GENETIC VARIATION AND DISEASE IN THE ROMA (GYPSIES)

DAVID JULIAN JOHN GRESHAM
(BSc., McGill University)

This thesis is submitted for the degree of Doctor of Philosophy

Undertaken at:
The Centre for Human Genetics
Faculty of Communication, Health and Science
Edith Cowan University, Western Australia

Principal supervisor: Associate Professor Luba Kalaydjieva
Associate supervisor: Professor Alan Bittles

Date of Submission: 30th August 2001

USE OF THESIS

ABSTRACT

The Roma (Gypsies) are a European people composed of a mosaic of culturally heterogeneous populations. Linguistic analyses point to their origins in the Indian subcontinent. Cultural diversity in extant Romani populations suggests that they are descended from a mixture of Indian populations. Previous population genetic studies of the Roma have supported this claim by demonstrating the genetic heterogeneity of Romani populations. More recently, medical genetic research has detected identical founder mutations in separated Romani populations, which provides evidence of their relatedness. In this thesis, the genetic heritage of the Roma and its significance for genetic disease and research is investigated. Male and female lineages were analysed in eight traditionally endogamous Romani populations. Asian specific Y chromosome haplogroup VI-68 and mitochondrial DNA (mtDNA) haplogroup M were detected in all populations and accounted for 39% and 25% of all lineages respectively. Diversity within haplogroups was assessed by genotyping Y chromosome short tandem repeats (Y STRs) and sequencing the mtDNA hypervariable segment 1 (HVS1). Lineages within haplogroups VI-68 and M were found to be closely related suggesting that Romani populations are predominantly descended from a single Indian ethnic population. The differing historical legacies of Romani populations and adherence to endogamous practices have resulted in genetic substructure and limited diversity within populations. Thus, the Roma are shown to comprise a conglomerate of related admixed population isolates. The unique genetic heritage of the Roma provides a powerful tool for the positional cloning of monogenic disease genes. This is demonstrated through the reduction of the critical chromosomal region for a novel genetic disorder, hereditary motor and sensory neuropathy type Lom (HMSNL). In the initial report, the HMSNL disease locus was defined as a 3cM region on chromosome 8q24. In this study, refined genetic mapping utilising historical and parental recombinations observed in Romani individuals from different populations reduced the HMSNL critical interval to 202kb. Sequence analysis of two genes contained within this genomic interval found all affected individuals to be homozygous for a C→T mutation in codon 148 of *N-myc downstream regulated gene 1* (*NDRG1*), resulting in a truncating R148X mutation. Investigation of

the population distribution of the R148X disease allele shows that it occurs in six of eight separated Romani populations. Another founder mutation, C283Y in the γ-sarcoglycan gene (*SGCG*), which causes limb girdle muscular dystrophy type 2C (LGMD2C), was found in two of eight Romani populations. Profound founder effects are apparent within Romani populations with a carrier frequency of 19.5% determined for the R148X mutation in the Lom population, and 6.25% for the C283Y allele in the Turgovzi population. High carrier frequencies for autosomal recessive diseases can be expected to pose a significant health risk for these communities. Thus, community-wide carrier testing represents a potential means of addressing this health problem. A pilot community based carrier-testing program was implemented in a Romani community of northeastern Bulgaria and relevant attitudes assessed by means of a questionnaire. Community-based carrier screening was demonstrated to be an appropriate approach to improving health amongst the Roma.

DECLARATION

I certify that this thesis does not, to the best of my knowledge and belief:

(i) incorporate with acknowledgement any material previously submitted for a degree or diploma in any institution of higher education;

(ii) contain any information previously published or written by another person except where due reference is made in the text; or

(iii) contain any defamatory material.

David Gresham

To my Father, who has engendered in me a passion for knowledge, and my Mother, who has taught me that knowledge takes many forms.

PERSONAL ACKNOWLEDGEMENTS

I wish to thank my principal supervisor Associate Professor Luba Kalaydjieva for providing me with a rich and stimulating intellectual experience. I have benefited greatly from her knowledge, passion, guidance, and the opportunities with which she has provided me.

I am indebted to my asssociate supervisor Professor Alan Bittles for his help and mentorship.

Work completed for the purposes of this thesis has been generously supported by The Wellcome Trust, the National Health and Medical Research Council, the Australian Research Council and Edith Cowan University.

Throughout the duration of this degree I have been aided and tutored by a number of scientists whose efforts have been instrumental in my development as a scientist. It has also been a pleasure to develop close personal relationships with them. They include Dr Anna Pérez-Lezaun , Dr Lisa Heather, Dr Dora Angelicheva, Dr Bharti Morar, David Chandler & Dr Wei Wang. As well, my virtual tutors, Dr Francesc Calafell and Dr Guiseppe Passarino have taught me much.

