The QIAquick spin column was placed into the provided 2 ml collection tube.
3. The DNA was bound by applying the sample to the column and centrifuging for 30 – 60 seconds.
4. The 'flow-through' was discarded.
5. 0.75 ml Buffer PE was added to the column as a wash, and centrifuged 30 – 60 seconds.
6. Flow-through was discarded. A further 1 minute centrifugation was required to remove residual fluid.
7. QIAquick column was placed into a clean 1.5 ml tube.
8. DNA was eluted by adding 50μl Buffer EB (10 mM Tris-HCl, pH 8.5) and centrifuged for 1 minute.

The DNA was then prepared for quantification and subsequent use in sequencing processes.

A.6.b DNA Dipstick Kit (Invitrogen)

This simple visual method of quantification, assayed the DNA per 1μl of each serial dilution.

The dilutions chosen for quantification were;

Undiluted, tenfold dilution and one hundred fold dilution.

Method:

1. Spot 1μl of each sample dilution on the dipstick membrane. Allow to air dry.
2. Place dipstick in cuvette 1, with 1ml Wash Solution for 10 seconds.
3. Place dipstick in cuvette 2, with 1ml Coupling solution for 3 minutes.
4. Rinse dipstick with deionized water for 20 seconds.
5. Replace in cuvette 1 containing Wash Solution for 4 minutes.
6. Prepare Devloping Solution in cuvette 3 by adding one drop of Developer to 1ml of Developer Stock. Cap cuvette and inert to mix.
7. Place dipstick in cuvette 3 containing Developing Solution for 2 minutes.
8. Gently rinse Dipstick in cuvette 1 contaiing Wash Solution for 20 seconds. Allow to dry.
9. Determination of nucleic acid concentration was done by compariosn of colour intensity with standards.

A.6.c Sequencing results

DNA sequencing was performed by Alta Bioscience at the University of Birmingham and six samples analysed courtesy of Mrs Fisher at Loughborough University.

The sequences were generated using the ABI Prism technology. Each nucleotide base is colour coded, whereby;

T is red, A is green, G is black and C is blue.

The graphical display expresses sequential peaks of different colours pertaining to the nucleotide base. Parallel to each peak, the base is written, forming a continous string of nucleotides and the DNA sequence. The peaks vary in height due to the varying signal intensity of each nucleotide. In positions where the nucleotide is indeterminate, a ‘N' is used to depict uncertainty.

As a guide to the size of the sequenced oligonucleotide, 10 base pair increments are also included on the print out. These will not give an accurate sizing measurement although they are useful to guide one to a region within the sequence.

The sequences have been organized clustering sequences of particular loci as listed in the table overpage;

SAMPLE	LOCUS	COMMENTS	FIGURE NUMBER
528+524	D10s526	As a test of heteroduplex formation these two DNA samples (of known homozygosity @ D10S526) were mixed and amplified. The PCR products containing 2 expected and 1 anomalous band were sequenced. All 3 bands contained the same the same repeat motif and similar nucleotide sequence. Thus the anomalous band was the suspected 'heteroduplex'	297
611	D12S297	All alleles expressed the same nucleotide sequence with no anomalies. Thus, extraneous high molecular weight (HMW) band was the 'heteroduplex'	298
518	D12S297	As above	299-301
82	D3S1514	As above	302-303
608	D4S2285	As above	304
615	D4S2285	Homozygous at this locus, with similar profile to sample 608	305
508	D5S592	This nucleotide sequence was similar to the sequence determined by the Utah Marker Development group (personal communication)	306-308
505	D5S592	Similar to the above profile	309-310
515	D10S520	All alleles expressed the same nucleotide sequence with no anomalies. Thus, extraneous high molecular weight (HMW) band was the 'heteroduplex'	311-312
521	D10S520	Homozygous at this locus, with similar profile to sample 515	313
63	D9S252	Homozygous at this locus, with a similar profile to that of the Utah Marker Development group (personal communication)	314-315
489	D7S1485	As above	316-317
62	D7S618	As above	318
112	D1S407	As above	319-320

ABI PRISM

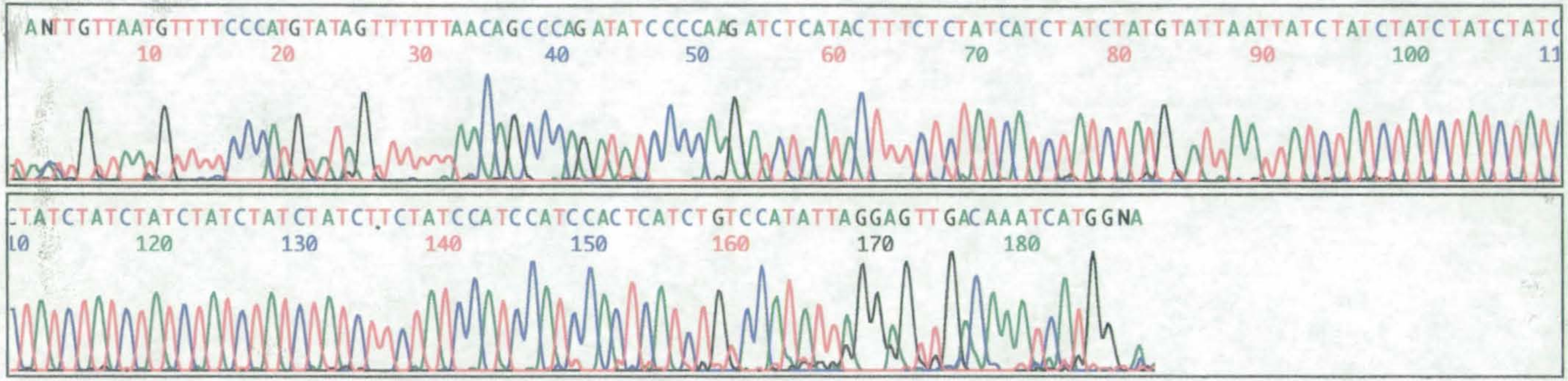
Model 377
Version 3.3
ABI200
Version 3.2

04•63 D9S252R

63 D9S252R
Lane 4

Signal G:42 A:151 T:41 C:74
DT {BD Set Any-Primer}
377 Matrix
Points 1200 to 2950 Pk 1 Loc: 1130

Page 1 of 1
Mon, Jun 14, 1999 11:36
Tue, Feb 09, 1999 18:33
Spacing: 9.00{-9.00}



BI
RISM™

Model 377
Version 3.2
ABI100
Version 3.2

N16075C=19-RELOAD
MISS EMMA WATKINS PRIMER 10s 520
N16075C=19-RELOAD
Lane 93

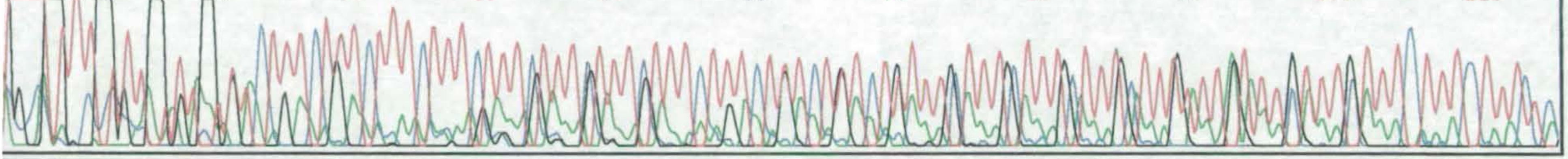
521

Signal G:843 A:688 T:1489 C:765
DT {BD Set Any-Primer}
dRhod-matrix
Points 1062 to 4000 Pk 1 Loc: 949

Page 1 of 1
Thu, Jan 28, 1999 10:13 AM
Wed, Jan 27, 1999 4:24 PM
Spacing: 9.00{-9.00}

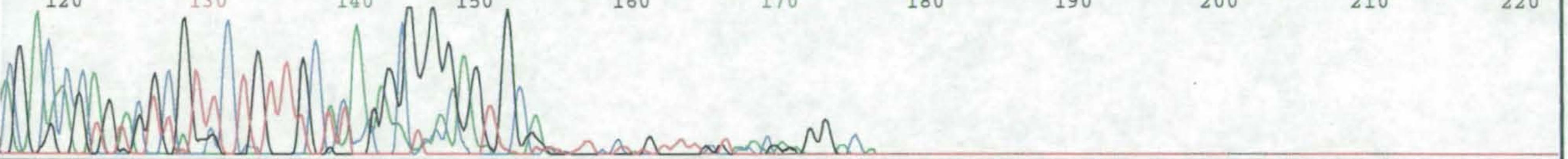
TTTGG TTGGNTTGAT NGNN NCTTTCTNTCTTTCTTTCTTTCTATTNTTTNTTTCTTTCTNTCTNTCTNTTTNTTTCTTTCTTNCTTNATTNNTTGNTTNTTTC TTT CTT CTC

10 20 30 40 50 60 70 80 90 100 110



NG AC N CAGANNCGTT CTGTTGC NNANAGGGNGGAGTGC ANN TCCNGTTTCTNACAAGGACANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN

120 130 140 150 160 170 180 190 200 210 220



NNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN

230 240 250 260 270 280 290 300

299

298

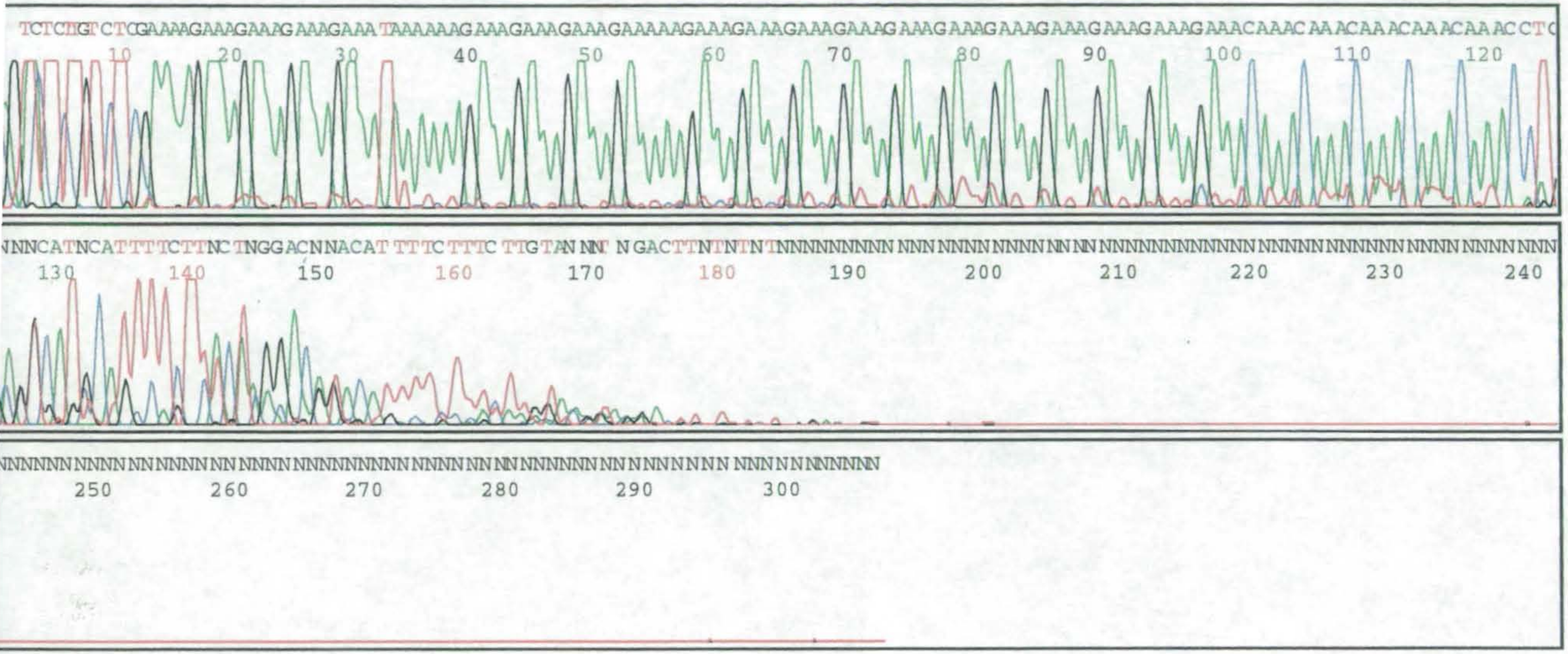
ABI PRISM

Model 377
Version 3.2
ABI100
Version 3.2

N16074H=16 5/5
MISS EMMA WATKINS PRIMER 10s 520
N16074H=16
Lane 45

Signal G:715 A:1283 T:318 C:626
DT {BD Set Any-Primer}
dRhod-matrix
Points 1042 to 4000 Pk 1 Loc: 943

Page 1 of 1
Thu, Jan 28, 1999 10:13 AM
Wed, Jan 27, 1999 4:24 PM
Spacing: 9.85{9.85}



TCTCTTGT CTCGAAAAGAAAGAAAG AAAGAAA TAAAAAAG AAAGAAAGAAAGAAAAAGAAAGA AAGAAAGAAAG AAAG AAAG AAAG AAAG AAAG AAAG AAACAAAC AA ACAA ACAA ACAA AC CTC

NNNCATNCATTTTCTTNCTNGGACNNACAT TTTCTTTC TTGTAN NNT N GACTTNTNTNTNNNNNNNNN NNN NNNN NNNN NN NN NNNNNNNNNNNNN NNNNN NNNNNN NNNN NNNN NNN

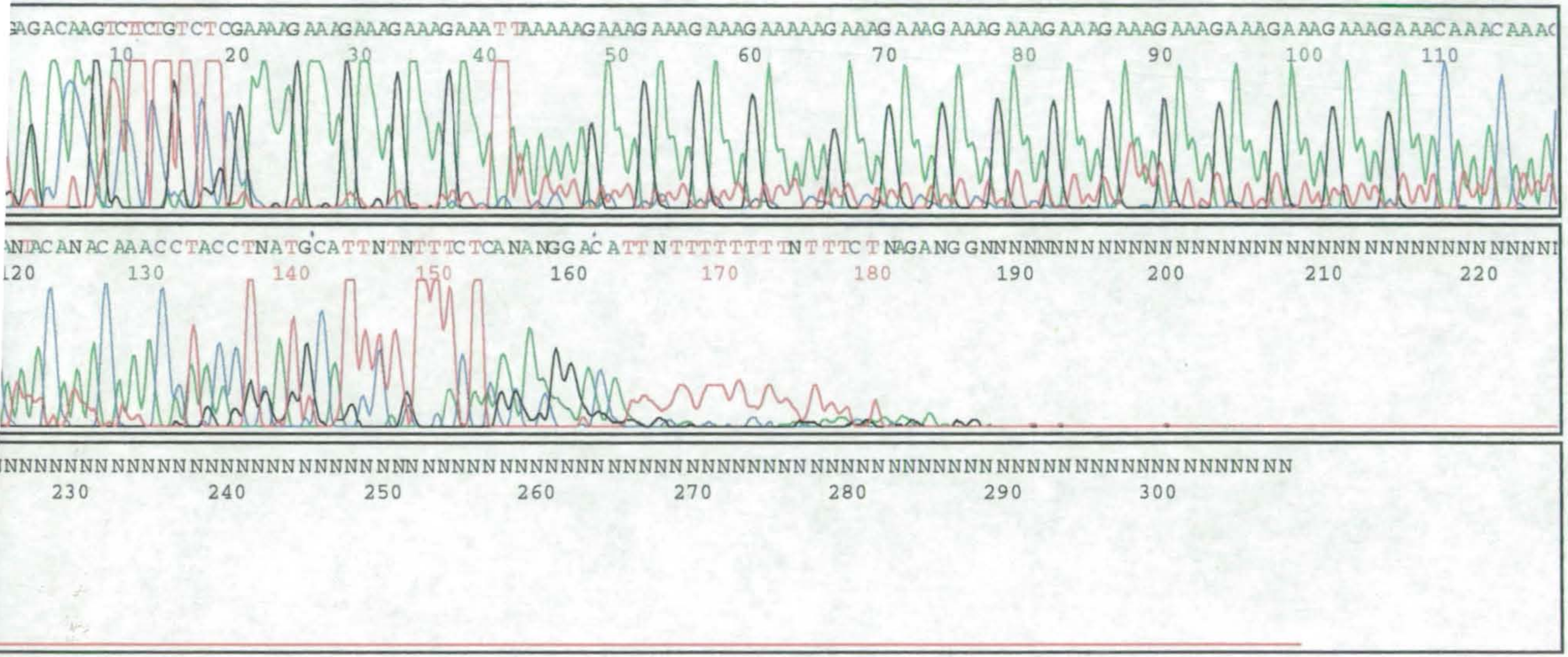
NNNNNN NNNN NN NNNNN NN NNNN NNNNNNNNNN NNNN NN NN NNNNNNN NN N NNNN N N N NNN N N NNNNNN