Thanks to Bec Gooding, with whom I shared the highs and lows of positional cloning.

Thankyou to all the staff and students at the Centre for Human Genetics.

And to Mum, Dad, Katie and Peter for their interest and support and good food.

ACKNOWLEDGEMENT OF WORK CARRIED OUT BY OTHER PERSONS

Section I

- Unique event polymorphisms on the Y chromosome and restriction fragment length polymorphisms in the mitochondrial genome were genotyped by Dr Peter Underhill and Dr Guiseppe Passarino, Department of Genetics, Stanford University.
- Dr Bharti Morar sequenced the mitochondrial DNA hypervariable segment 1 for 16 Intreni samples.

Section II

- Technical assistance was provided by Rebecca Gooding in constructing the bacterial artificial chromosome (BAC) contig.
- Genotyping analysis of polymorphic repeats in the HMSNL region was performed in conjunction with Associate Professor Luba Kalaydjieva, Dr Dora Angelicheva and David Chandler, Centre for Human Genetics, Edith Cowan University.
- Two PAC clones were identified by Ros de Jonge, Academic Medical Centre, Amsterdam, The Netherlands.
- Sequencing of genomic clones spanning the HMSNL region was performed at the Jena Centre for Molecular Biotechnology under the auspices of the Human Genome Project.

Section IV

- The questionairre was constructed by Professor Assen Jablensky, Department of Behavioural Science and Psychiatry, University of Western Australia, and Associate Professor Luba Kalaydjieva, Centre for Human Genetics, Edith Cowan University.
- Genetic counselling, sample collection and adminsteration of the questionairre was performed by a team led by Associate Professor L. Kalaydjieva and Dr I. Tournev of the Medical University, Sofia, Bulgaria.

TABLE OF ABBREVIATIONS

%	Percent
<	Less than
>	Greater than
A	Adenosine
amps	Amperes
BACs	Bacterial artificial chromosome(s)
bp	Base pairs
BP	Before present
BSA	Bovine Serum Albumin
°C	Celsius
C	Cysteine
C	Cytosine
cDNA	Complementary DNA
cm	Centimetre
cM	CentiMorgan
dH_2O	Distilled water
DNA	Deoxyribonucleic acids
dNTPs	Dinucleotide phosphate(s)
DTT	Dithiothreitol
E	Glutamic acid
EDTA	Ethylenediaminetetraacetic acid
EST(s)	Expressed sequence tag(s)
G	Guanine
gen(s)	Generation(s)
HVS1	Hypervariable segment 1
IBD	Identical By Descent
IBS	Identical By State
K	Lysine
kb	Kilobase

kDa	Kilodalton
LB	Luria-Bertani medium
LD	Linkage disequilibrium
Lod	Logarithm of odds
M	Molar
Mb	Megabase
$MgCl_2$	Magnesium chloride
min(s)	Minute(s)
mL	Milllitre
mM	Millimolar
MtDNA	Mitochondrial DNA
N	Sample Size
NaI	Sodium Iodide
NaOH	Sodium Hydroxide
ng	Nanogram
NIH	National Institute of Health (USA)
nM	Nanomolar
P	Proline
$^{32}\gamma$-P-ATP	Adenosine triphospate labelled with radioactive phosphate
$^{32}\alpha$-P-CTP	Cytosine triphophate labelled with radioactive phosphate
PAC	P1 atificial chromosome
PBS	Phosphate-buffered saline
PCR	Polymerase chain reaction
pers comm	Personal comunication
pM	Picomolar
R	Arginine
RFLP	Restriction fragment length polymorphism
Rh	Rhesus
RNA	Ribonucleic acids

rpm	Revolutions per minute
s	Seconds
SNP	Single nucleotide polymorphism
SSC	Sodium chloride/sodium citrate solution
STR	Short tandem repeat
STS(s)	Sequence tagged site(s)
T	Threonine
T	Thymine
TAE	Tris-acetate/EDTA electrophoresis buffer
TBE	Tris-borate/EDTA electrophoresis buffer
TE	Tris-EDTA
TEMED	N, N, N', N'-tetramethylethylenediamine
U	Unit
UEPs	Unique event polymorphism(s)
UL	Microlitre
v	Version
V	Volt
x	Times (in reference to solution concentration)
X	Stop codon
Y	Tyrosine
YAC(s)	Yeast atrificial chromosome(s)
Y STR	Y chromosome short tandem repeat

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Human populations that are genetically isolated by geography or culture provide a unique resource for identifying the genetic basis of Mendelian disorders. In these populations, there is typically a reduction in the diversity of factors underlying inherited disease. In addition, monogenic traits that are otherwise rare or absent in other populations often occur at increased frequencies. This combination of factors permits approaches to the discovery of disease genes in population isolates that are inapplicable in heterogeneous populations. Determination of these gene defects generally leads to diagnostic and predictive testing tools and rational approaches to disease management. In addition, the identification of malfunctioning genes provides an entry point from which gene function and cellular processes can be investigated. Therefore, knowledge gained from studies of rare genetic diseases in minority populations extends beyond the scope of the specific disorder.

During the last ten to fifteen years, studies of Mendelian disorders in population isolates have been successful in identifying a large number of novel disease genes. Frequently, the success of these studies has required the construction of hypotheses informed by knowledge of population history and social structure. The investigation of polymorphic genetic markers provides a means of correlating history and social phenomena with the genetic composition and structure of populations. Knowledge gained from such investigations has been essential to the design of studies that seek to determine the aetiology of genetic disorders within the population. At the same time, studies of heritable markers within a population can serve to illuminate its history, which otherwise may remain ill defined in the absence of historical or archaeological records. Through the characterisation of genetic variation, the origins, histories and social practices of a population can be inferred. Using molecular genetic tools, it is now possible to examine questions of varying depths in time, and to compare the possibly contrasting histories of males and females within a population.

Scriver (1992) has distinguished two causative factors resulting in the manifestation of genetic disease. The ultimate cause is the biological component, namely the disease gene. The more proximate causes are dependent on the

circumstances of an individual's life. At a population level, the proximate causes of a particular genetic disorder include demographic circumstances and cultural practices. Hence, the investigation of genetic disease within populations requires the combined study of biological and social phenomena. My personal interests lie in biochemistry and molecular genetics, and in anthropology and archaeology. The study of genetic variation and disease in human populations represents a field in which these seemingly disparate disciplines can be integrated. Whilst I am fascinated by the physical phenomena that constitute life at the cellular level, my particular passion is discovering the rich diversity within human populations and in their histories, their cultures, and their legacies. Thus, in this doctoral thesis I have attempted to conduct a study that encompasses my diverse interests through the examination of the proximate and ultimate causes of genetic disease.

From a variety of perspectives, the Roma of Europe are a complex and fascinating study population. On the basis of linguistics and social anthropology, Romani populations are believed to have originated in the Indian subcontinent and historical records point to their arrival in Europe at least 800 years ago (Fraser, 1992). The legacy of the Roma in Europe has been one of marginalisation and persecution which, combined with a strong internal social cohesion manifest in the practice of endogamy, has resulted in the maintenance of group identities distinct from those of other European populations. Today, the extant Romani populations of Europe represent an amalgam of geographically and socially separated groups. These groups are culturally diverse, with complex intergroup affinities that often exclude neighbouring communities while transcending national boundaries. The historical relatedness of these many groups has hitherto remained unclear.

Previous genetic studies of the Roma have demonstrated that the social and cultural diversity of specific sub-populations is reflected in their genetic composition. Numerous studies of the Roma have sought evidence of biological affinities with Indian populations. However, the heterogeneity of Romani populations and the methodologies employed in these studies have effectively precluded the formation of rigorous conclusions regarding the population origin of the Roma. Many studies purport to show that Romani groups are genetically distinct from other European populations, and

comparisons between different Romani populations suggest that many are only distantly related. This conclusion is not supported by the presence of identical founder mutations in geographically (Abicht et al., 1999; Lasa et al., 1998; Piccolo et al., 1996; Todorova, Ashikov, Beltcheva, Tournev, & Kremensky, 1999) and socially (Kalaydjieva et al., 1996) separated Romani populations, which provides evidence of their genetic relatedness. Thus, the precise nature of the genetic relationships between Romani populations remains unclear.

Illumination of the genetic structure of the Roma can be expected to be of benefit to research into genetic disorders in these populations. Novel genetic disorders identified in the Roma have all shown evidence of homogeneous aetiologies (Angelicheva et al., 1999; Kalaydjieva et al., 1996; Rogers et al., 2000). However, the appropriate strategies for refining genomic regions and elucidating the causative gene defects in the Roma have been unknown. Although identical disease-causing mutations have been identified in different Romani populations, their distribution has not been systematically examined. Knowledge of the distribution of deleterious alleles in the Roma should provide a useful resource for disease diagnosis and predictive testing. Populations in which specific disease alleles occur at high frequencies lend themselves to targeted carrier testing. This has never been undertaken in Romani populations and thus the appropriate approach, and the salient social and psychological factors that would impact on predictive testing, are yet to be determined.