Model 377
Version 3.2
ABI100
Version 3.2

N16075A=17 515
MISS EMMA WATKINS PRIMER 10s 520
N16075A=17
Lane 5

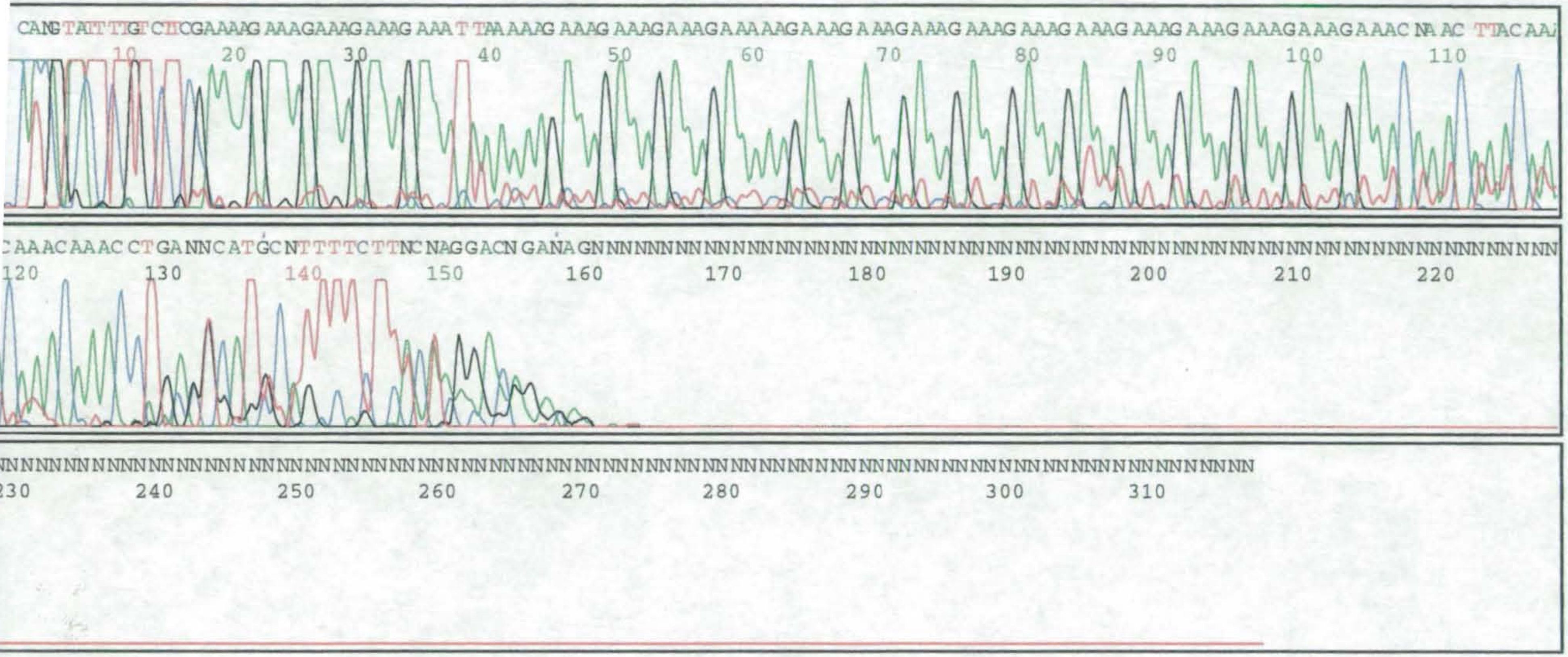
Signal G:1267 A:2134 T:698 C:1142
DT {BD Set Any-Primer}
dRhod-matrix
Points 1002 to 4000 Pk 1 Loc: 874

Page 1 of 1
Thu, Jan 28, 1999 10:13 AM
Wed, Jan 27, 1999 4:24 PM
Spacing: 9.00{-9.00}



300

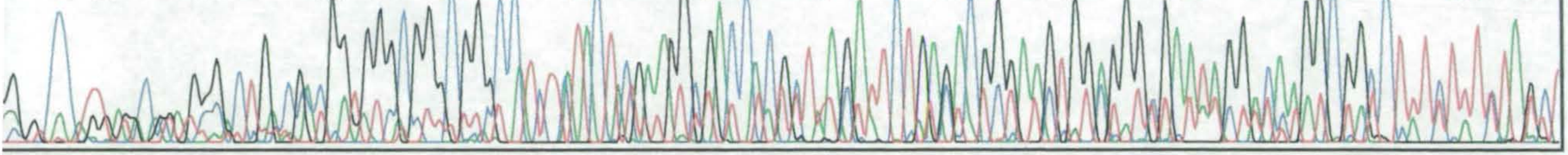
Model 377 | N16075B=18 5/5 | Signal G:1161 A:1975 T:523 C:934 | Page 1 of 1
Version 3.2 | MISS EMMA WATKINS PRIMER 10s 520 | DT {BD Set Any-Primer} | Thu, Jan 28, 1999 10:13 AM
ABI100 | N16075B=18 | dRhod-matrix | Wed, Jan 27, 1999 4:24 PM
Version 3.2 | Lane 11 | Points 1022 to 4000 Pk 1 Loc: 874 | Spacing: 9.00{-9.00}



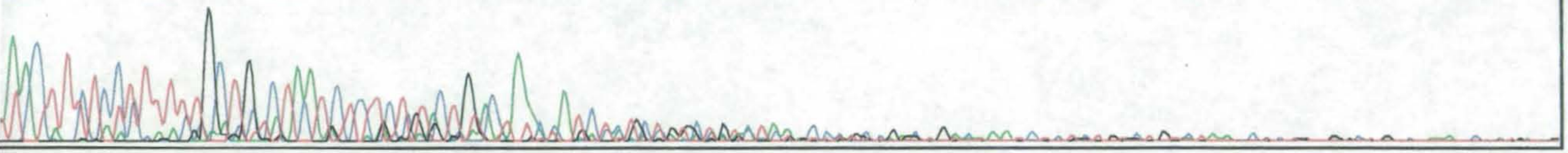
```
CANGTATTTTGTCTTCGAAAAG AAAGAAAGAAAG AAAT TAAAAAG AAAG AAAGAAAGAAAAGAAAGAAAGAAAG AAAG AAAG AAAG AAAG AAAG AAAGAAAGAAAC NAAC TTACAA
CAAACAAACCTGANNCATGCNTTTTCTTNCNAGGACNGANAGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

```
Model 377       N16074F=14  505                        Signal G:356 A:280 T:426 C:492      Page 1 of 1
Version 3.2     MISS EMMA WATKINS  PRIMER 5s 592       DT {BD Set Any-Primer}              Thu, Jan 28, 1999  10:13 AM
ABI100          N16074F=14                             dRhod-matrix                        Wed, Jan 27, 1999  4:24 PM
Version 3.2     Lane 33                                Points 1002 to 4000   Pk 1 Loc: 946 Spacing: 10.17{10.17}
```

```
GAGNCCANTGA NCAN AGNGACT GA CGNC GGTGGGCGGGCGGGCC TCT NTACT CGA AGGC GAC CAC GNTAA GATTCTGANACGGGAAGTGGNG GGNGA ATA GGNCAAGGCGGC CTTT NTTT NTAN CT
          10        20        30        40        50        60        70        80        90        100       110       120
```

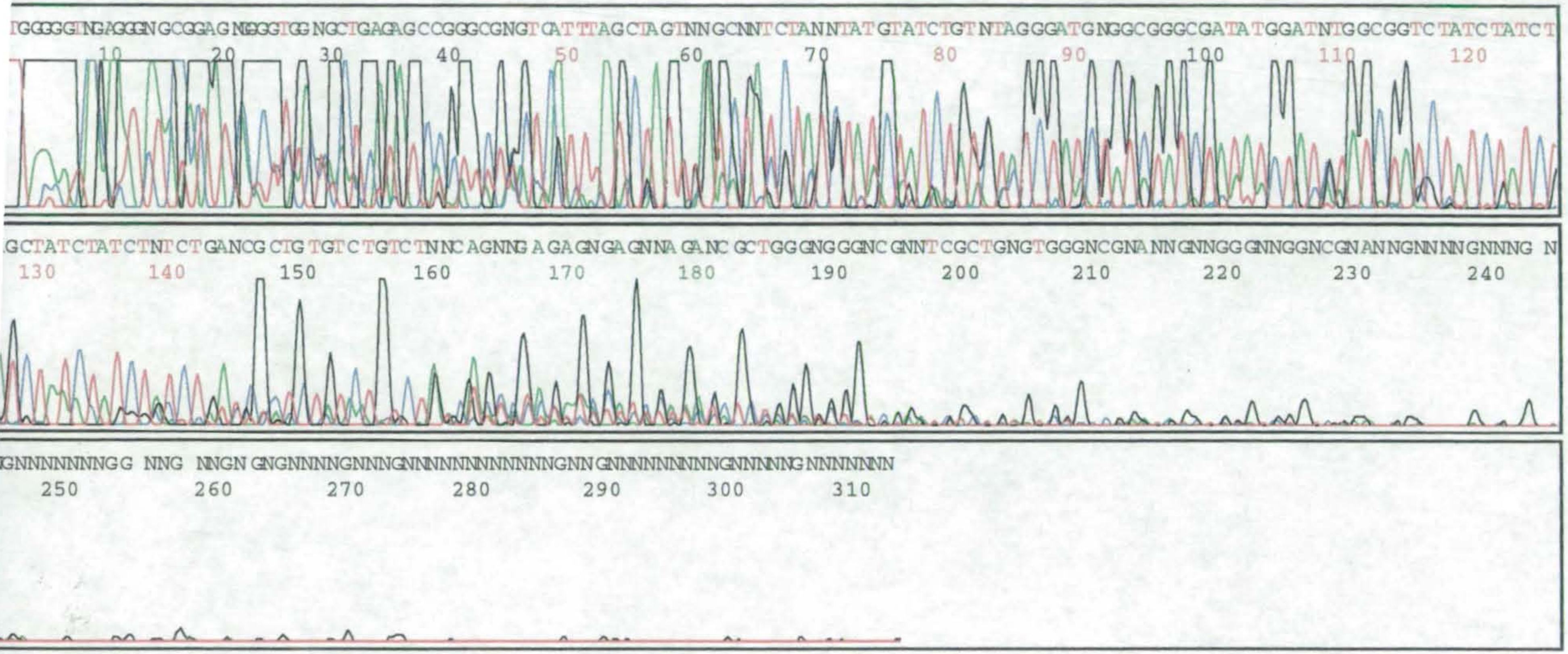


```
AAC TT NTC CNTTTTTGCTG TCTAATC TC TCTG ACT G ACTAACCA TCTNNCTGNCTGTNTA ANCCNNCC GACNGACNAANCNCTCTG  CTGNCTANNCN NNNNGNNNG  NNNANCNNNNNNN
130       140       150       160       170       180       190       200       210       220       230       240
```



```
NNNNNN NNNNNNN NN NNNNNNNNNNNNNNN NN NN NNNNNNNNNNNNNNN NN NNNNNNNNNNN NNNNN NNN
 250       260       270       280       290       300
```

```
Model 377        N16074G=15  505                    Signal G:377 A:427 T:709 C:721          Page 1 of 1
Version 3.2      MISS EMMA WATKINS  PRIMER 5s 592   DT {BD Set Any-Primer}                  Thu, Jan 28, 1999  10:13 AM
ABI100           N16074G=15                         dRhod-matrix                            Wed, Jan 27, 1999  4:24 PM
Version 3.2      Lane 39                            Points 982 to 4000   Pk 1 Loc: 944      Spacing: 10.06{10.06}
```



```
TGGGGTNGAGGGNGCGGAGNGGGTGGNGCTGAGAGCCGGGCGNGTCATTTAGCTAGTNNGCNNTCTANNTATGTATCTGTNTAGGGATGNGGCGGGCGATATGGATNTGGCGGTCTATCTATCT
          10        20        30        40        50        60        70        80        90        100       110       120

GCTATCTATCTNTCTGANCGCTGTGTCTGTCTNNCAGNNGAGAGNGAGNNAGANCGCTGGGNGGGNCGNNTCGCTGNGTGGGNCGNANNGNNGGGNNGGNCGNANNGNNNNGNNNG N
  130       140       150       160       170       180       190       200       210       220       230       240