In this thesis, I have explored the hypothesis that separated Romani populations comprise a mosaic of related genetic isolates. Through the use of molecular genetic tools, the origin and nature of the genetic heritage of the Roma has been examined. The populations included in the study are culturally and geographically diverse allowing an examination of the relationship between social and genetic structure. I hypothesise that knowledge of the genetic structure of the Roma is a useful tool for application to the positional cloning of disease genes. Furthermore, this structure is predicted to impact on the distribution and frequency of disease alleles, and therefore is an important consideration in approaches to predictive testing and health. To investigate the central hypothesis of the thesis and associated issues, I have considered specific aspects of genetic variation and disease in the Roma. These are represented by four inter-related

studies that collectively form a complementary approach to the study of proximate and ultimate causes of genetic disease in the Roma. In the body of the dissertation, these studies are presented in the following format:

In section I, the genetic composition, structure and diversity of the Roma is assessed through the characterisation of maternal and paternal lineages. This allows insights into sex-specific histories and social practices. Comparison of these lineages with other worldwide populations provides a means of disentangling population origins. Within the Roma, the examination of lineages in different populations affords insights into their relatedness, and their differing historical legacies. Analysis of lineages within Romani populations examines the implication of the cultural practice of endogamy on intrapopulation genetic diversity.

In section II, the relevance of genetic structure in the Roma to the investigation of genetic disease is examined. This is achieved through the refined genetic mapping and positional cloning of the gene defect underlying a novel autosomal recessive genetic disorder, hereditary motor and sensory neuropathy type Lom (HMSNL). HMSNL was first identified by Kalaydjieva et al., (1996) with a conserved disease haplotype spanning 3cM on chromosome 8q24 pointing to an identical founder mutation in three socially separated Romani populations. In that study, limited haplotype diversity within a large kindred was useful for mapping the disease gene. However, the refinement of a disease locus requires variation in disease chromosomes that has resulted from recombination. Thus, the most appropriate approach to identifying the disease gene was not immediately apparent. In this study, I examine the relevance of population history and structure to refined genetic mapping and positional cloning in the Roma.

Section III comprises an investigation of the history and distribution of disease alleles in different Romani populations. Haplotype analysis provides a means of tracing the evolution and history of a mutation within populations. This approach allows additional insights into population history, thus complementing the use of paternal and maternal lineages described in section I. Furthermore, the history and diversification of disease haplotypes is an important consideration for disease gene mapping and positional cloning. This history is also related to the distribution of disease alleles.

Systematic examination of disease alleles in different Romani populations is useful for diagnostic purposes and the design of carrier testing programs.

In section IV, a pilot genetic carrier screening study for a private mutation causing limb girdle muscular dystrophy type 2C (LGMD2C) is described and assessed. Genetic determinants are usually traced within families; however, a high disease allele frequency within an isolated Romani population has been shown to result in an entire population being at increased risk for a rare genetic disorder (Plasilova et al., 1999). In such cases, it is possible that carrier testing should not be limited to family members of index cases, but provided to entire communities. Effectiveness of carrier testing within a community is largely dependent on the prevailing cultural and psychological attitudes within the community. To investigate these attitudes, I have analysed the results of a questionnaire administered to the Romani community participating in the pilot carrier testing study.

The four sections of the thesis provide a comprehensive approach to the study of genetic variation and disease in the Roma and are presented in a natural order that allows the knowledge gained in each study to be applied to the following section(s). It should be noted that the sections have not been presented in the strict order in which the studies were performed and where possible, I have attempted to avoid repetition of data and ideas from each section. At this point it is also worth mentioning the use of terminology within the thesis. The term "Gypsy" has derogatory connotations that merit its discontinuance. Therefore, I have used the term *Roma* as the noun describing the population of study. The adjective used to describe the population is *Romani*. To avoid confusion, I refer to the language spoken by these people as *Romany*.