GNNNNNNNNGG NNG NNGN GNGNNNNGNNNGNNNNNNNNNNNNNGNNGNNNNNNNNNNGNNNNNGNNNNNNN
    250       260       270       280       290       300       310
```

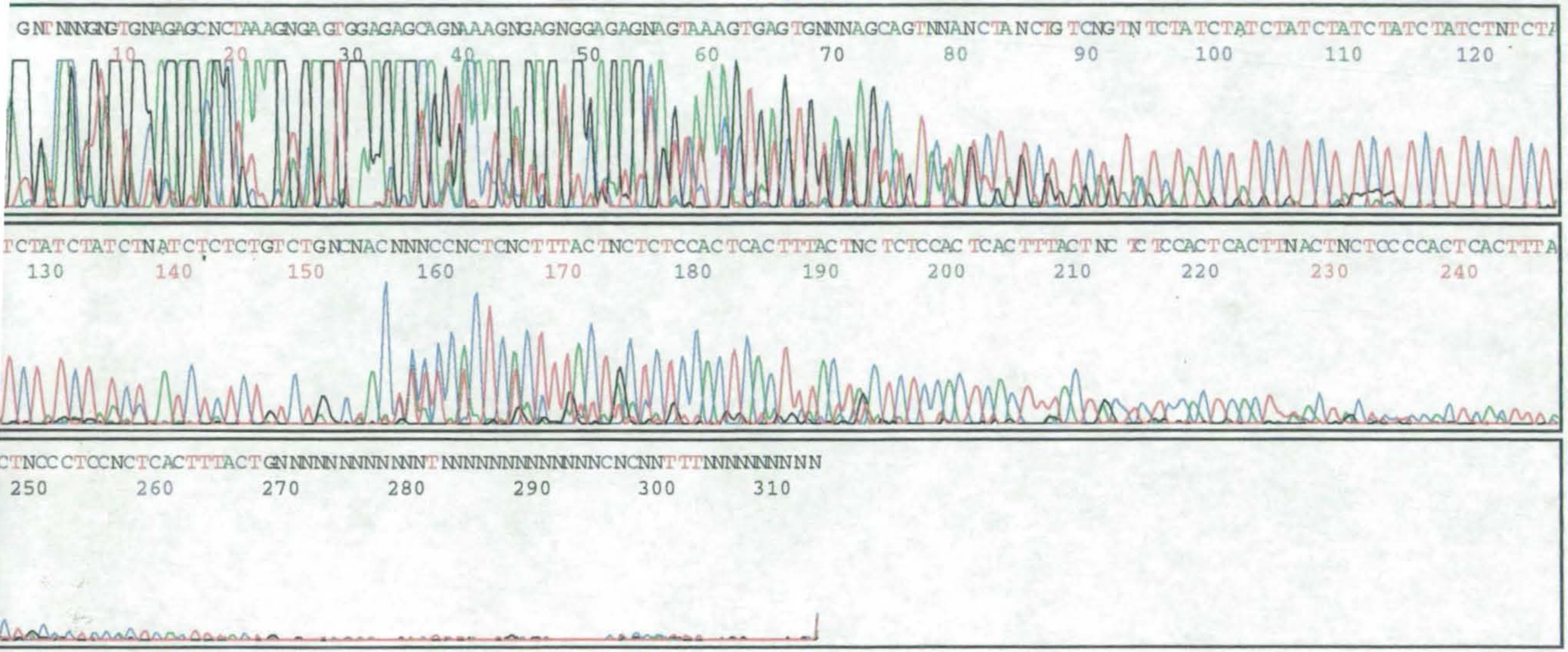
BI
RISM™

Model 377
Version 3.2
ABI100
Version 3.2

N16074E=13 508
MISS EMMA WATKINS PRIMER 5s 592
N16074E=13
Lane 27

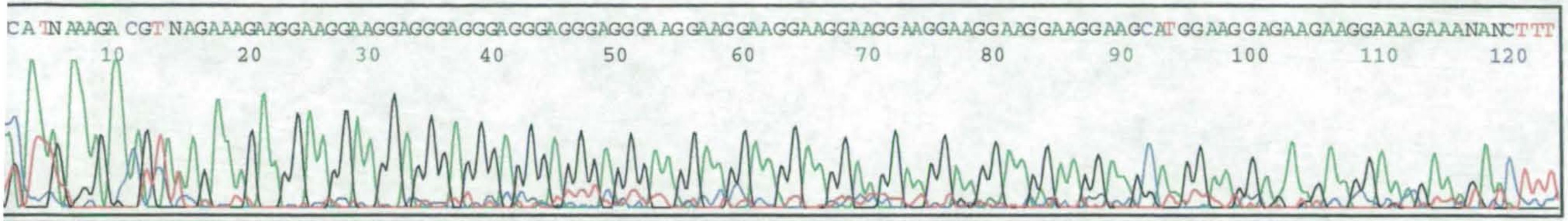
Signal G:305 A:278 T:294 C:290
DT {BD Set Any-Primer}
dRhod-matrix
Points 1002 to 4000 Pk 1 Loc: 874

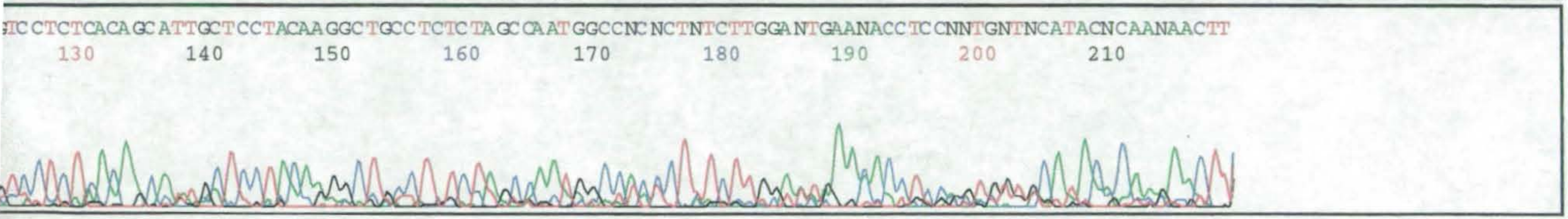
Page 1 of 1
Thu, Jan 28, 1999 10:13 AM
Wed, Jan 27, 1999 4:24 PM
Spacing: 10.10{10.10}



615

Model 377	N16074C=11/REP	Signal G:2170 A:1596 T:634 C:1505	Page 1 of 1
Version 3.2	MISS EMMA WATKINS PRIMER 4s2285	DT {BD Set Any-Primer}	Mon, Feb 1, 1999 10:39 AM
ABI100	N16074C=11/REP	dRhod-matrix	Fri, Jan 29, 1999 4:34 PM
Version 3.2	Lane 61	Points 1042 to 3000 Pk 1 Loc: 948	Spacing: 9.85{9.85}





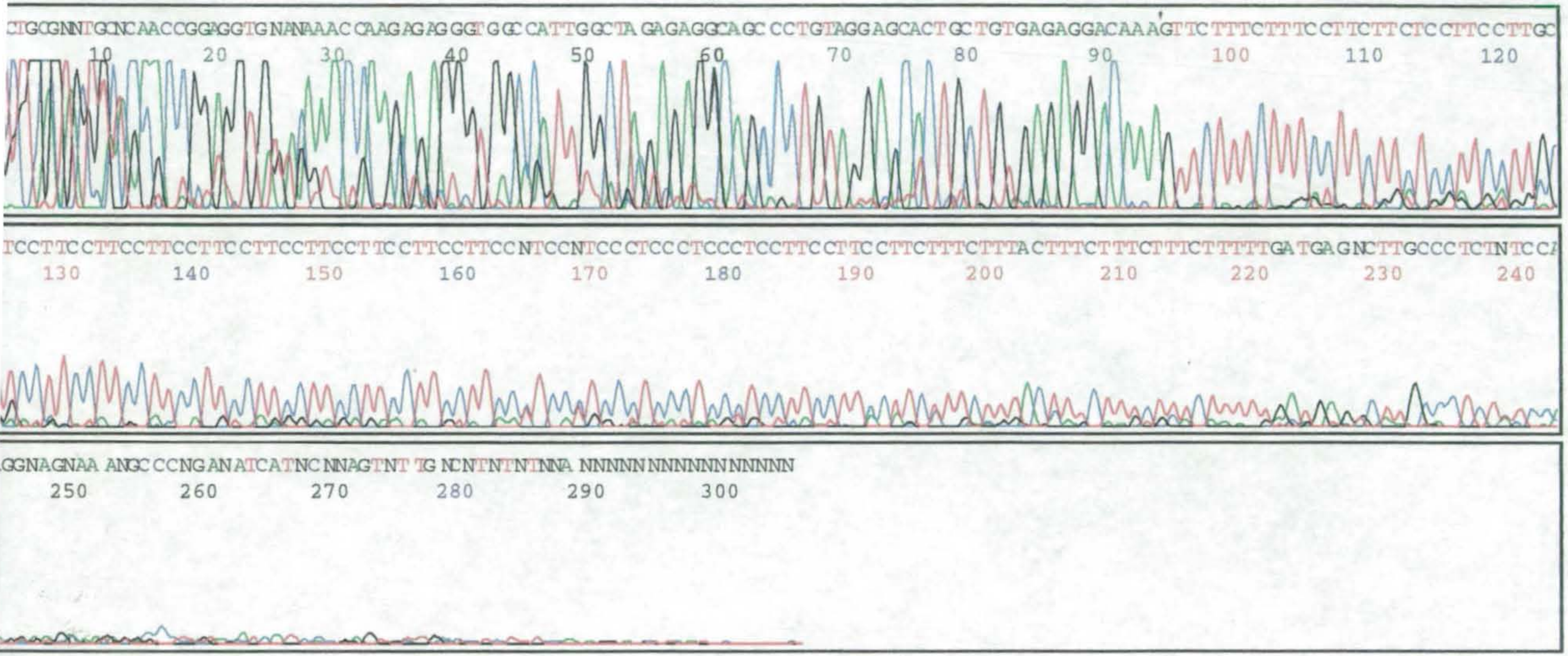
608

Model 377
Version 3.2
ABI100
Version 3.2

N16073H=8
MISS EMMA WATKINS PRIMER 4s 2285
N16073H=8
Lane 43

Signal G:427 A:281 T:241 C:531
DT {BD Set Any-Primer}
dRhod-matrix
Points 1062 to 4000 Pk 1 Loc: 943

Page 1 of 1
Thu, Jan 28, 1999 10:13 AM
Wed, Jan 27, 1999 4:24 PM
Spacing: 9.96{9.96}



306

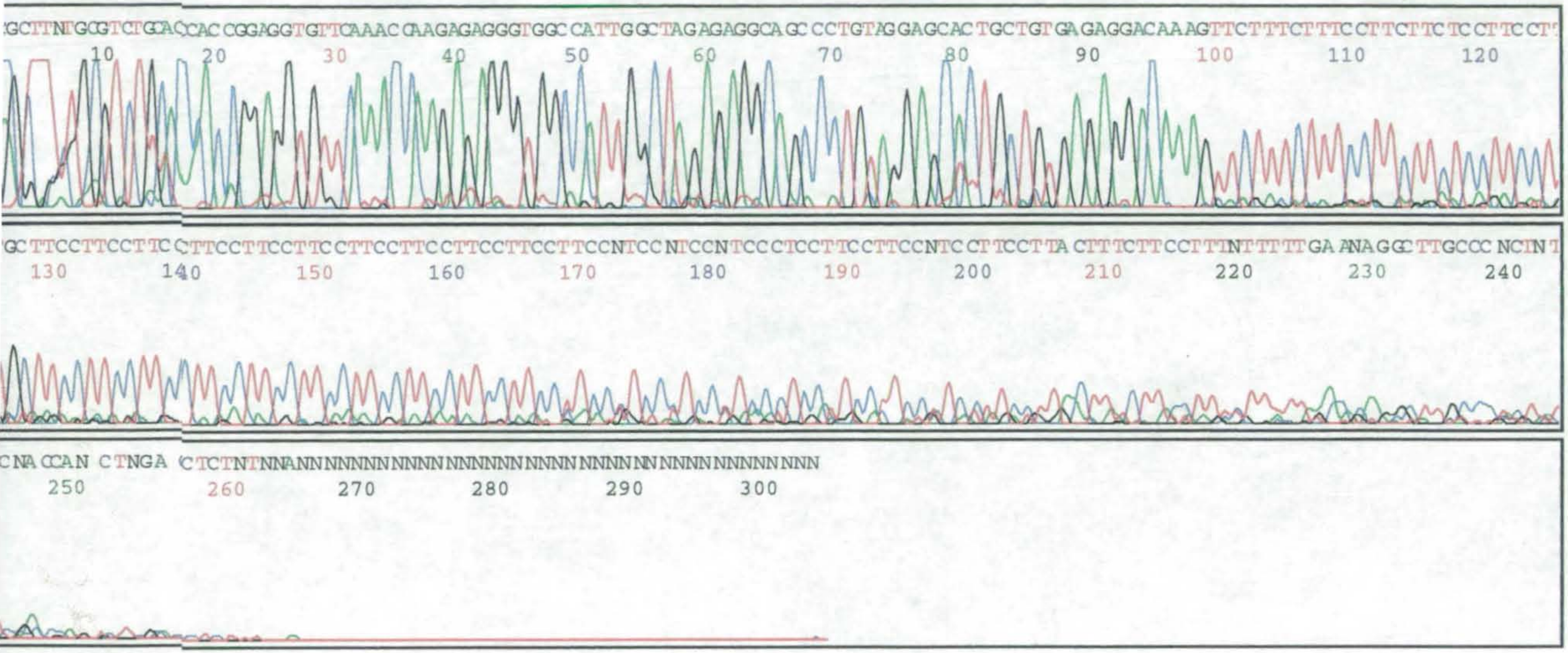
ABI PRISM™

Model 377
Version 3.2
ABI100
Version 3.2

N16074A=9 608
MISS EMMA WATKINS PRIMER 4s 2285
N16074A=9
Lane 3

Signal G:546 A:375 T:406 C:784
DT {BD Set Any-Primer}
dRhod-matrix
Points 1042 to 4000 Pk 1 Loc: 873

Page 1 of 1
Thu, Jan 28, 1999 10:13 AM
Wed, Jan 27, 1999 4:24 PM
Spacing: 9.91{9.91}



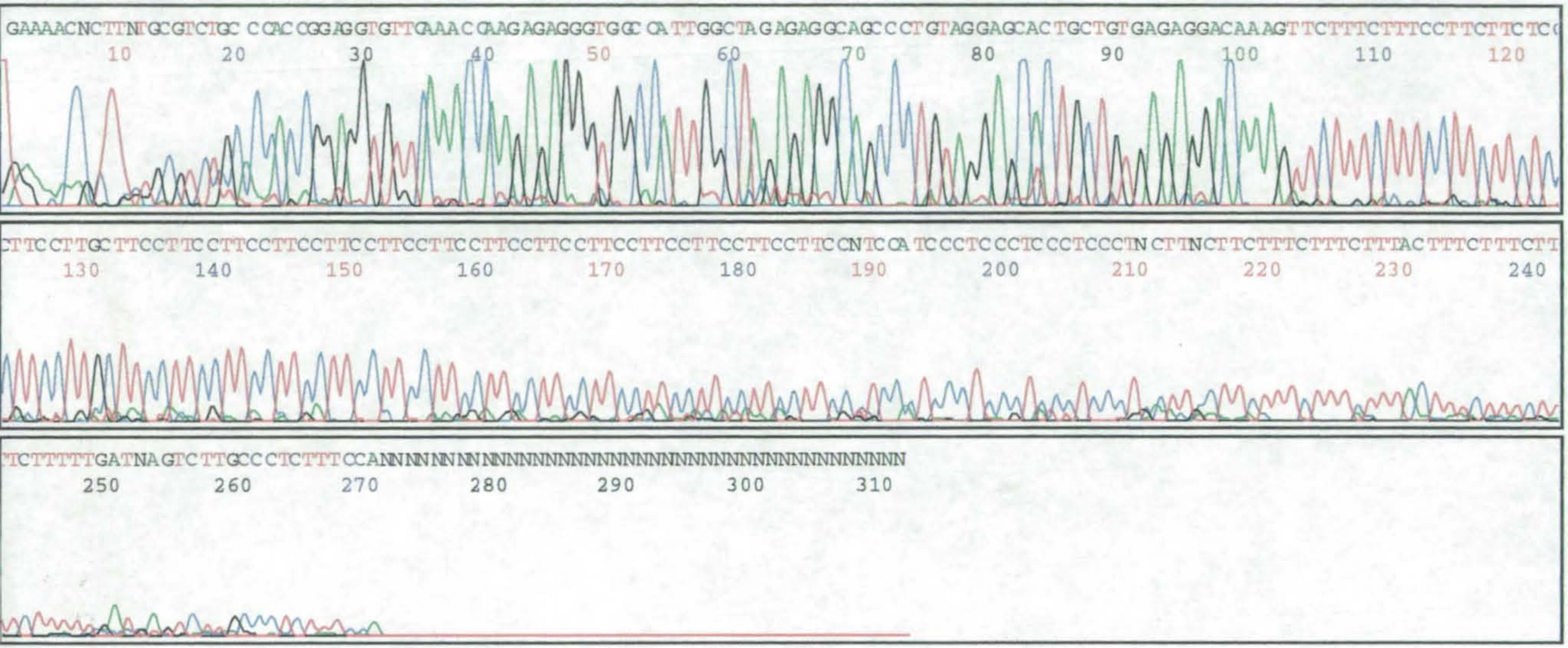
ABI PRISM™

Model 377
Version 3.2
ABI100
Version 3.2

N16074B=10 608
MISS EMMA WATKINS PRIMER 4s 2285
N16074B=10
Lane 9

Signal G:464 A:314 T:378 C:684
DT {BD Set Any-Primer}
dRhod-matrix
Points 1002 to 4000 Pk 1 Loc: 948

Page 1 of 1
Thu, Jan 28, 1999 10:13 AM
Wed, Jan 27, 1999 4:24 PM
Spacing: 9.97{9.97}



```
GAAAACNCTTNTGCGTCTGCCCACCGGAGGTGTTGAAACGAAGAGAGGGTGGCCATTGGCTAGAGAGGCAGCCCTGTAGGAGCACTGCTGTGAGAGGACAAAGTTCTTTCTTTCCTTCTTCTC
CTTCCTTGCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCNTCCATCCCTCCCTCCCTCCCTNCTTNCTTCTTTCTTTCTTTACTTTCTTTCTT
TCTTTTTGATNAGTCTTGCCCTCTTTCCANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

82

ABI PRISM™

Model 377
Version 3.2
ABI100
Version 3.2

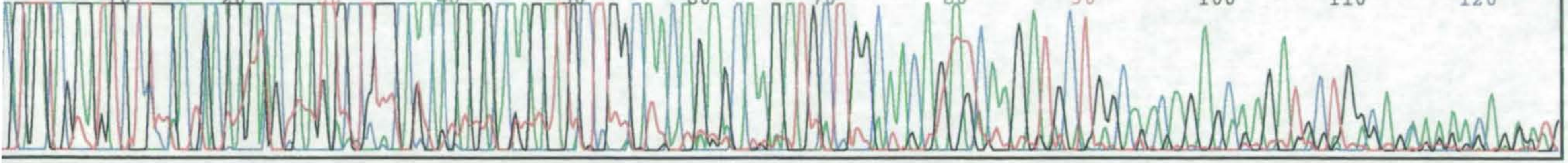
N16073F=6
MISS EMMA WATKINS PRIMER 3s 1514
N16073F=6
Lane 31

Signal G:505 A:530 T:350 C:1055
DT {BD Set Any-Primer}
dRhod-matrix
Points 1042 to 4000 Pk 1 Loc: 872

Page 1 of 1
Thu, Jan 28, 1999 10:12 AM
Wed, Jan 27, 1999 4:24 PM
Spacing: 9.82{9.82}

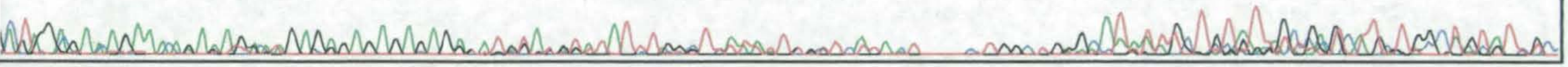
GTTCAGGATGC TGGCAACAGAAGCAAGGATGC TGGCAACAGAGCAAGATGC TGGCAACAGAGCAAGATGC TGGCAACAGATCAAGATGC TGGCAACAGAGCAAGATGC TGGCANNAAAACAGNATG

10 20 30 40 50 60 70 80 90 100 110 120



NTAGAGACAGAAAAGAAAGNC NGAGAGAGAGAGAGANA TGTANNTATATNTGGCTNATATTNNNTCATACTNNNNNTGGC TGNGCATANTNC TGTNTTGCTGNNCNTCTGGCTGTATNGC

130 140 150 160 170 180 190 200 210 220 230 240



GNTCTTNTGNC TGTNTNGC GNTC TTNTGNCNGTATTNC GNTC TTNTGGCNNTNTNTNGNTNNTNTNG

250 260 270 280 290 300 310

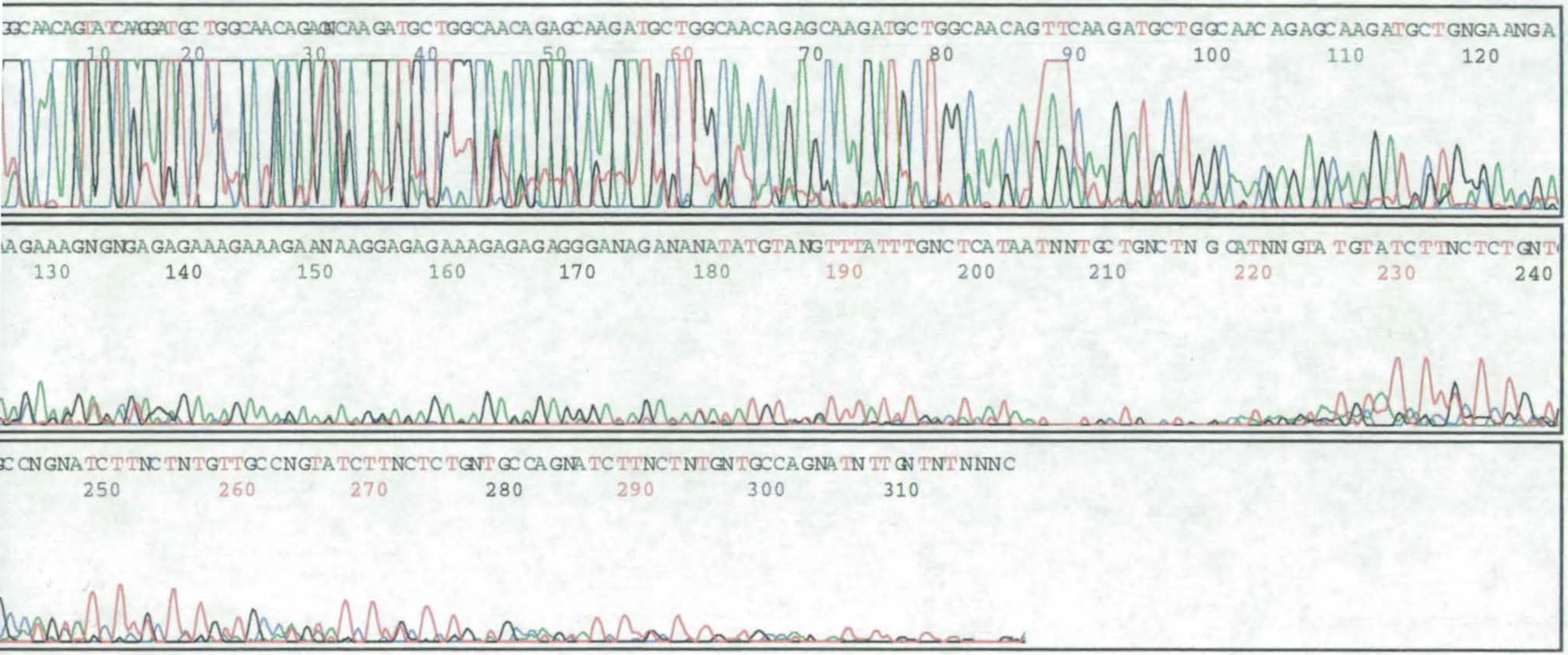
309

Model 377
Version 3.2
ABI100
Version 3.2

N16073G=7 82
MISS EMMA WATKINS PRIMER 3s 1514
N16073G=7
Lane 37

Signal G:553 A:537 T:379 C:1057
DT {BD Set Any-Primer}
dRhod-matrix
Points 1002 to 4000 Pk 1 Loc: 943

Page 1 of 1
Thu, Jan 28, 1999 10:12 AM
Wed, Jan 27, 1999 4:24 PM
Spacing: 9.68{9.68}