CHAPTER 1

REVIEW OF LITERATURE ON THE ROMA

1.1 History Of The Roma

1.1.1 Introduction

The Roma are a people found throughout Europe and former European colonies. The total size of the Romani population is difficult to determine, but estimates range from 4 to 10 million within Europe (Fraser, 1992; Liégeois, 1994) and possibly 12-15 million worldwide (Liégeois, 1994). The Roma comprise numerous socially and culturally distinct groups. These groups may live in close geographical proximity but speak different dialects and languages, practice different traditional trades, conform to different religions, adhere to different cultural customs and have vastly different historical legacies. At the same time, many Romani populations have retained cultural characteristics that are common to disparate groups. Salient commonalties include the Romany language in its many dialects, the preservation of the "Group", and cultural features such as strict hygiene laws and a belief in spiritual power and fate (Rishi, 1976). For much of their history in Europe, the Roma have been the target of state-sanctioned discriminatory policies and faced persecution from neighbouring peoples. Therefore, whether by choice or of necessity, the Roma have lead a largely peripatetic existence (Fraser, 1992). The adoption of specialised trades required by the macro-society enabled the Roma to fill economic niches. Today, the Roma have generally adopted sedentary lifestyles and abandoned their traditional trades. Nevertheless, significant migrations still occur in present times, typically as a result of social upheavals in the macro-society. Recent examples of events precipitating major migrations of Romani people include the political changes in Eastern Europe in the early 1990s and the 1999 war in Kosovo.

The origins of the Roma have been a disputed academic and social question virtually since their appearance in Europe. The Roma have no historical records and it is generally believed that they are unaware of their origins. According to Fraser (1992) three scholars are credited with more or less simultaneously identifying Romany as having Indo-Aryan roots; Vali (1753-4), Rüdiger (1782) and Bryant (1785). However, the use of linguistics to study the Roma did not enter the realm of serious scholarship until 1870 when the gypsiologist, Paspati, stated that "the key to the history of the Roma should be sought in the study of the Romany language" (Fraser, 1992). Thus, for the past 150 years and largely on the basis of linguistic evidence, the Roma have been considered as being of Indian origin. Some researchers have disputed this claim, for example Okely (1983), who contends that the British Roma are displaced serfs from the agricultural revolution. The issue is not unimportant since, according to the Romani scholar Ian Hancock (1991), the claims of European origins of the Roma imply that the historical and current practices of discrimination against them are based on social, rather than racist grounds.

Discriminatory practices against the Roma have been common throughout their history in Europe. At one time or another, in many of the countries of Europe it has been illegal to be a "Gypsy" and such a crime was punishable by death. Policies and practices of forcible expulsion, internment, sedenterisation, assimilation and extermination have ensured that the Gypsies have been oppressed for the majority of their history in Europe. This has been punctuated by particularly heinous atrocities, such as the 500 years of enslavement in the Danubian principalities of Wallachia and Moldavia and the annihilation of the Roma during the Third Reich. Genetic studies played a notorious role in the Holocaust (referred to by the Roma as the Pojaramos, the Great Devouring) through the "scientific" classification of people of "Gypsy blood" bound for the death camps (Müller-Hill, 1998).

1.1.2 On the Origins and Exodus of the Roma

Original historical records, which describe the Roma, generally refer to them as being of Egyptian ancestry. However, scholars have discounted these records as perpetuating a historical misnomer. It is possible that this arose because one of the first

places in Europe where the Roma resided was known as “Little Egypt” in Greek Albania (Fraser, 1992). However, it is also likely that the foreign immigrants were simply deemed to be Egyptians due to their physical appearance. This incorrectly ascribed Egyptian origin is believed to provide an explanation for the etymology of the Greek word, *Astinagoi*, the English word, *Gypsy*, and related appellations in other languages (eg. Tsigani, Gitano).

In the absence of a written history, evidence for an Indian origin of the Roma has been sought in a variety of social and cultural domains. Several scholars have claimed that the cultural practices of the Roma provide the most irrefutable evidence of Indian origins. Hancock (1999a) states that “[i]t is in the area of spiritual and physical well-being that the Indian origin of the Romani people is most clearly seen”. Marushiakova and Popov (1997) have claimed that the “Gypsy group…[provides] the most convincing evidence of the Indian origins” of the Roma. Other shared social and cultural practices with Indian populations are found in marriage customs, female warrior goddess worship (Shaktism), hygiene laws and, according to one author, a love of buffalo milk (Rishi, 1976). Similarly, parallels between musical styles used by populations in India to those of the Roma in Europe suggest at least cultural if not ethnic affinities (Gatliff, 1993).

Linguistic analyses have thus far provided the most robust and informative evidence regarding the Indian origin of the Roma. The similarities between Romany and Sanskrit were first noted over 200 years ago. Since then, linguistic analyses have been used to reconstruct the time and route of migration from India and to refine the origin of the Roma within Indian populations. Whilst Romany holds many similarities to Sanskrit, it also has similarities to more modern Indic and Dardic languages, indicating that Romany dates from post-Sanskrit times (Fraser, 1992). Hancock (2001a) has asserted that the distribution of genders in the Romany language points to a departure from India after 1000AD. However, others have concluded that the proto-Roma had moved out of India into Persian territories before 300 BC (Kaufman, 1984). Support for claims of an early Indian exit is weak and, although the time of departure from India remains unresolved, there is general agreement that it occurred around 1,000 years ago.