518

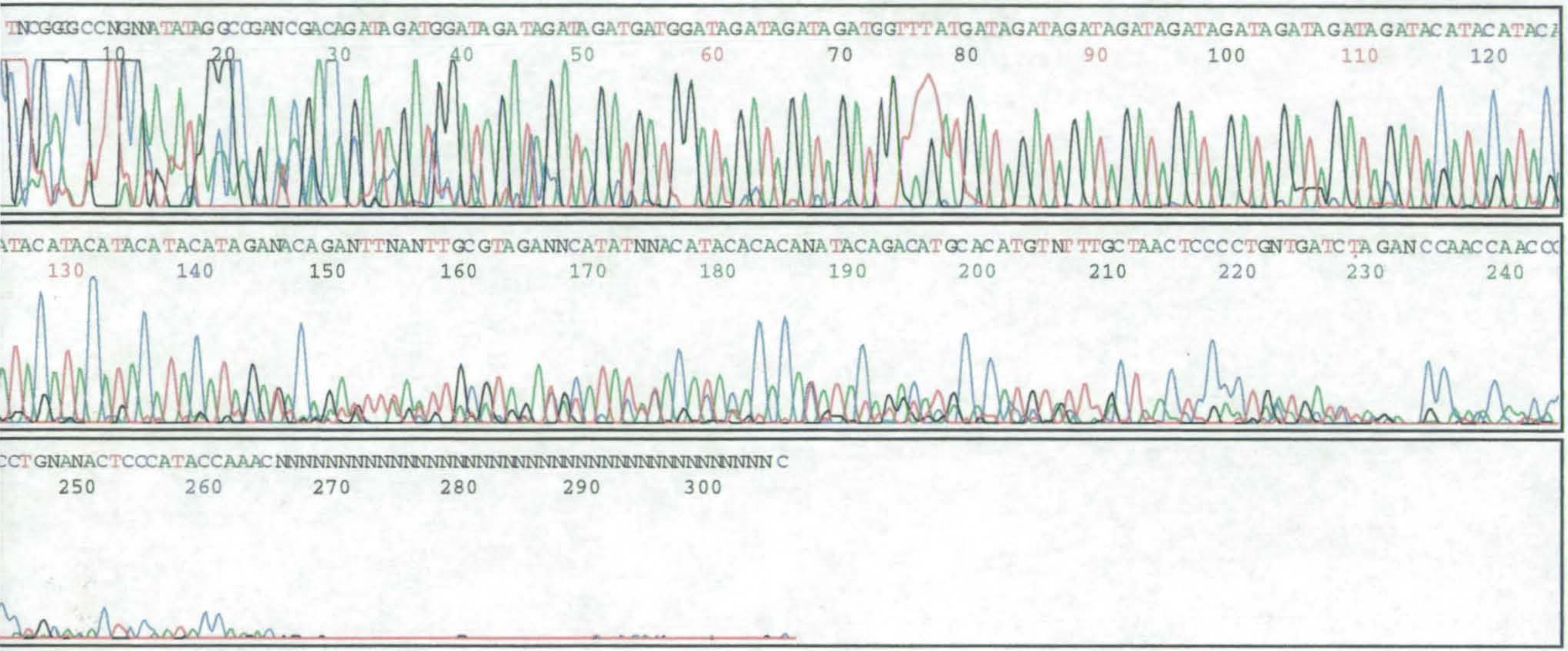
Model 377
Version 3.2
ABI100
Version 3.2

N16073C=3
MISS EMMA WATKINS PRIMER 12s 297
N16073C=3
Lane 13

Signal G:320 A:350 T:318 C:305
DT {BD Set Any-Primer}
dRhod-matrix
Points 1062 to 4000 Pk 1 Loc: 947

Page 1 of 1
Thu, Jan 28, 1999 10:12 AM
Wed, Jan 27, 1999 4:24 PM
Spacing: 10.00{10.00}

```
TNCGGGCCNGNNATATAGGCCGANCGACAGATAGATGGATAGATAGATAGATGATGGATAGATAGATAGATGGTTTATGATAGATAGATAGATAGATAGATAGATAGATAGATACATACATACA
ATACATACATACATACATAGANACAGANTTNANTTGCGTAGANNCATATNNACATACACACANATACAGACATGCACATGTNTTTGCTAACTCCCCTGNTGATCTAGANCCAACCAACCC
CCTGNANACTCCCATACCAAACNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNC
```



518

ABI PRISM

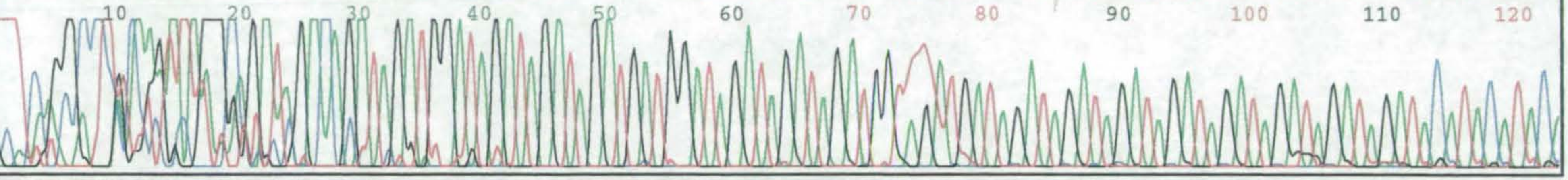
Model 377
Version 3.2
ABI100
Version 3.2

N16073D=4
MISS EMMA WATKINS PRIMER 12s 297
N16073D=4
Lane 19

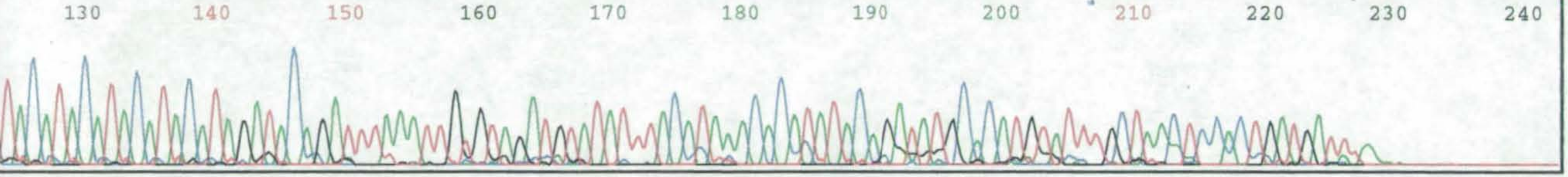
Signal G:510 A:618 T:456 C:445
DT {BD Set Any-Primer}
dRhod-matrix
Points 1062 to 4000 Pk 1 Loc: 870

Page 1 of 1
Thu, Jan 28, 1999 10:12 AM
Wed, Jan 27, 1999 4:24 PM
Spacing: 9.87{9.87}

TTCAGGCCTNCANATAGGCAGATAGACAGATAGATGGATAGATAGATAGATGATGGATAGATAGATAGATGGTNTATGATAGATAGATAGATAGATAGATAGATAGATAGATACATACATACA



TACATACATACATACATAGATACAGATTTAAATTGTGTAGATGTATATTTACATAAACACATATACAGATATGCACATGTATTTGCTAACTCCACTGATGATTTAANNNNNNNNNNNNNN



NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN

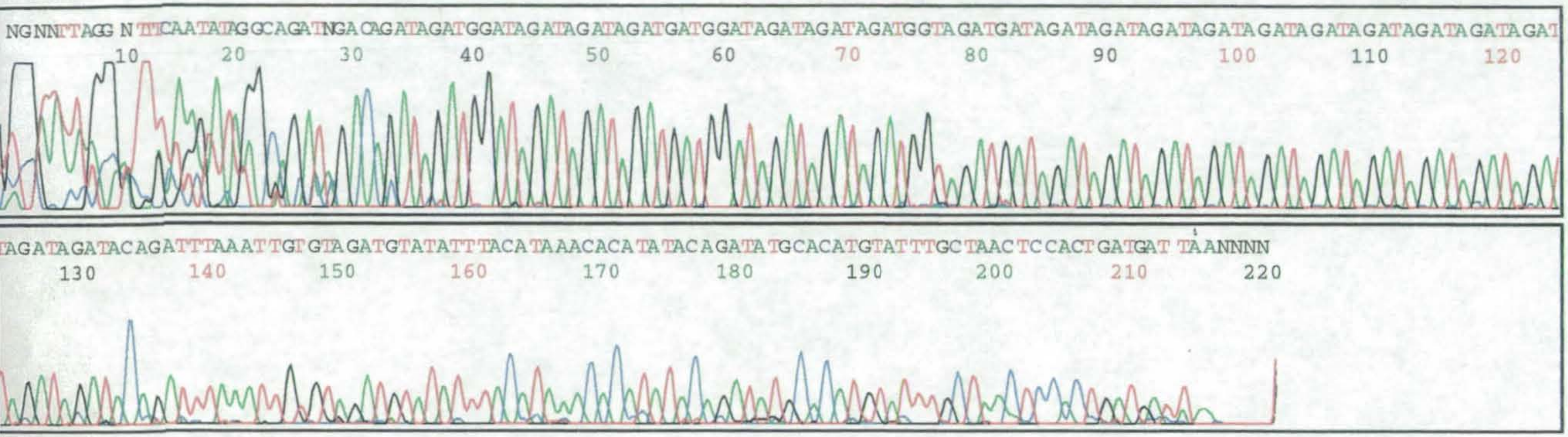
ABI PRISM™

Model 377
Version 3.2
ABI100
Version 3.2

N16073E=5/REP
MISS EMMA WATKINS PRIMER 12s297
N16073E=5/REP
Lane 55

Signal G:2001 A:2108 T:1383 C:952
DT {BD Set Any-Primer}
dRhod-matrix
Points 1042 to 3000 Pk 1 Loc: 808

Page 1 of 1
Mon, Feb 1, 1999 10:39 AM
Fri, Jan 29, 1999 4:34 PM
Spacing: 9.49{9.49}



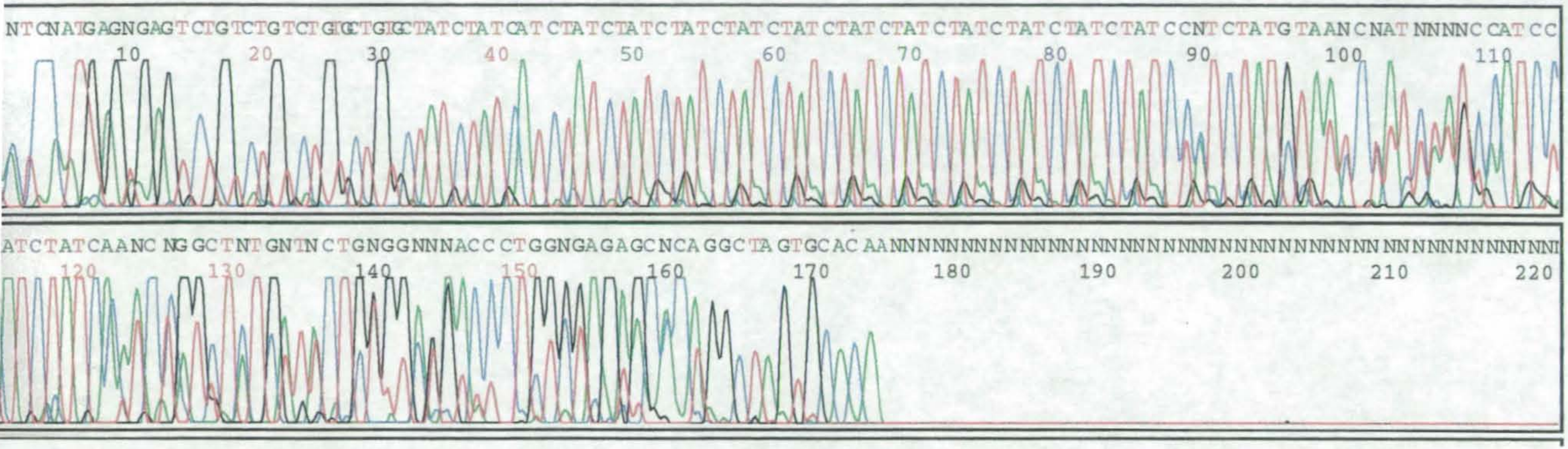
ABI PRISM

Model 377
Version 3.2
ABI100
Version 3.2

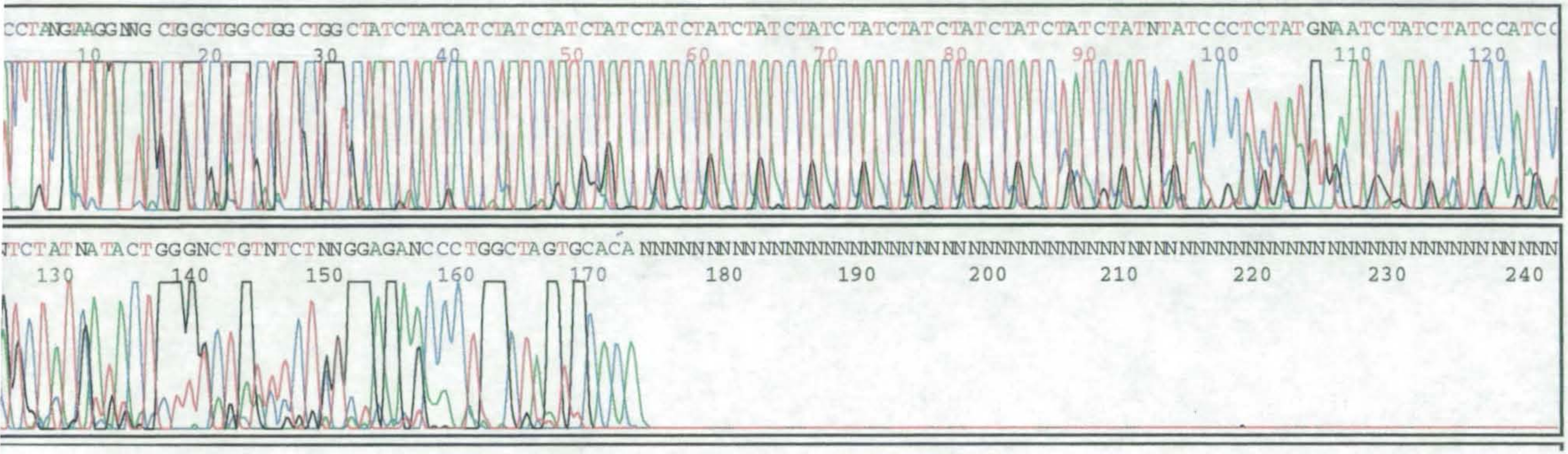
N16075F=22
MISS EMMA WATKINS PRIMER 10s 526
N16075F=22
Lane 35

Signal G:155 A:175 T:303 C:335
DT {BD Set Any-Primer}
dRhod-matrix
Points 1022 to 8260 Pk 1 Loc: 947

Page 1 of 2
Thu, Jan 28, 1999 10:13 AM
Wed, Jan 27, 1999 4:24 PM
Spacing: 9.00{-9.00}



Version 3.2 | MISS EMMA WATKINS PRIMER 10s 526 | DT {BD Set Any-Primer} | Thu, Jan 28, 1999 10:13 AM
ABI100 | N16075E=21 | dRhod-matrix | Wed, Jan 27, 1999 4:24 PM
Version 3.2 | Lane 29 | Points 1022 to 8260 Pk 1 Loc: 828 | Spacing: 9.75{9.75}



ABI PRISM

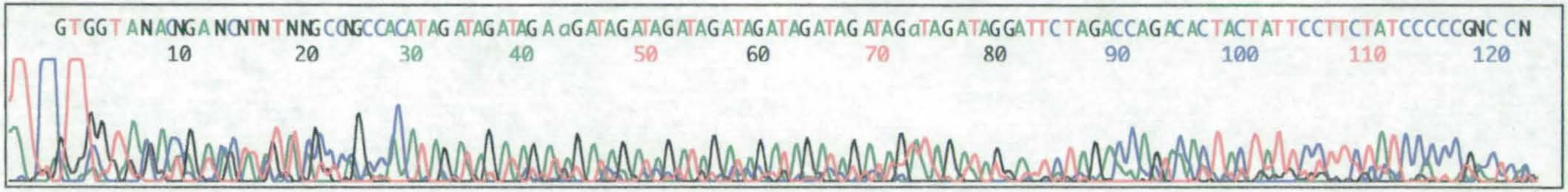
Model 377
Version 3.3
ABI200
Version 3.2

09•112 D1S407F

112 D1S407F
Lane 9

Signal G:42 A:62 T:23 C:42
DT {BD Set Any-Primer}
377 Matrix
Points 1070 to 2190 Pk 1 Loc: 300

Page 1 of 1
Mon, Jun 14, 1999 11:45
Tue, Feb 09, 1999 18:33
Spacing: 9.55{9.55}



316

Model 377
Version 3.3
ABI200
Version 3.2

08•112 D1S407R

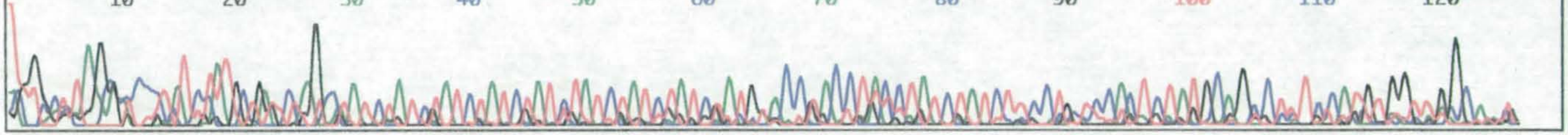
112 D1S407R
Lane 8

Signal G:29 A:53 T:22 C:41
DT {BD Set Any-Primer}
377 Matrix
Points 1150 to 2230   Pk 1 Loc: 1064

Page 1 of 1
Mon, Jun 14, 1999 11:43
Tue, Feb 09, 1999 18:33
Spacing: 10.37{10.37}

NGCCNTAGNNCCTATCTATGCGNCTAGCtATCTATCtATCTATCtATcTATCTATCTATCtATGtACCtACCTaCCTATCTATCTTCCTG TCCACTCTATGCAGCCTTTCCATGTGGTTGGCANGG

10 20 30 40 50 60 70 80 90 100 110 120



ABI PRISM

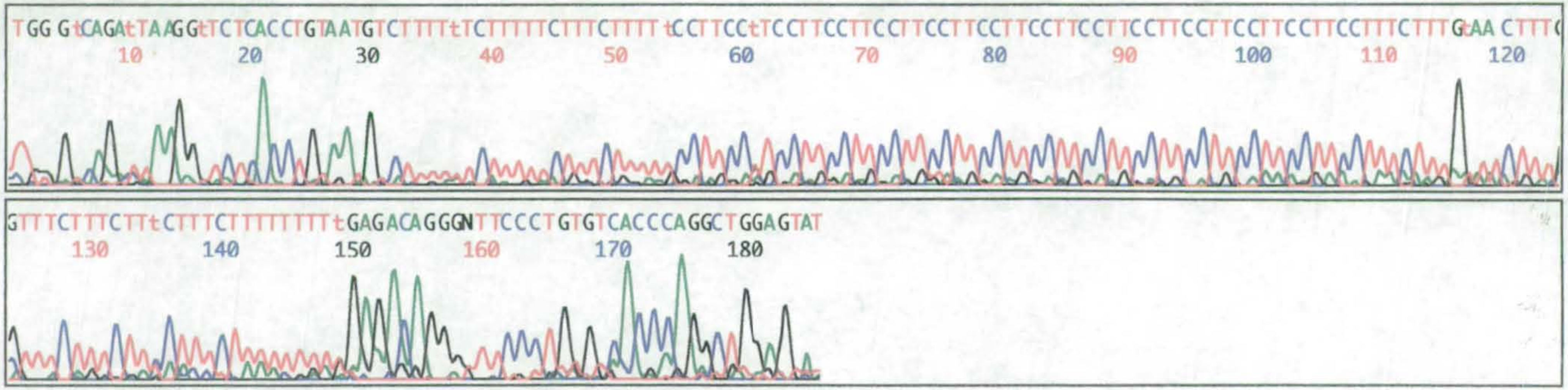
Model 377
Version 3.3
ABI200
Version 3.2