No historical records from India have ever been shown to describe the Roma and their emigration from India. Therefore, the reasons that the Roma left India, and with which populations in India they are most closely related, are the subjects of much speculation. One commonly retold scenario, which has entered popular folklore, contends that the Roma are descended from 10,000 musicians that were given as a gift to Bahram Gur, the ruler of Persia, in 439 AD (Fraser, 1992). Today, people known as the *Luri,* who speak an Indian-based language, live throughout the Middle East and are believed to be descendants of those Sindhian musicians (Hancock, 1999a). Whilst this has been proposed as an explanation for the origins of the Roma, others claim that the *Luri* are of no relation to the European Roma (Hancock, 1999a).

Within India, possible ethnic affiliations have been suggested to a nomadic people, the Dom, who are a caste of musicians (Fraser, 1992). However, this assertion appears to be based only on the superficial similarity between the word Dom and Rom. Moreover, Fraser (1992) himself states that "too often the assumption has been made…that any reference to a migrant group pursuing a Gypsy-like occupation can for that reason be equated with them".

Another hypothesis asserts that the Roma are derived from a high caste warrior group. Hancock (1999b) has rejuvenated this hypothesis, now over a century old, which claims that the Roma are descendants of an Indian military force comprised of a conglomerate of non-Aryan people. This military group was called the Rajputs, but also contained individuals from the Lohars, Gujjars, Tands, and Siddhis (Hancock, 1999b). In addition, Hancock (1999b) proposes the inclusion of East Africans immigrants in the military force. The aim of the assembled force was to fight off the incursion of the Islamic forces of Mahmud of Ghazi. This campaign was eventually unsuccessful and the Rajputs were forced to exit India via the Hindu Kush. As they moved further from India, class distinctions became less clear and the traditional *jatis,* the sub-caste groupings, all but disappeared. Thus, the Roma reached Europe as a socially cohesive group of people of relatively diverse ethnic origins.

As evidence for this claim of heterogeneous population origins, Hancock (2000) has argued that the Romany language is a *koine,* a product of the mixing of linguistic subsystems, which emerged outside of India. Though a military origin of the Roma is

not widely accepted, there is general consensus that the proto-Roma were ethnically diverse. Indeed, it has been asserted that "[m]ost scientists think that Gypsies belonged to the lowest social layer in their homeland and did not constitute a separate ethnic group" (Marushiakova & Popov, 1997). Claims of heterogeneous origins of the Roma present the possibility that a number of independent migrations by Indian populations may have occurred. However, this theory has not been widely promoted amongst Romani scholars.

Linguistic analysis has been applied to reconstruct the migration route followed by the Roma from India to Europe. Hancock (1999a) has summarised a possible route out of India on the basis of acquired linguistic features found in present-day Romany (figure 1-1). This reconstruction relies on the assumption that no major changes in linguistic territories have occurred. According to work performed by Hancock and others, the migrants passed through the Hindu Kush, along the southern shoreline of the Caspian Sea and through Persian linguistic territory. Fraser (1992) suggests a prolonged stay in Persia based on the presence of a significant number of Persian loanwords in Romany. Conversely, the absence of any Afghani contribution to the language provides evidence of this region having been avoided (Lee, 1998). This is compatible with claims that the Roma were a military force as Afghani territory was home to the Islamic invaders. The migrants continued through the southern Caucasus spending a considerable amount of time in Armenia, as evidenced by the significant number of Armenian loanwords in Romany (Fraser, 1992). The invasion of Armenia by the Seljuk Turks possibly provided the impetus to move into the Asian regions of the Byzantine Empire (Marushiakova & Popov, 1997) and further still into Europe (Fraser, 1992; Marushiakova & Popov, 1997). A large Medieval Greek contribution points to the likelihood that considerable time was spent within the Byzantine Empire before the Roma dispersed throughout Europe (Fraser, 1992; Hancock, 1999a).