02•489 D7S145R

489 D7S145R
Lane 2

Signal G:41 A:45 T:50 C:71
DT {BD Set Any-Primer}
377 Matrix
Points 1150 to 2850 Pk 1 Loc: 1074

Page 1 of 1
Mon, Jun 14, 1999 11:32
Tue, Feb 09, 1999 18:33
Spacing: 10.02{10.02}



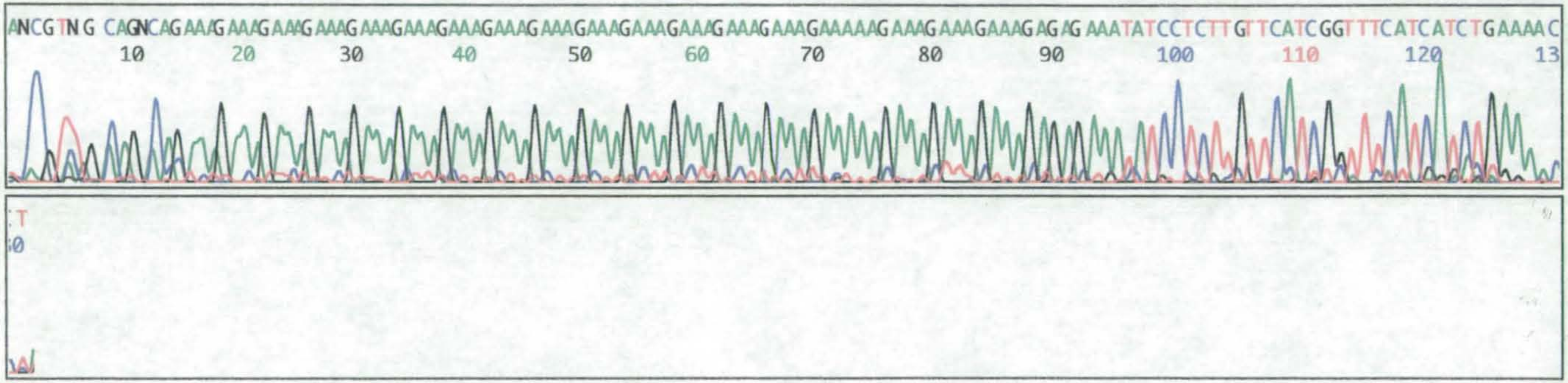
Model 377
Version 3.3
ABI200
Version 3.2

07•62 D7S618F

62 D7S618F
Lane 7

Signal G:152 A:370 T:35 C:71
DT {BD Set Any-Primer}
377 Matrix
Points 1080 to 2190 Pk 1 Loc: 1035

Page 1 of 1
Mon, Jun 14, 1999 11:40
Tue, Feb 09, 1999 18:33
Spacing: 10.28{10.28}



Model 377
Version 3.3
ABI200
Version 3.2

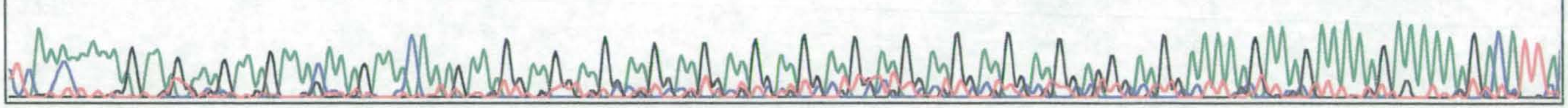
03•489 D7S145F

489 D7S145F
Lane 3

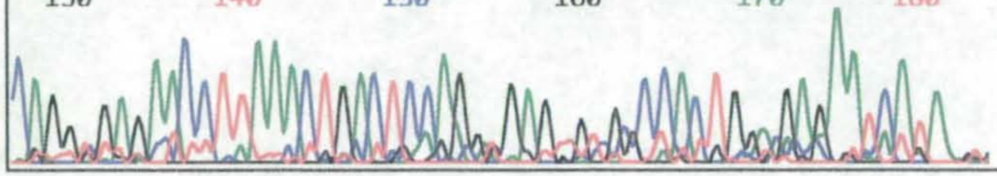
Signal G:89 A:154 T:19 C:44
DT {BD Set Any-Primer}
377 Matrix
Points 1160 to 2800 Pk 1 Loc: 1041

Page 1 of 1
Mon, Jun 14, 1999 11:34
Tue, Feb 09, 1999 18:33
Spacing: 9.88{9.88}

```
TCAANAAAAAGGAANNAAAGAAAGAAACAAAGAAACAAAGAAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAaGGAAGGAAGGAcGGAGGGAAGGAAAAAGAAAGAAAAAGAAAAAAGACATTA
         10     G 20       30        40        50        60        70        80        90        100       110       120
```



```
CAGGTGAGAACCTTAAACTGACTCCAGGTGAGTGCCCCACTGGACGAGAATCATANT
  130       140       150       160       170       180
```



A.7 Autosomal Raw Data

The raw genotype data for each sample population at each locus has been given. Each allele size (in base pairs) is listed and the two alleles separated by a comma. 'NR' indicates No Result, whereby either the PCR process failed or the size of the alleles could not be confidently or accurately determined after repeated gel electrophoresis steps. If the alleles could not be correctly sized they were omitted, so as to preserve accurate and reliable data.

U.K. Leicestershire Population Data

Sample number ID.	**D5S592**	**D2S262**	**D7S1485**	**D4S2285**	**D1S407**	**D7S618**	**D9S252**	**D12S297**	**D10S520**	**D3S1514**
1	NR	207,199	NR	NR	NR	138,134	226,218	265,253	NR	218,206
2	182,174	NR	NR	281,277	152,152	138,138	218,214	261,245	182,178	226,210
3	190,186	215,211	NR	NR	148,144	138,134	218,218	NR	174,166	226,218
5	NR	203,199	216,212	285,285	148,148	142,130	222,214	NR	178,170	218,210
24	186,178	195,195	NR	NR	140,140	134,130	226,214	NR	NR	214,210
38	178,166	199,195	NR	281,269	148,132	130,130	218,214	237,221	182,178	222,222
54	186,174	215,199	NR	281,269	152,148	NR	NR	237,237	182,178	222,230
87	182,174	NR	212,208	NR	148,132	142,122	218,214	NR	190,182	NR
88	186,178	207,203	208,204	285,281	144,144	138,126	222,218	NR	166,162	226,218
90	NR	207,203	204,200	285,269	148,144	138,126	226,218	NR	NR	222,234
91	182,174	199,195	208,200	285,273	148,144	134,130	226,214	NR	182,170	226,222
100	182,178	NR	NR	293,277	NR	134,130	218,214	241,237	170,162	218,214
104	NR	199,191	NR	281,273	148,144	138,130	NR	233,233	182,182	NR
105	NR	203,199	NR	285,281	148,132	138,134	NR	241,205	NR	226,230

U.K. Leicestershire cntd.

Sample number ID.	D5S592	D2S262	D7S1485	D4S2285	D1S407	D7S618	D9S252	D12S297	D10S520	D3S1514
107	174,166	NR	NR	289,269	NR	138,130	NR	233,225	174,170	218,214
109	NR	NR	NR	285,269	148,144	138,138	NR	237,205	182,178	226,210
110	194,186	207,203	208,200	289,281	NR	142,130	NR	NR	178,178	NR
113	186,174	NR	208,200	NR	152,148	NR	NR	241,237	174,174	218,206
115	170,166	191,191	NR	NR	NR	142,142	NR	NR	182,170	218,214
117	178,170	191,187	NR	281,269	148,140	142,134	NR	257,205	182,174	230,230
124	182,170	203,203	NR	289,269	148,148	134,130	NR	241,205	182,174	222,218
125	178,174	NR	NR	289,269	148,132	142,126	NR	NR	186,182	218,230
133	174,166	199,195	NR	285,281	NR	142,122	NR	241,237	170,162	214,210
144	182,174	211,207	NR	NR	148,148	138,138	NR	NR	182,178	218,210
161	178,166	NR	NR	273,269	152,148	138,134	NR	253,245	NR	214,230
162	178,174	NR	NR	NR	152,148	142,134	222,214	245,237	178,174	NR
163	NR	203,199	220,216	285,269	148,140	130,126	222,214	237,233	178,174	226,222
164	186,174	199,187	NR	277,273	NR	142,122	NR	241,233	NR	226,214
165	174,170	203,203	NR	289,281	NR	138,126	222,214	237,233	NR	222,222
167	186,182	211,195	NR	NR	152,148	134,134	226,214	NR	182,182	210,230
170	182,178	NR	NR	293,285	NR	134,126	222,218	245,237	182,170	226,230
172	190,174	NR	224,224	273,269	NR	134,126	222,218	237,205	174,166	NR
173	190,186	195,195	212,208	281,269	148,144	142,134	218,218	249,241	178,178	226,230
174	182,178	207,203	224,220	289,269	148,148	142,138	NR	241,237	NR	226,210
175	186,178	207,207	224,212	281,269	NR	NR	222,218	NR	NR	226,214

U.K. Leicestershire cntd.

Sample number ID.	D5S592	D2S262	D7S1485	D4S2285	D1S407	D7S618	D9S252	D12S297	D10S520	D3S1514
177	186,178	203,195	212,212	281,269	144,140	NR	230,214	237,205	182,182	226,234
180	190,174	NR	220,212	285,285	NR	NR	222,214	265,233	190,182	NR
182	182,178	NR	NR	285,281	152,148	138,138	NR	237,233	NR	214,206
183	170,162	195,183	NR	277,273	148,144	138,126	NR	245,241	186,178	210,222
184	186,178	NR	NR	273,269	144,144	142,134	226,218	NR	190,178	226,218
187	NR	NR	216,204	NR	NR	NR	222,218	NR	NR	226,214
188	178,174	195,195	NR	273,269	NR	142,138	226,226	NR	178,178	218,222
189	186,178	203,203	216,216	NR	156,156	138,138	222,218	237,213	178,174	218,214
190	194,182	203,195	216,204	289,269	152,148	142,126	218,218	237,233	194,178	234,234
191	186,182	207,195	220,212	301,289	156,156	142,130	NR	241,241	174,170	226,230
192	186,182	199,199	NR	289,269	156,156	138,138	222,218	233,229	178,174	214,230
193	186,174	NR	NR	285,277	160,148	142,130	234,226	237,237	186,170	218,238
194	190,178	NR	204,200	289,289	156,156	138,126	218,214	241,229	186,182	226,226
195	186,182	NR	208,200	293,277	148,136	138,134	226,222	237,209	178,170	226,218
196	186,182	207,207	216,200	273,273	156,156	138,134	222,218	241,237	186,186	218,234
197	198,182	215,203	NR	269,269	160,160	126,126	226,226	257,241	186,182	234,222
198	186,182	NR	NR	301,269	156,156	134,134	230,222	241,233	182,174	230,222
199	186,182	NR	NR	289,289	156,136	138,130	226,226	241,233	190,178	234,230
200	194,182	NR	NR	289,289	152,152	138,122	230,218	245,241	178,166	234,222
201	194,190	NR	NR	289,269	156,148	142,138	226,218	NR	178,170	226,222
202	174,174	NR	NR	289,273	156,152	146,142	234,226	237,237	178,174	210,230

U.K. Leicestershire cntd.

Sample number ID.	D5S592	D2S262	D7S1485	D4S2285	D1S407	D7S618	D9S252	D12S297	D10S520	D3S1514
203	NR	NR	NR	301,269	164,156	126,126	NR	237,233	186,158	222,222
204	190,178	NR	NR	301,301	164,156	142,142	NR	NR	182,182	218,214
205	182,178	NR	NR	297,297	160,136	NR	NR	241,233	178,170	226,218
206	194,194	NR	NR	293,273	156,152	NR	NR	261,237	182,174	230,222

Admixed Islanders

Sample ID.	D5S592	D7S1485	D4S2285	D2S262	D1S407	D7S618	D12S297	D3S1514	D10S520	D9S252
98	174,170	212,208	273,269	207,207	NR	142,134	241,233	230,218	186,166	226,218
102	174,166	NR	293,269	215,203	148,144	134,134	261,205	226,226	178,174	226,222
104	178,174	212,204	NR	211,203	152,148	134,134	245,205	234,210	174,166	226,218
107	178,170	NR	289,273	203,191	156,148	134,122	237,205	222,218	182,170	222,218
108	186,182	212,208	293,273	207,203	152,148	142,138	257,229	230,222	182,182	218,218
113	178,174	NR	285,269	203,191	156,152	138,138	237,237	NR	NR	NR
114	NR	220,216	NR	199,199	NR	138,138	257,237	226,218	186,178	226,218
115	182,178	212,208	NR	199,195	160,152	134,134	237,233	NR	186,178	226,218
116	NR	216,212	NR	215,211	NR	142,130	245,237	NR	174,166	NR
601	178,174	208,208	NR	NR	NR	142,134	249,237	230,218	NR	222,214
605	182,174	208,200	NR	199,195	NR	130,130	245,205	226,222	NR	230,226
612	174,166	216,200	301,269	211,211	NR	138,130	265,205	226,214	178,178	214,210
614	190,178	216,208	NR	NR	144,144	134,130	265,245	230,226	174,174	214,210
617	178,178	216,212	269,269	203,203	152,144	146,134	265,241	230,226	178,174	222,218
621	NR	216,212	NR	199,195	144,144	142,138	NR	NR	174,170	210,210
625	182,170	216,208	285,277	NR	152,144	NR	261,257	222,218	182,170	222,218

Non-admixed Maori

Sample Number ID.	D5S592	D7S1485	D4S2285	D2S262	D1S407	D7S618	D12S297	D3S1514	D10S520	D9S252
66	182,174	212,212	185,185	203,203	156,148	134,134	265,245	178,170	222,234	226,222
74	190,174	212,212	297,269	207,203	148,148	142,134	NR	186,174	218,226	222,222
75	178,174	NR	297,293	203,199	152,152	142,138	NR	174,170	214,214	222,218
76	186,178	216,212	289,269	NR	164,152	138,134	261,205	NR	218,234	226,218
90	178,174	216,208	285,269	203,203	148,148	142,134	257,245	174,174	226,210	226,226
401	186,174	216,208	269,269	211,207	152,144	138,130	245,229	186,186	NR	230,222
404	182,178	220,216	281,269	219,199	156,152	138,138	269,265	190,182	NR	226,222
406	190,178	216,212	305,289	NR	144,144	142,130	269,257	NR	NR	222,218
410	NR	NR	285,281	211,203	148,144	134,130	265,253	186,174	NR	222,218
411	NR	212,208	?,?	NR	152,148	134,130	249,241	NR	NR	218,218
418	178,170	216,216	281,269	203,195	156,152	NR	249,237	186,178	222,214	222,214
420	NR	224,212	289,269	199,199	152,148	138,122	253,205	186,166	218,218	222,210
421	NR	220,212	NR	203,203	148,144	134,134	249,245	186,182	226,214	222,218
422	186,182	220,212	301,293	203,203	NR	NR	NR	182,174	222,226	226,214
426	186,174	220,208	269,269	NR	156,152	146,122	245,241	170,166	218,226	218,214
427	NR	216,216	NR	211,207	NR	142,130	257,253	186,178	234,226	218,214
429	178,174	216,208	305,289	207,199	NR	138,138	245,205	174,174	222,226	226,218
432	186,182	212,208	293,269	207,199	152,148	138,134	261,257	178,174	NR	218,214
433	182,174	216,216	293,269	199,199	152,148	142,130	249,205	NR	206,206	230,226

Non-admixed Maori contd.

Sample Number ID.	D5S592	D7S1485	D4S2285	D2S262	D1S407	D7S618	D12S297	D3S1514	D10S520	D9S252
436	186,178	NR	297,289	NR	NR	NR	NR	NR	NR	218,214
437	178,174	220,216	293,285	215,199	156,148	134,130	257,253	NR	NR	218,214
439	174,174	220,216	NR	207,199	144,144	138,130	253,253	NR	NR	226,222
440	190,178	216,212	NR	NR	152,152	134,134	NR	194,178	234,210	226,222
441	174,174	208,204	NR	211,207	148,144	142,134	257,205	186,174	NR	222,218
448	182,174	212,208	289,269	NR	152,148	138,134	237,237	178,178	NR	218,214
449	NR	220,212	NR	207,199	148,144	NR	269,241	182,178	230,226	222,210
450	174,162	204,200	269,269	207,203	152,148	142,122	269,261	NR	NR	218,214
451	174,166	224,216	293,269	219,207	156,148	134,130	257,241	178,170	234,206	218,214
452	174,170	212,208	289,269	NR	144,144	138,134	237,233	182,182	218,234	222,218
453	190,174	220,208	277,269	207,203	144,144	134,134	257,233	186,178	218,230	214,214
454	NR	220,216	NR	207,203	152,148	142,130	205,205	186,186	NR	214,214
455	NR	212,208	285,277	203,203	152,148	138,134	257,245	182,174	214,210	222,218
456	174,174	216,212	293,269	207,203	152,152	134,134	261,241	182,178	NR	218,218
457	NR	212,204	273,269	NR	144,140	?,130	257,249	182,178	218,226	214,214
462	NR	212,208	293,285	NR	152,152	?,130	NR	182,174	226,210	214,210
463	182,174	216,208	269,269	203,195	148,144	NR	249,205	186,174	218,230	226,222
464	178,174	216,208	301,293	203,203	152,148	NR	245,241	NR	218,210	222,218
466	NR	216,208	293,285	NR	152,148	138,130	NR	178,170	230,210	214,214
468	174,170	216,208	289,269	NR	144,144	134,134	NR	178,170	218,210	222,214

Non-admixed Maori cntd.