Competing claims that "Gypsies" are a behaviourally defined segment of the European population have been alluded to since the early 1500s (Hancock, 2001b). The anthropologist Judith Okely (1983), has asserted that the Roma represent displaced serfs from the agricultural revolution. The contention of this theory is that peasant workers who were unable to adapt to the upheaval of the agricultural revolution assumed a

nomadic existence. The hypothesis of indigenous European ancestry of the Roma has found support amongst linguists (Wexler, 1997) and other scholars (Sandland, 1996) who are critical of the evidence of Indian origins. Okely (1983) claims that the Indian origins have been used to provide the Roma with a mythical charter and are spawned by the exoticisation of the Roma by non-Romani researchers. The implicit premise of this theory is that the Roma stem from an indigenous European social class rather than a non-European ethnic group. It is possible that this theory is symptomatic of the complexity of European itinerant groups. Several other populations are found in Europe, besides the Roma, who lead a nomadic existence. These include groups such as the *Travellers* in Ireland, the *Tartars* in Scandinavia, the *woowagenbewoners* in Holland, and the *quinquis* of Spain (Fraser, 1992). Contacts between the Roma and some of these groups are evident from the existence of Romani loanwords in their vocabulary. However, whether such groups are genetically related to the Roma is not known. Therefore, a theory of European origins of the Roma may be a result of incorrect extrapolations from findings in other itinerant groups, since much of Okely's work is based in the United Kingdom where there may be "Gypsies" who are not related to the Roma.

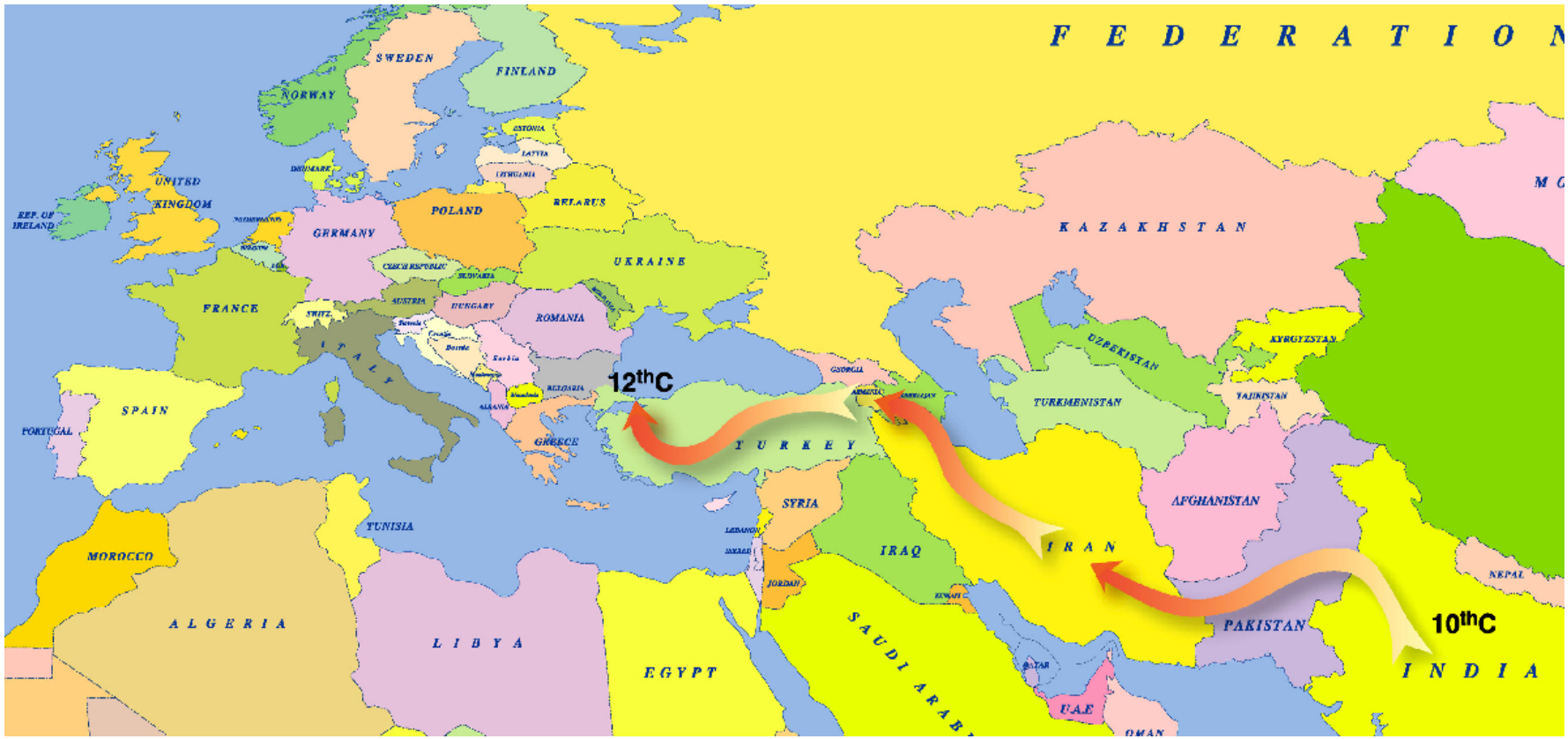

Figure 1-1 Migration Route of the Roma. The migration route of the Roma from India to Europe has been reconstructed on the basis of linguistic analysis. Extended stays in Persia (Iran) and Armenia are supported by significant representation of these languages in Romany.