Sample Number ID.	D5S592	D7S1485	D4S2285	D2S262	D1S407	D7S618	D12S297	D3S1514	D10S520	D9S252
469	182,178	216,208	293,277	NR	152,144	134,134	NR	182,174	218,226	226,218
470	178,174	216,208	269,269	207,199	NR	142,138	261,257	178,178	226,202	214,214
474	182,174	220,208	297,269	NR	152,152	130,122	NR	190,182	218,226	226,222
476	178,174	216,212	305,301	NR	NR	134,134	NR	178,170	210,210	222,222
477	178,174	208,208	285,269	207,203	NR	134,134	NR	182,174	218,234	226,218
479	174,170	208,208	?,?	207,199	NR	138,134	253,249	182,174	234,230	214,214
480	182,174	216,212	293,289	199,199	NR	142,134	261,257	182,178	230,226	226,218
484	174,170	220,216	293,269	NR	148,144	142,134	257,205	178,174	218,218	222,218
485	186,178	220,212	?,?	207,199	152,148	134,130	261,205	NR	NR	218,214
488	186,182	NR	269,265	207,207	NR	142,130	NR	178,178	222,210	NR
500	174,170	220,208	285,269	207,203	NR	134,134	261,257	182,170	218,210	NR
501	178,174	220,212	289,269	203,199	NR	134,134	257,205	186,182	NR	210,210
502	182,174	224,216	293,281	NR	144,144	142,134	261,257	174,174	230,206	218,214
503	194,174	216,212	289,269	215,203	NR	134,130	257,245	194,170	222,230	222,214
504	190,182	216,212	289,265	207,199	NR	138,134	257,245	182,174	222,226	226,222
505	182,174	216,216	281,269	207,199	NR	138,134	NR	186,178	222,210	218,214
506	182,174	216,208	285,269	207,207	NR	134,130	257,249	186,178	214,214	NR
507	174,166	228,208	289,269	203,199	NR	134,130	NR	186,170	NR	218,214
509	NR	216,216	289,269	199,191	NR	138,134	NR	182,174	NR	218,214
510	NR	212,212	289,281	203,199	NR	134,130	NR	182,174	NR	230,218

Non-admixed Maori cntd.

Sample Number ID.	D5S592	D7S1485	D4S2285	D2S262	D1S407	D7S618	D12S297	D3S1514	D10S520	D9S252
511	182,178	216,212	297,269	207,199	NR	130,130	NR	182,178	214,210	214,210
512	178,178	216,208	293,269	203,199	NR	134,134	245,241	178,174	NR	NR
514	186,170	216,208	285,281	211,199	NR	134,130	257,249	182,174	226,210	222,218
516	178,174	212,208	277,269	207,199	NR	142,138	261,249	186,174	222,226	218,214
517	170,170	224,208	289,281	203,203	NR	NR	NR	182,178	218,210	222,218
520	NR	NR	289,269	NR	NR	134,134	261,257	186,186	222,218	226,226
521	NR	224,208	277,269	207,203	NR	142,142	265,241	178,178	222,234	226,214
523	NR	216,208	289,269	207,207	NR	134,130	261,237	174,174	226,210	NR
524	NR	208,208	273,269	199,199	NR	134,134	249,241	186,178	230,226	NR
525	NR	216,208	297,293	NR	NR	134,134	261,205	174,174	222,234	NR
527	170,170	212,208	NR	211,207	NR	NR	265,261	178,170	226,210	222,218
529	178,174	212,208	NR	211,203	NR	142,134	261,245	182,170	222,218	NR
530	NR	220,216	285,281	207,199	NR	134,134	241,205	174,170	218,210	222,218

Admixed Maori raw data

Sample Number ID	D7S1485	D5S592	D4S2285	D2S262	D1S407	D7S618	D12S297	D3S1514	D10S520	D9S252
59	182,178	NR	297,269	NR	152,148	134,134	237,237	NR	NR	226,222
60	NR	220,216	289,285	207,199	152,152	142,126	NR	234,226	182,174	218,214
61	178,174	220,208	293,289	211,203	NR	138,134	205,205	226,226	178,174	218,214
63	182,178	216,216	289,285	207,199	NR	138,138	NR	226,214	186,178	226,226
64	178,166	216,212	285,273	207,203	NR	142,138	249,249	222,222	182,174	218,218
65	NR	NR	293,269	207,207	152,148	142,130	NR	222,226	186,182	230,218
67	174,174	NR	293,273	203,203	NR	134,130	NR	218,234	182,170	230,218
72	190,182	216,208	293,269	203,203	148,148	NR	269,257	234,230	186,178	226,214
73	178,178	216,208	285,269	211,199	152,144	142,138	261,237	218,234	178,174	222,218
77	178,174	NR	297,269	211,203	148,144	138,130	NR	218,226	182,174	230,218
78	182,178	208,200	297,289	NR	NR	134,126	237,233	226,210	178,178	214,214
79	190,186	208,208	277,269	199,199	NR	142,126	NR	218,206	182,170	226,222
82	182,178	208,200	297,293	NR	144,144	142,138	245,245	238,226	186,174	230,214
83	178,174	224,220	281,269	207,203	152,148	134,126	261,205	222,210	NR	218,218
84	182,174	220,204	281,269	219,207	152,148	134,134	NR	222,226	186,178	222,218
86	182,178	212,208	297,285	207,203	160,152	142,138	205,205	NR	194,178	222,218
88	186,174	220,216	281,273	NR	148,148	138,130	241,233	222,222	182,174	226,222
91	174,166	212,208	293,269	NR	NR	NR	261,245	222,218	190,170	222,214
92	190,178	216,208	289,285	207,199	152,148	142,134	233,233	NR	182,162	222,222

Admixed Maori raw data cntd.

Sample Number ID	D7S1485	D5S592	D4S2285	D2S262	D1S407	D7S618	D12S297	D3S1514	D10S520	D9S252
409	186,178	216,212	NR	215,207	156,152	142,134	249,233	218,214	182,178	214,214
412	182,174	NR	305,285	215,199	NR	134,130	253,205	NR	186,170	222,210
413	182,174	216,216	297,273	NR	152,152	138,126	237,225	NR	182,174	NR
414	182,170	216,212	285,277	203,199	156,148	NR	253,245	222,210	182,182	218,218
417	182,174	224,216	281,273	NR	148,148	138,130	237,233	222,226	174,174	226,218
419	NR	216,216	301,293	215,207	152,152	134,134	269,265	218,234	174,174	226,222
423	186,178	220,212	269,269	NR	156,152	134,134	253,205	222,210	186,170	218,214
424	178,174	212,212	293,269	203,203	148,148	142,134	249,237	218,226	178,174	218,214
425	186,174	216,208	293,289	203,187	156,152	138,134	249,205	226,210	182,174	226,222
428	174,170	220,212	289,281	199,199	156,148	138,130	NR	NR	178,174	222,214
434	178,170	NR	289,269	NR	NR	NR	NR	NR	NR	218,214
435	186,174	216,208	297,269	211,203	160,148	134,130	261,257	NR	178,174	222,218
442	186,174	220,216	NR	207,207	148,148	134,130	253,245	NR	186,174	214,214
443	182,170	212,204	305,269	207,203	144,144	134,134	257,241	NR	178,178	222,214
444	190,174	216,212	289,285	203,203	148,140	146,138	261,245	NR	186,174	222,214
446	NR	208,200	281,269	NR	156,156	134,130	233,205	NR	190,178	222,214
458	178,174	224,208	NR	211,203	152,152	138,138	261,241	NR	178,174	218,214
459	NR	220,216	293,289	207,203	144,140	138,134	261,205	234,226	174,170	218,218
460	174,170	212,212	293,277	207,207	152,148	NR	245,241	226,214	174,174	226,222
465	182,178	212,212	305,289	211,195	148,140	613,134	NR	226,206	NR	214,214
467	186,178	216,212	269,269	NR	152,148	134,130	245,205	226,214	NR	214,210

Admixed Maori raw data cntd.

Sample Number ID	D7S1485	D5S592	D4S2285	D2S262	D1S407	D7S618	D12S297	D3S1514	D10S520	D9S252
471	178,174	212,208	289,285	207,207	NR	134,134	NR	226,214	186,170	222,222
472	178,174	220,204	297,269	NR	NR	142,134	NR	222,230	178,174	230,218
473	174,170	220,220	269,269	NR	144,144	142,134	NR	NR	178,170	230,230
475	NR	220,208	297,273	203,195	NR	134,130	NR	222,230	178,178	226,222
478	182,170	208,204	285,281	NR	NR	134,130	NR	218,218	178,174	222,218
481	174,166	NR	293,289	NR	NR	NR	NR	222,222	NR	230,218
482	182,174	212,208	297,285	203,199	NR	146,138	NR	218,230	178,178	222,222
487	186,170	216,208	293,289	NR	NR	138,130	NR	222,214	NR	222,214
489	190,170	216,208	293,273	199,199	NR	142,134	257,205	222,222	186,178	230,218
508	178,174	216,208	281,269	NR	NR	134,130	NR	222,218	186,174	214,210
513	174,170	NR	277,269	203,203	NR	138,126	249,249	NR	178,174	222,218
515	186,178	216,216	297,281	207,195	NR	142,122	257,253	222,230	182,178	222,218
518	NR	216,208	289,269	203,199	NR	142,138	265,245	222,214	178,170	218,218
528	174,170	216,212	269,269	199,199	NR	146,142	269,205	218,226	178,174	NR
533	170,166	212,212	293,269	211,203	NR	138,130	241,237	NR	NR	NR
600	NR	212,208	297,269	NR	NR	134,130	249,241	226,226	NR	NR

A.8 Genetic distance measures

Rst and Delta mu squared ($\delta\mu^2$), distance matrices of the sample populations across all loci.

Rst distance matrix is situated below diagonal and the $\delta\mu^2$ distance above diagonal.

	NAI	AM	NAM	U.K. Leics.	AI
NAI		1.787	1.556	1.660	1.533
AM	0.208		0.137	0.061	0.050
NAM	0.183	0.003		0.210	-0.112
U.K. Leics.	0.179	0.012	0.033		0.075
AI	0.149	0.005	-0.006	0.016	

A.9 Y chromosome haplotypic raw data: The population-group letters, P, M and C refer to Polynesian Islanders, New Zealand Maori and U.K. Leicestershire caucasians, respectively. Loci, repeat numbers, microsatellite and UEP (haplogroup) codes are given for each Y chromosome.

Sample_ID	Pop_group	Pop_subgroup		_YAP	92R7	sY81	SRY465	SRY4064	TAT	M9	M13	M17	M20	SRY10831	DYS19	DYS388	DYS390	DYS391	DYS392	DYS393	MS_Code	UEP_Code
163	C		U.K. Leicestershire	NEGATIVE	T	A	C	G	T	G	G	G+	A	G	13	12	24	10	13	13	IHTFII	hg1
162	C		U.K. Leicestershire	NEGATIVE	T	A	C	G	T	G	G	G+	A	G	14	12	23	11	13	13	JHSGII	hg1
173	C		U.K. Leicestershire	NEGATIVE	T	A	C	G	T	G	G	G+	A	G	14	12	23	11	13	13	JHSGII	hg1
177	C		U.K. Leicestershire	NEGATIVE	T	A	C	G	T	G	G	G+	A	G	14	12	23	11	13	13	JHSGII	hg1
164	C		U.K. Leicestershire	NEGATIVE	T	A	C	G	T	G	G	G+	A	G	14	12	23	11	14	13	JHSGJI	hg1
124	C		U.K. Leicestershire	NEGATIVE	T	A	C	G	T	G	G	G+	A	G	14	12	24	10	13	13	JHTFII	hg1
182	C		U.K. Leicestershire	NEGATIVE	T	A	C	G	T	G	G	G+	A	G	14	12	24	10	13	13	JHTFII	hg1
99	C		U.K. Leicestershire	NEGATIVE	T	A	C	G	T	G	G	G+	A	G	14	12	24	11	13	13	JHTGII	hg1
188	C		U.K. Leicestershire	NEGATIVE	T	A	C	G	T	G	G	G+	A	G	14	12	24	11	13	13	JHTGII	hg1
115	C		U.K. Leicestershire	NEGATIVE	T	A	C	G	T	G	G	G+	A	G	14	12	24	11	13	14	JHTGIJ	hg1
172	C		U.K. Leicestershire	NEGATIVE	T	A	C	G	T	G	G	G+	A	G	14	12	24	12	13	13	JHTHII	hg1
186	C		U.K. Leicestershire	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	14	14	22	10	11	13	JJRFGI	hg2
133	C		U.K. Leicestershire	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	14	14	23	10	11	14	JJSFGJ	hg2
165	C		U.K. Leicestershire	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	15	14	23	10	11	13	KJSFGI	hg2
167	C		U.K. Leicestershire	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	15	14	24	10	11	13	KJTFGI	hg2
117	C		U.K. Leicestershire	POSITIVE	C	A	C	A	T	C	G	G+	A	G	13	12	24	10	11	13	IHTFGI	hg21
175	C		U.K. Leicestershire	POSITIVE	C	A	C	A	T	C	G	G+	A	G	13	12	24	10	11	13	IHTFGI	hg21
183	C		U.K. Leicestershire	POSITIVE	C	A	C	A	T	C	G	G+	A	G	13	12	24	10	11	13	IHTFGI	hg21

Sample_ID	Pop_group	Pop_subgroup	_YAP	92R7	sY81	SRY465	SRY4064	TAT	M9	M13	M17	M20	SRY10831	DYS19	DYS388	DYS390	DYS391	DYS392	DYS393	MS_Code	UEP_Code
90	C	U.K. Leicestershire	POSITIVE	C	A	C	A	T	C	G	G+	A	G	14	12	24	10	11	13	JHTFGI	hg21
87	C	U.K. Leicestershire	NEGATIVE	C	A	C	G	T	G	G	G+	A	G	14	12	23	10	13	13	JHSFII	hg26
108	C	U.K. Leicestershire	NEGATIVE	C	A	C	G	T	G	G	G+	A	G	14	12	24	11	14	13	JHTGJI	hg26
54	C	U.K. Leicestershire	NEGATIVE	T	A	C	G	T	G	G	G-	A	A	15	12	25	11	11	13	KHUGGI	hg3
404	M	100% Native	NEGATIVE	T	A	C	G	T	G	G	G+	A	G	13	12	22	10	15	13	IHRFKI	hg1
488	M	100% Native	NEGATIVE	T	A	C	G	T	G	G	G+	A	G	14	12	23	10	14	12	JHSFJH	hg1
65	M	62.5% MAORI/EURO	NEGATIVE	T	A	C	G	T	G	G	G+	A	G	14	12	23	11	13	13	JHSGII	hg1
91	M	40%MAORI/SCOT/FREN	NEGATIVE	T	A	C	G	T	G	G	G+	A	G	14	12	23	11	13	14	JHSGIJ	hg1
460	M	50%	NEGATIVE	T	A	C	G	T	G	G	G+	A	G	14	12	23	12	13	13	JHSHII	hg1
420	M	100% Native	NEGATIVE	T	A	C	G	T	G	G	G+	A	G	14	12	24	10	13	13	JHTFII	hg1
429	M	100% Native	NEGATIVE	T	A	C	G	T	G	G	G+	A	G	14	12	24	10	13	13	JHTFII	hg1
413	M	25%MAORI/EURO	NEGATIVE	T	A	C	G	T	G	G	G+	A	G	14	12	24	11	13	13	JHTGII	hg1
433	M	100% Native	NEGATIVE	T	A	C	G	T	G	G	G+	A	G	14	12	24	11	13	13	JHTGII	hg1
444	M	75% MAORI/EURO	NEGATIVE	T	A	C	G	T	G	G	G+	A	G	14	12	24	11	13	13	JHTGII	hg1
503	M	100% Native	NEGATIVE	T	A	C	G	T	G	G	G+	A	G	14	12	24	11	13	13	JHTGII	hg1
533	M	50% MAORI/EURO	NEGATIVE	T	A	C	G	T	G	G	G+	A	G	14	15	24	11	13	13	JKTGII	hg1
79	M	25% MAORI/EURO	NEGATIVE	T	A	C	G	T	G	G	G+	A	G	15	12	24	10	13	14	KHTFIJ	hg1
77	M	50% MAORI/DUTCH/GERM	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	14	14	23	10	11	13	JJSFGI	hg2
509	M	100% Native	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	14	14	23	10	11	13	JJSFGI	hg2
510	M	100% Native	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	14	15	23	10	11	13	JKSFGI	hg2
88	M	75%MAORI/EURO	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	14	16	22	10	11	13	JLRFGI	hg2
474	M	100% Native	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	15	13	22	10	11	14	KIRFGJ	hg2

Sample_ID	Pop_group	Pop_subgroup	_YAP	92R7	sY81	SRY465	SRY4064	TAT	M9	M13	M17	M20	SRY10831	DYS19_ASize	DYS388_ASize	DYS390_ASize	DYS391_ASize	DYS392_ASize	DYS393_ASize	MS_Code	UEP_Code
472	M	50%MAORI/EURO	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	15	14	23	10	11	13	KJSFGI	hg2
514	M	100% Native	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	15	15	20	10	12	14	KKPFHJ	hg2
471	M	50%MAORI/ENG	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	15	15	20	11	12	14	KKPGHJ	hg2
457	M	100% Native	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	15	16	22	10	11	13	KLRFGI	hg2
66	M	100% Native	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	16	15	20	10	12	14	LKPFHJ	hg2
72	M	62.5% MAORI/EURO	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	16	15	20	10	12	14	LKPFHJ	hg2
75	M	100% Native	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	16	15	20	10	12	14	LKPFHJ	hg2
84	M	25%MAORI/EURO	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	16	15	20	10	12	14	LKPFHJ	hg2
86	M	50%MAORI	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	16	15	20	10	12	14	LKPFHJ	hg2
417	M	75% MAORI/ENGLISH	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	16	15	20	10	12	14	LKPFHJ	hg2
421	M	100% Native	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	16	15	20	10	12	14	LKPFHJ	hg2
443	M	75% MAORI/EURO	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	16	15	20	10	12	14	LKPFHJ	hg2
478	M	50%MAORI/EURO	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	16	15	20	10	12	14	LKPFHJ	hg2
489	M	50%MAORI/EURO	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	16	15	20	10	12	14	LKPFHJ	hg2
500	M	100% Native	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	16	15	20	10	12	14	LKPFHJ	hg2
512	M	100% Native	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	16	15	20	10	12	14	LKPFHJ	hg2
520	M	100% Native	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	16	15	20	10	12	14	LKPFHJ	hg2
527	M	100% Native	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	16	15	20	10	12	14	LKPFHJ	hg2

Sample_ID	Pop_group	Pop_subgroup		YAP	92R7	sY81	SRY465	SRY4064	TAT	M9	M13	M17	M20	SRY10831	DYS19_ASize	DYS388_ASize	DYS390_ASize	DYS391_ASize	DYS392_ASize	DYS393_ASize	MS_Code	UEP_Code
606	P		SAMOAN	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	16	15	19	10	12	14	LKOFHJ	hg2
93	P		W SAMOAN/N GUINEA	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	16	15	20	10	12	14	LKPFHJ	hg2
106	P		SAMOA	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	16	15	20	10	12	14	LKPFHJ	hg2
604	P		RAROTONGA	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	16	15	20	10	12	14	LKPFHJ	hg2
608	P		W SAMOA	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	16	15	20	10	12	14	LKPFHJ	hg2
609	P		W SAMOA	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	16	15	20	10	12	14	LKPFHJ	hg2
611	P		TONGA	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	16	15	20	10	12	14	LKPFHJ	hg2
617	P		50%COOK/MAORI/SCOT	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	16	15	20	10	12	14	LKPFHJ	hg2
94	P		100% Native	NEGATIVE	C	A	C	G	T	G	G	G+	A	G	15	13	24	10	13	13	KITFII	hg26
104	P		50%HOMEA/TAHITI	NEGATIVE	C	A	C	G	T	G	G	G+	A	G	15	13	24	10	13	13	KITFII	hg26
610	P		TONGA	NEGATIVE	C	A	C	G	T	G	G	G+	A	G	15	14	24	10	13	13	KJTFII	hg26
616	P		TOKELAU	NEGATIVE	C	A	C	G	T	G	G	G+	A	G	16	14	24	9	15	13	LJTEKI	hg26
103	P		TOKELAU	NEGATIVE	C	A	C	G	T	G	G	G+	A	G	16	14	24	9	16	13	LJTELI	hg26
607	P		SAMOAN	NEGATIVE	C	A	C	G	T	G	G	G+	A	G	16	14	24	9	16	13	LJTELI	hg26

Sample_ID	Pop_group	Pop_subgroup	_YAP	92R7	sY81	SRY465	SRY4064	TAT	M9	M13	M17	M20	SRY10831	DYS19_ASize	DYS388_ASize	DYS390_ASize	DYS391_ASize	DYS392_ASize	DYS393_ASize	MS_Code	UEP_Code
529	M	100% Native	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	16	15	20	10	12	14	LKPFHJ	hg2
531	M	100% Native	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	16	15	20	10	12	14	LKPFHJ	hg2
532	M	100% Native	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	16	15	20	10	12	14	LKPFHJ	hg2
422	M	100% Native	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	16	15	20	11	12	14	LKPGHJ	hg2
469	M	100% Native	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	16	15	20	11	13	13	LKPGII	hg2
406	M	100% Native	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	16	15	21	10	12	14	LKQFHJ	hg2
78	M	62.5% MAORI/EURO	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	17	13	24	11	12	13	MITGHI	hg2
64	M	25% MAORI/25%CHINESE/EURO	NEGATIVE	C	A	C	G	T	G	G	G+	A	G	13	12	24	11	11	13	IHTGGI	hg26
83	M	50% MAORI/CHINESE	NEGATIVE	C	A	C	G	T	G	G	G+	A	G	14	10	24	10	14	12	JFTFJH	hg26
82	M	25% MAORI/EURO	NEGATIVE	C	A	C	G	T	G	G	G+	A	G	14	12	23	10	13	13	JHSFII	hg26
534	M	100% Native	NEGATIVE	C	A	C	G	T	G	G	G+	A	G	14	12	25	11	14	13	JHUGJI	hg26
515	M	12.5%MAORI/IRISH/EURO	NEGATIVE	C	A	C	G	T	G	G	G+	A	G	15	13	24	10	13	13	KITFII	hg26
107	P	50% W SAMOAN/GERM	NEGATIVE	T	A	C	G	T	G	G	G+	A	G	14	12	23	11	13	13	JHSGII	hg1
605	P	50% COOK ISLAND	NEGATIVE	T	A	C	G	T	G	G	G+	A	G	14	12	23	11	13	13	JHSGII	hg1
612	P	50%RAROTONGA/EURO	NEGATIVE	T	A	C	G	T	G	G	G+	A	G	14	12	23	11	13	13	JHSGII	hg1
95	P	W SAMOA	NEGATIVE	C	A	C	G	T	C	G	G+	A	G	15	15	20	10	11	14	KKPFGJ	hg2

References

Altheide, T. and Hammer, M. (1997) Evidence for a possible Asian origin of YAP + Y chromosomes. Am. J. Hum. Genet. 61, 462 – 466

Amos, W. (1999) A comparative approach to the study of microsatellite evolution. In: Microsatellites: Evolution and Application (eds. Goldstein, D. and Schlotterer, C.) Pp66 Oxford University Press

Amos, W., and Rubinsztein, D. (1996) Microsatellites are subject to directional evolution. Nat. Gen. 12, 13 – 14

Amos, W., Sawcer, S., Feakes, R. and Rubinsztein, D. C. (1996). Microsatellites show mutational bias and heterozygote instability. Nat. Gen. 13, 390 – 391

Andrews, A., T. (1988) Electrophoresis: Theory, techniques and biomedical and clinical applications. Oxford, Clarendon Press.

Armour. J.. Alegre, S., Miles, S., Williams, L. and Badge, R. (1999) Microsatellites and mutation processes in tandemly repetative DNA. In: Microsatellites: Evolution and Application (eds. Goldstein, D. and Schlotterer, C.) pp24 Oxford University Press

Balding, D., Greenhalgh, M. and Nichols, R. (1996) Population genetics of STR loci in Caucasians. Int. J. Leg. Med. 108, 6, 300 – 305

Barker, M. and Cook, C. (1976) Pears Cyclopedia. A book of background information and reference for everyday use. Eighty-fifth edition (eds. Barker, M and Cook. C.) The Chaucer Press, Suffolk.

Beaglehole, J. (1975) Magellan. In: The exploration of the Pacific 3rd Edition. (ed. Beaglehole, J.) Pp5 Adam and Charles Black, London

Bellwood (1989) The colonization of the Pacific: some current hypotheses. In: The colonization of the pacific: a genetic trail (eds. A.V.S. Hill. & S.W. Serjeantson). pp. 1-60. Clarendon Press: Oxford.

Bianchi, N., Catanesi, C., Bailliet, G., Martinez-Marignac, V., Bravi, C., Vidal-Rioja, R. and Lopez-Camelo, J. (1998) Characterization of ancestral and derived Y chromosome haplotypes of New World native populations. Am. J. Hum. Genet. 63, 1862 – 1871

Bosch, E., Calafell, F., PerezLezaun, A., Comas, D., Benchemsi, N., Tyler-Smith, C. and Bertranpetit, J. (1999). Variation in short tandem repeats is deeply structured by genetic background on the human Y chromosome. Am. J. Hum. Genet. 65, 6, 1623 – 1638

Botstein, D., R.L. White, M. Skolnick and R.W. Davis. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am. J. Hum. Genet. 32:314-331.

Bowcock, A., Kidd, J., Mountain, J., Hebert, J., Carotenuto, L., Kidd, K. and Cavalli-Sforza, L. L. (1991) Drift, admixture and selection in human evolution: A study with DNA polymorphisms. Proc. Natl. Acad. Sci. USA 88, 839 – 843

Bowcock, A., Ruiz-Linares, A., Tomfohrde, J., Minche, E., Kidd, J., and Cavallisforza, L. (1994) High-resolution of human evolutionary trees with polymorphic microsatellites. 368, 6470, 455-457

Brenner, C. and J. Morris. (1990). Paternity index calculations in single locus hypervariable DNA probes: validation and other studies, pp. 21-53. In: Proceedings for the International Symposium on Human Identification 1989. Promega Corporation, Madison, WI.

Brinkmann, B., Sajantila, A., Goedde, H., Matsumoto, H., Nish, K., Weigand, P. (1996) Population genetic comparisons among eight populations using allele frequency and sequence data from three microsatellite loci. Eur. J. Hum Genet. 4, 175 – 182

Bruford, M. and Wayne, R. (1993). Microsatellites and their application to population genetic studies. Current Opinion in Genetics and Development 3, 939 – 943

Burgess, C. (1974) The Bronze age. In: British Prehistory: A new outline (ed. Renfrew, C.) Pp175 Duckworth press, Unwin brothers Surrey

Cameron, K. (1977) English place-names. (ed. Cameron. K. 4th ed.) London:Batsford press

Carvalho-Silva, D., Santos, F., Hutz, M., Salzano, F. and Pena, S. (1999) Divergent human Y chromosome microsatellite evolution rates. J. Mol. Evol. 49, 204 – 214

Chakraborty, R. (1992) Sample size requirements for addressing the population genetic issues of Forensic use of DNA typing. Hum. Bio. 64, 2, 141 – 159

Chakraborty, R., Kimmel, M., Stivers, D., Davison, L., and Deka, R. (1997). Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. Proceedings of the National Academy of Science USA, 94, 1041-1046.

Chu, J., Huang, W., Kuang, S., Wang, J., Xu, J., Chu, Z., Yang, Z., Lin, K., Li, P., Wu, M., Geng, Z., Tan, C., Du, R. and Jin, L. (1998) Genetic relationship of populations in China. Proc. Natl. Acad, Sci. USA 95, 11763 – 11768

Clark, A., Hamilton, F., & Chambers, G. (1995). Inference of population subdivision from the VNTR distributions of New Zealanders. Genetica, 96, 37-49.

Cook, L. (1976) Population genetics. (ed. Cook, L.) London:Chapman and Hall Publishers.

Cowan, J. (1930) Maori and Polynesian Origins. In:The Maori, yesterday and today. (eds. Cowan, J.) Whitcombe and Tombs Limited Auckland
Craig, J., Fowler, S., Burgoyne, L., Scott, A., & Harding, H. (1988). Repetitive deoxyribonucleic acid (DNA) and human genome variation - A concise review relevant to forensic biology. J. For. Sci. 33(5), 1111-1126.

Cunliffe, B. (1974) The Iron age. In: British Prehistory: A new outline (ed. Renfrew, C.) Pp241 Duckworth press, Unwin brothers Surrey

de Knijff, P., Kayser, M., Caglia, A., Corach, D., Fretwell, N., Gehrig, C., Graziosi, G, Heidorn, F, Herrmann, S., Herzog, B., Hidding, M., Honda, K., Jobling, M., Krawczak, M., Leim, K., Meuser, S., Meyer, E., Oesterreich, W., Pandya, A., Parson, W., Penacino, G., PerezLezaun, A., Piccinini, A., Prinz, M., Schmitt, C., Schneider, P. M., Szibor, R., TeifelGreding, J., Weichhold, G., Roewer, L. (1997) Chromosome Y microsatellites: population genetics and evolutionary aspects. Int. J. Legal Med. 110, 134-140

Deka, R., Jin, L., Shriver, M., Yu, L., Saha, N., Barrantes, R., Chakraborty, R. and Ferrell, R. (1996) Dispersion of Human Y chromosome haplotypes based on five microsatellites in global populations. 6, 1177 – 1184

Deka, R., Shriver, M. D., Yu, L., Ferrell, R. and Chakraborty, R. (1995) Intra- and inter-population diversity at short tandem repeat loci in diverse populations of the world. Electrophoresis 16, 1659 – 1664

Diamond, J. (1988) Express train to Polynesia. Nature, 336, 307 – 308

Drozd, M., Archard, L., Lincoln, P., Morling, N., Nellemann, L., Phillips, C., Soteriou, B., and Syndercombe Court, D. (1994). An investigation of the HUMVWA31A locus in British Caucasians. Forensic Science International, 69, 161-170.

Durrans, B. (1979) Ancient pacific voyaging: Cook's views and the development of interpretation. In: Captain Cook and the South Pacific (ed. Mitchell, T) pp137 British Museum Publications, Jolly and Barber Ltd.

Edwards, A., Hammond, H., Jin, L., Caskey, T. and Chakraborty, R. (1992) Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. Genomics 12, 241 – 253

Estoup, A. and Cournet, J-M. (1999) Microsatcllite evoltuion: inferences from population data In: Microsatellites: Evolution and Application (eds. Goldstein, D. and Schlotterer, C.) Pp49 Oxford University Press

Evett, I., W., Gill, P., Scrange, J. and Weir, B. S. (1996). Establishing the robustness of Short Tandem Repeat statistics for Forensic Applications. Am. J. Hum. Genet. 58, 398 – 407

Excoffier, L. and Smouse, P. (1994) Using allele frequencies and geographic subdivision to reconstruct gene trees within a species – Molecular variance parsimony. Gen. 136, 1, 343 - 359

Excoffier, L. and Smouse, P. (1994) Using allele frequencies and geographic subdivision to reconstruct gene trees within a species: Molecular variance parsimony. Genetics 136, 343 – 359

Finney, B. (1977) Voyaging canoes and the settlement of Polynesia. Science 196, 1277 – 1284

Forster, P., Kayser, M., Meyer, E., Roewer, L., Pfeiffer, H., Benkmann, H. and Brinkmann, B. (1998). Phylogenetic resolution of complex mutational features at Y-STR DYS390 in Aboriginal Australians and Papuans. Mol. Biol. Evol. 13, 1213-1218

Freimer, N. and Slatkin, M. (1996) Microsatellites: evolution and mutational processes. In: Variation in the Human Genome. Pp51-72. Ciba Foundation Symposium 197, Wiley Chichester press.

Gibbs. M., Stanford. J., McIndoe. R., Jarvik. G., Kolb. S., Goode, E., Chakrabarti, L., Schuster, E., Buckley, V., Miller, E., Brandzel, S., Li, S., Hood, L. and Ostrander, E. (1999) Evidence for a rare prostate cancer-susceptibility locus at chromosome 1p36. Am. J. Hum. Genet. 64, 776 – 787

Gill, P., & Evett, I. (1995). Population genetics of short tandem repeat (STR) loci. Genetica, 96, 69-87.

Goldstein. D. and Pollock, D (1997) Launching microsatellites: A review of mutation processes and methods of phylogenetic inference. J. Hered. 88, 335 – 342

Goldstein, D. and Schlotterer, C. (1999) Microsatellites: Evolution and Application (eds. Goldstein, D. and Schlotterer, C.) Pp114 Oxford University Press

Goldstein, D., Ruiz Linares, A., Cavalli-Sforza, L. and Feldman, M. (1995). Genetic absolute dating based on microsatellites and the origin of modern humans. Proc. Natl. Acad. Sci. USA. 92, 6723-6727

Haeseler, A., Sajantila, A. and Paabo, S. (1995) The genetical archaeology of the human genome. Nat. Genet. 14, 135 – 140

Hagelberg, E., & Clegg, J. (1993) Genetic polymorphisms in prehistoric Pacific Islanders determined by analysis of ancient bone DNA. Proc. R. Soc. Lond. B., 252, 163-170.

Hagelberg, E., Kayser, M., Nagy, M., Roewer, L., Zimdahl, H., Krawczak, M., Lio, P. and Schiefenhovel, W. (1999) Molecular genetic evidence for the human settlement of the Pacific: analysis of mitochondrial DNA, Y chromosome and HLA markers. Proc. Roy. Soc. Lond. B. 354, 1379, 141 – 152

Hagelberg, E., Quevedo, S., Turbon, D. and Clegg, J. (1994) DNA from ancient Easter Islanders. Nature 369, 25 – 26

Hamilton, J., Starling, L., Cordiner, S., Monahan, D., Buckleton, J., Chambers, G. and Weir, B. (1996) New Zealand population data at five VNTR loci: validation as databases for forensic identity testing. Science and Justice. 36, 2, 109 – 117

Hammer, M. (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, M. and Horai, S. (1995) Y chromosomal DNA variation and the peopling of Japan. Am. J. Hum. Genet. 56, 951 – 962

Hammer, M. and Zegura, S. (1996) The role of the Y chromosome in human evolutionary studies. Evol. Anth. 5, 116 – 134

Hammer, M. F., Spurdle, A., Karafet, T., Bonner, M., Wood, E., Novelletto, E., Malaspina, P., Mitchell, R., Horai, S., Jenkins, T. and Zegura, S. (1997). The geographic distribution of Human Y chromosome variation. Genet 145, 787 – 805

Hammer, M., Karafet, T., Rasanayagam, A., Wood, E. T., Altheide, T. K., Jenkins, T., Griffiths, R. C. (1998) Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. Mol. Biol. Evol. 15, 427-441

Hammond, H., Jin, L., Zhong, Y., Caskey, T. and Chakraborty, R. (1994). Evaluation of 13 short tandem repeat loci for use in personal identification applications. Am. J. Hum. Genet. 55, 175 – 189

Hancock, J. (1999) Microsatellites and other simple sequences: genomic context and mutational mechanisms In: Microsatellites: Evolution and Application (eds. Goldstein, D. and Schlotterer, C.) Pp238 Oxford University Press

Hartl, D. L. (1981) The causes of evolution. In: A primer of Population genetics (ed. Hartl, D.), pp 101. Sinauer Associates Inc, Sunderland Massachusetts

Hartl, D. Principles of population genetics. (1989) (eds. Hartl, D. Clark, G. 2nd edition) Sinauer Associates Inc, Sunderland Massachusetts

Hawkes,J and Hawkes, C. (1958). Prehistoric Britain (eds. Hawkes, J. and Hawkes, C.) Penguin publishers

Heyer, E., Puymirat, J., Dieltjes, P., Bakker, E. and 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, A. and Serjeantson, S. (1989) The Colonization of the Pacific: A genetic trail. (ed. Hill, A. and Serjeantson) pp5. Clarendon Press, Oxford

Hoelzel, A. and Bancroft, D. (1992) Statistical analysis of variation. In:Molecular genetic analysis of populations, a practical approach. (edited by Hoelzel) pp304 Oxford University press.

Hummel, S., Schultes, T., Bramanti, B. and Herrmann, B. (1999) Ancient DNA profiling by megaplex amplifications. Electrophoresis 20, 8, 1717 – 1721

Hurles, M., Irven, C., Nicholson, J., Taylor, P., Santos, F., Loughlin, J., Jobling, M., and Sykes, B. (1998) European Y-Chromosomal Lineages in Polynesians: A Contrast to the Population structure revealed by mtDNA. Am. J. Hum. Genet. 63, 1793-1806.

Imaizumi, Y. (1974) Genetic structure in the United Kingdom. Hum. Hered. 24, 151 – 159

Irwin, G. (1992) The prehistoric exploration and colonisation of the Pacific. Cambridge:Cambridge University Press

Jin. L.. Macaubas, C., Hallmayer, J., Kimura. A. and Mignot, E. (1996). Mutation rate varies among alleles at a microsatellite locus: Phylogenetic evidence. Proc. Natl. Acad. Sci. USA 93, 15285 – 15288

Jobling, M. and Tyler-Smith, C. (1995) Fathers and sons: the Y chromosome and human evolution. Trends. Genet. 11: 449-456

Jobling, M., Pandya, A. and Tyler-Smith, C. (1997) The Y chromosome in forensic analysis and paternity testing Int. J. Leg. Med. 110, 118 – 124

Jones, D.A. (1972). Blood samples: Probability of discrimination. J. Forensic Sci. Soc.12:355-359

Jorde, L., Carey, J., and White, R. (1995). Introduction. In: Medical Genetics (eds Underdown, E.) pp. 3 Mosby press.

Jorde, L., Rogers, A., Bamshad, M., Watkins, S., Krakowiak, P., Sung, S., Kere, J. and Harpending, H. (1997) Microsatellite diversity and the demogaphic history of modern humans Proc. Natl. Acad. Sci. USA 94, 3100 – 3103

Kaeppler, A. (1979) Tracing the history of Hawaiian Cook voyage artefacts in the Museum of Mankind. In: Captain Cook and the South Pacific (ed. Mitchell, T) pp167 British Museum Publications, Jolly and Barber Ltd.

Karafet. T., Osipova, L., Posukh, O.. Weibe, V. and Hammer, M. (1999a) Y chromosome microsatellite haplotypes and the history of Samoyed-speaking populations in north-west Siberia. In: Microsatellites: Evolution and Application (eds. Goldstein, D. and Schlotterer, C.) Pp249 Oxford University Press

Karafet, T., Zegure, S., Posukh, O., Osipova, L., Bergen, A., Long, J., Goldman, D., Klitz, W., Harihara, S., de Knijff, P., Wiebe, V., Griffiths, R., Templeton, A. R. and Hammer, M. (1999) Ancetral Asian source(s) of New World Y-Chromosome founder haplotypes. Am. J. Hum. Genet. 64, 817-831

Kashi, Y. and Soller, M. (1999) Functional roles of microsatellites and minisatellites. In: Microsatellites: Evolution and Application (eds. Goldstein, D. and Schlotterer, C.) Pp10 Oxford University Press

Kayser. M., Caglia, A., Corach, D., Fretwell, N., Gehrig, C., Graziosi, G., Heidorn. F., Herrman, S. Herzog, B., Hidding, M., Honda, K., Jobling, M., Krawczak, M., Leim, K., Meuser, S., Meyer, E., Oesterreich, W., Pandya, A., Parson, W., Penacino, C., Perez-Lezaun, A., Piccinini, A., Prinz, M., Schmitt, C., Schneider, P. M., Szibor, R., Teifel-Greding, J., Weichhold, G., de Knijff, P., and Roewer, L. (1997) Evaluation of Y-chromosomal STRs: a multicenter study. Int. J. Legal Med. 110, 125-133

Kimura, M. and Crow, J. (1964) The number of alleles that can be maintained in a finite population. Genetics 49, 725 – 738

Kirk, R. (1989) Population genetic studies in the Pacific: red cell antigen, serum protein, and enzyme systems. In: The Colonization of the Pacific: A genetic trail. (ed. Hill, A. and Serjeantson) pp 60 Clarendon Press, Oxford

Kittles, R. A., Perola, M., Peltonen, L., Bergen, A. W., Aragon, R. A., Virkkunen, M., Linnoila, M., Goldman, D., and Long, J.C. (1998) Dual origins of Finns revealed by Y Chromosome haplotype variation. Am. J. Hum. Genet. 62, 1171-1179

Kopec, A. (1970) The distribution of blood groups in the United Kingdom. Oxford University Press, Oxford.

Laing, L. and Laing, J. (1980) The Origins of Britain (ed. Wheatcroft, R. and Kegan, P.), London and Henley press

Lareu, M., Barral, S., Salas, A., Rodriguez, M., Pestoni, C. and Carracedo, A. (1998) Further exploration of new STRs of interest for forensic genetic analysis. Prog. For. Gen. 7, 198 – 200

Leeds, T. (1913) Introductory and geographical considerations. In:The archaeology of the anglo-saxon settlements. (ed. Hogarth, D.) pp13 Clarendon Press Oxford

Lehmann, H., North, A. and Staveley, J. (1958) Absence of the Diego Blood Group and abnormal Haemoglobins in 92 Maoris. Nat. 18, 791 – 792

Levinson, G. and Gutman, G. (1987) Slipped-strand mispairing – a major mechanism for DNA – sequence evolution. Mol. Biol. Evol. 4, 3, 203 – 221

Long, C., Darke, C. and Marks, R. (1998). Celtic ancestry, HLA phenotype and increased risk of skin cancer. Brit. J. Dermat. 138, 4, 627 – 630

Lum, K.. Cann., Martinson, J. and Jorde, L. (1998) Mitochondrial and nuclear genetic relationships among pacific Island and Asian populations. Am. J. hum. Genet. 63, 613 – 624

Lum, K., Rickards, O., Ching, C. and Cann, R. (1994) Polynesian mitochondrial DNAs reveal three deep maternal lineage clusters. Hum. Bio. 66, 4, 567 – 590

Maiste, P. and Weir, B. S. (1995). A comparison of tests for independence in the FBI RFLP data bases. Genetica 96, 125 – 138

Mastana, S. and Sokol, R., J. (1998) Genetic variation in the East Midlands. Ann. Hum. Bio. 25, 1, 43-69

Mathias, N., Bayes, M. and Tyler-Smith, C. (1994) Highly informative compound haplotypes for the human Y chromosome. Hum. Mol. Genet. 3, 115-123

McConkey, E. (1993). Introduction. In Human Genetics: The molecular revolution (pp. 2). Boston: Jones and Bartlett.

Melton, T., Peterson, R., Redd, A., Saha, N., Sofro, A., Martinson, J. and Stoneking, M. (1995) Polynesian genetic affinities with southeast Asian populations as identified by mtDNA analysis. Am. J. Hum. Genet. 57, 403 – 414

Michalakis, Y. and Excoffier, L. (1996) Generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. Gen. 142, 1061 – 1064

Mitchell, R., Vandenberg, N., Van Oorschot, R. and Tyler-Smith, C. (1999) Y chromosome specific microsatellite variation in Australian Aboriginal people. Am. J. Hum. Biol. 11, 1, 155

Murray-McIntosh, R., Scrimshaw, B., Hatfield, P., and Penny, D. (1998) Testing migration patterns and estimating founding population size in Polynesia by using human mtDNA sequences. Proc. Natl. Acad. Sci. USA. 95, 9047-9052.

National Research Council (1996) The evaluation of DNA evidence (eds. NRC council) National Academy Press, Washington D.C.

Nei, M. (1987) Genetic Variation within Species. In: Molecular Evolutionary Genetics (pp. 176-207). New York: Columbia University Press.

Nellemann, L., Moller, A. and Morling, N. (1994) PCR typing of DNA fragments of the short tandem repeat (STR) system HUMTHO1 in Danes and Greenland Eskimos. Forensic. Sci. Int. 68, 45 – 51

Ohta. T. and Kimura, M. (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. 22, 201 – 204

Perez-Lezaun. A., Calafell, F., Mateu, E., Comas, D., Ruiz-Pacheco, R. and Bertranpetit, J. (1997b) Microsatellite variation and the differentiation of modern humans. Hum. Gen. 99, 1 – 7

Perez-Lezaun, A., Calafell, F., Seielstad, M., Mateu, E., Comas, D., Bosch, E. and Bertranpetit, J. (1997a) Population Genetics of Y-Chromosome Short Tandem Repeats in Humans. J. Mol. Evol. 45, 265-270.

Phillips, G. (1931) The Blood groups of the Maori. Hum. Biol. 3, 282-287

Primmer, C., Ellegren, H., Saino, N. and Moller, A. (1996) Directional evolution in germline microsatellite mutations. Nat. Gen. 13, 4, 391-393

Pritchard, J., Seielstad, M., Perez-Lezaun, A. and Feldman, M. (1999) Population growth of Human Y chromosomes: A study of Y chromosome microsatellites. Mol. Biol. Evol. 16, 12, 1791 – 1798

Renfrew , C. (1974) British Prehistory: A new outline (ed. Renfrew, C.) Duckworth press. Unwin brothers Surrey

Rholf, F. (1993) NTSYS-pc numerical taxonomy and multivariate analysis system version 1.80. Exeter Software. Applied Biostatistics Inc.©, New York

Richards, M., Oppenheimer, S. and Sykes, B. (1998) mtDNA suggests Polynesian origins in Eastern Indonesia. Am. J. Hum. Genet. 63, 1234 – 1236

Roberts, D. (1973). Genetic Variation in Britain. (eds. Roberts, D. and Sunderland, E.) Pp 1 –16 Taylor and Francis, London

Roewer, L., Kayser, M., Dieltjes, P., Nagy, M., Bakker, E., Krawczak, M. and de Knijff, P. (1996) Analysis of molecular variance (AMOVA) of Y-chromosome-specific microsatellites in two closely related human populations. Hum. Mol. Genet. 5: 7, 1029-1033

Rostedt, I., Lalu, K., Lukka, M. and Sajantila, A. (1996). Genotyping of five short tandem repeat loci via triplex and duplex PCR. Forensic Science International 82, 217 – 226

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

Rubinsztein, D. (1999) Trinucleotide expansion mutations cause diseases which do not conform to classical Mendelian expectations In: Microsatellites: Evolution and Application (eds. Goldstein, D. and Schlotterer, C.) Pp80 Oxford University Press

Rubinsztein, D., Amos, W., Leggo, J., Goodburn, S., Jain, S., Li, S., Margolis, R., Ross, C. and Furguson-Smith, M. (1995) Microsatellite evolution – evidence for directionality and variation in rate between species. Nat. Gen. 10, 337 – 343

Ruiz-Linares, A., Nayar, K., Goldstein, D., Hebert, J., Seielstad, M., Underhill, P., Lin, A., Feldman, M. and Cavalli-Sforza, L. L. (1996) Geographic clustering of human Y chromosome haplotypes. Ann. Hum. Genet. 60, 401 – 408

Ruiz-Linares, A., Ortiz-Barrientos, D., Figueroa, M., Mesa, N., Munera, J. A., Bedoya, G., Velez, I. D., Garcia, L. F., Perez-Lezaun, A., Bertranpetit, J., Feldman. M. W. and Goldstein, D. B. (1999). Microsatellites provide evidence for Y chromosome diversity among founders of the New World. Proc. Natl. Acad. Sci. USA 96, 6312-6317

Saferstein, R. (1986). Forensic Science Handbook Vol. II. (ed. Saferstein, R.) Prentice Hall, New Jersey.

Santos, E., Epplen, J. and Epplen, C. (1997) Extensive gene flow in human populations as revealed by protein and microsatellite DNA markers. Hum. Hered. 47, 165 – 172

Santos, F., Pandya, A., Tyler-Smith, C., Pena, S., Schanfield, M., Leonard, W., Osipova, L., Crawford, M. and Mitchell, J. (1999). The central Siberian origin for native American Y chromosomes. Am. J. Hum. Genet 64, 619 – 628

Santos, F. and Tyler-Smith, C. (1996). Reading the human Y chromosome: the emerging DNA markers and human genetic history. Braz. J. Genet. 19, 665-670

Satow, A., and Akiyama, T. (1993). Simultaneous determination of the migration coefficient of each base in heterogeneous oligo-DNA by gel filled capillary electrophoresis. Journal of Chromatography, 652, 23-30.

Schlotterer, C. and Wiehe, T. (1999) Microsatellites, a neutral marker to infer selective sweeps In: Microsatellites: Evolution and Application (eds. Goldstein, D. and Schlotterer, C.) Pp238 Oxford University Press

Schneider, S., Kueffer, J-M., Roessli, D. and Excoffier, L. (1997). Arlequin ver 2.0b: a software for population genetic data analysis. Genetics and Biometry Laboratory, University of Geneva, Geneva

Scozzari, R., Cruciana, F., Santolamazza, P., Malaspina, P., Torroni, A., Sellitto, D., Arredi, B., Destro-Bisol, De Stefano, G., Rickards, O., Martinez-Labarga, C., Modiano, D., Biondi, G., Moral, P., Olckers, A., Wallace, D. and Novelletto, A. (1999). Combined use of biallelic and microsatellite Y-chromosome polymorphisms to infer affinities among African populations. Am. J. Hum. Genet. 65, 829 – 846

Seielstad, M. T., Herbert, J. M., Lin, A. A., Underhill, P. A., Ibrahim, M., Vollrath, D., Cavali-Sforza, L. L. (1994). Construction of human Y-chromosomal haplotypes using a new polymorphic A to G transition. Hum. Mol. Genet. 3, 2159-2161

Seielstad, M., Bekele, E., Ibrahim, M., Toure, A. and Traore, M. (1999) A view of Modern Human origins from Y chromosome microsatellite variation Gen. Res. 9, 558 – 567

Sensabaugh, G. (1982) Biochemical markers of individuality. In: Forensic Science Handbook Volume I. (ed. Saferstein) Pp339 Prentice Hall Regents. New Jersey

Sjerps, M., Van der Geest, N., Pieron, C., Gajadhar, M., & Kloosterman, A. (1995). A Dutch population study of the STR loci HUMTHO1, HUMFES/FPS, HUMVWA31/1 and HUMF13A1, conducted for forensic purposes. Int. J. Leg. Med. 108, 127-134.

Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. Genetics 139, 457-462

Sneath, P. and Sokol, R. (1973) Numerical Taxonomy. Freeman publishers, San Francisco

Sokol, R. and Michner, C. (1958) A statistical method for evaluating systematic relationships. University of Kansas Sci. Bull. 28, 1409 – 1438

Spurdle, A. and Jenkins, T. (1992). The search for Y-chromosome polymorphism is extended to negroids. Hum. Mol. Genet. 1, 169-170

Spurdle, A., Woodfield, D., Hammer, M. and 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. and Richards, M. (1995) The origins of the Polynesians: An interpretation from mitochondrial lineage analyis. Am. J. Hum. Gen. 57, 1463 – 1475

Terrell, J. (1986) Science and Prehistory. In: Prehistory in the Pacific Islands (ed. Terrell, J.) Cambridge University Press

Thomas, M. G., Bradman, N. and Flinn, H. M. (1999). High throughput analysis of 10 microsatellite and 11 diallelic polymorphisms on the human Y-chromosome. Hum. Genet. 105; 6, 577-581

Thomas, M. G., Parfitt, T., Weiss, D. A., Skorecki, K., Wilson, J. F., Le Roux, M., Bradman, N. and Goldstein, D. B. (2000). Y Chromosomes travelling South: The Cohen modal haplotype and the Origins of the Lemba-the "Black Jews of Southern Africa". Am. J. Hum. Genet. 66, 674-686

Thomas, M., Skorecki, K., Ben-Ami, H., Parfitt, T., Bradman, N., & Goldstein, D. (1998). Origins of Old Testament priests. Nature. 394, 138-140.

Thompson, W. C. (1995). Subjective interpretation, laboratory error and the value of Forensic DNA evidence: three case studies. Genetica 96, 153 – 168

Thurn. E. (1931) A study of primitive character. In: Transactions of Anthropology – presidential address. Pp515 – 519. (ed. Sir Everard Thurn) London Press.

Underhill, P.A., Jin, L., Lin, A.A, Mehdi, S.Q., Jenkins, T., Vollrath, D., Davis, R.W, CavalliSforza, L.L. Oefner, 0. (1997) Detection of numerous Y-chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. Gen. Res. 7, 996-1005

Urquhart. A., Oldroyd, N., Kimpton, C. and Gill, P. (1995) Highly discriminating heptaplex short t andem repeat PCR system for forensic identification. Biotechniques. 18, 116 – 121

Utah marker Development Group (1995) A collection of ordered tetranucleotide-repeat markers from the Human Genome. Am. J. Hum. Genet. 57, 619 – 628

Wall. W., Williamson, R., Petrou, M., Papaioannou, D., and Parkin, B. (1993). Variation of STR repeats within and between populations. Human Molecular Genetics, 2(7), 1023-1029.

Ward, R. G. (1972) Man in the Pacific Islands. Oxford, London: Clarendon Press

Watson, E., Gill, P. and Mastana, S. (1998) Genetic diversity at the HUMTHO1 locus. Ann. Hum. Biol. 25, 6, 563-580

Weaver, R. (1992) Basic genetics: a contemporary perspective (eds. Robert F. Weaver, Philip W. Hedrick) Oxford Press.

Weber, J. and Wong, C. (1993) Mutation of human short tandem repeats. Hum. Mol. Gen. 2, 8, 1123 – 1128

Weir, B.. S. (1996) Genetic Data Analysis 2nd Edition. Sunderland, Mass. Press. Sinauer Associates.

Weir, B. S. (1994) The effects of inbreeding on forensic calculations. Annu. Rev. Genet. 28. 597 – 621

Whitfield, L.S., Sulston, J.E. and Goodfellow, P.N. (1995). Sequence variation of the human Y chromosome. Nature 378, 379-380

Wilson, I. and Balding, D. (1998) Genealogical inference from microsatellite data. Genetics 150, 499 – 510

Woodfield, G., Simpson, L., Seber, G. and McInerney, P. (1987) Blood grouups and other genetic markers in New Zealand Europeans and Maoris. Ann. Hum. Bio. 14, 1. 29 – 39

Xin, H., Matt, D., Qin, J-Z., Burg, G. and Boni, R. (1999). The sebaceous nevus: A nevus with deletions of the PTCH gene. Cancer Res. 59, 8, 1834 – 1836

Zeiger, R., Salomon, R., Dingman, C., and Peacock, A. (1972). Role of base composition in the electrophoresis of microbial and crab DNA in polyacrylamide gels. Nat. New Biol. 238, 65-69.

Zerjal, T., Dashnyam, B., Pandya, A., Kayser, M., Peower, L., Santos, F. R., Schiefenhovel, W., Fretwell, N., Jobling, M. A., Harihara, S., Shimizu, K., Semjidmaa, D., Sajantila, A., Salo, P., Crawford, M. H., Ginter, E. K., Evgrafov, O. V. and Tyler-Smith, C. (1997). Genetic relationships of Asians and northern Europeans, revealed by Y-chromosomal