1.1.3 Early Historical Records of the Roma in Europe

The earliest possible reference to the Roma in Europe comes from monastic records in Constantinople in 1068 AD (Fraser, 1992). However, this is not consistent with the suggestion that the Roma fled Armenia following the invasion by the Seljuk Turks in 1071 AD (Fraser, 1992; Marushiakova & Popov, 1997). The next historical reference that clearly refers to the Roma in the Byzantine capital dates to the 12th century (Fraser, 1992). Reconstructions from historical records suggest the Roma proceeded into Thrace and Greece and then further into the Balkans. Church records point to the settlement of the Roma throughout the Balkans in the 13th and 14th century with possible earlier incursions (Marushiakova & Popov, 1997). In Serbia, the earliest historical records of the Roma date to 1348 AD and 1362 AD (Fraser, 1992). In 1378 AD, the Bulgarian Tsar is recorded as giving some villages, partly inhabited by sedentary Roma, to the Rila monastery (Fraser, 1992). The Roma are first described in the Danubian principalities of Wallachia and Moldavia in 1385 AD (Fraser, 1994). The first tax registry taking the Roma into account in Rumelia (the Balkan provinces of the Ottoman Empire) dates from 1475 (Marushiakova & Popov, 1997).

Based on historical records, the penetration of the Roma into Western Europe appears to have begun during the 15th century. The Roma are first described in German records in 1407 AD (Hancock, 1991) and in France in 1419 AD (Fraser, 1992). The first document providing evidence of the Roma in Spain is dated 1425 AD, when King Juan II of Aragon provided a pass for travelling Roma (Fraser, 1992). Thus, by the mid-fifteenth century, historical records indicate the presence of Roma throughout Western Europe. It is apparent that this movement was followed by more widespread migrations into Europe, as the Roma are first mentioned in England and Poland-Lithuania in the early 1500s (Patrin, 1999). Historical records from this early period of residence in Western Europe invariably describe the Romani populations as comprising 30-400 people lead by a "Duke" or "King" and presenting letters of Imperial (or even Papal) safe conduct, which introduced them as penitents wandering the world to expiate their sins (Fraser, 1992). Fraser (1992) has described this period as a sort of reconnaissance conducted by numerous bands with a seeming "unity of action and close connection with

each other". The historical descriptions of small groups are of particular relevance for genetic studies, as this population segmentation provided the template from which current population structure was forged.

Early historical descriptions in Europe do not record the Roma as being Indian migrants. Where their origins are recorded they are usually described as being Egyptian. However, India was certainly not an unknown entity to Europe at that time. Indeed, the 1492 voyage of Columbus set out with the express aim of finding a passage to India. Contact between India and the Phoenicians dates to as early as 925 BC (Rawlinson, 1975). The earliest contact between India and Greece occurred about 510 BC and Indians formed part of the Persian military force that invaded Greece in 480 BC (Rawlinson, 1975). In 305 BC, the marriage between a member of Indian royalty and a Greek princess cemented a political alliance (Rawlinson, 1975). Thus, contact between Europe and India would have existed for well over a thousand years before the proposed arrival of the Roma in Europe. It is perplexing that Indian immigrants were not recognised as such, despite the social and cultural evidence.

1.1.4 History of Roma in Europe: 1500AD to Present

A historical overview of the last 500 years of the Roma in Europe would be extensive and is beyond the purpose of this thesis. However, several historical events can be expected to have profoundly impacted on the genetic composition of the Roma and are outlined below.

The present day distribution of the Roma can be considered as the product of four major migrations (figure 1-2). The first was the arrival of the Roma in the Balkans during the 12th and 13th centuries. Roma who have remained there ever since are referred to as the Balkan Roma. This was followed by the migration of Roma into Western Europe, during the 15th and 16th centuries. The third major migration followed the emancipation of the Roma from slavery in Wallachia and Moldavia in the mid-nineteenth century (Marushiakova & Popov, 1997). During this migration, Roma moved south into the Balkans and west into Europe in a process that resembled the initial westward migration of the Roma some 400 years earlier (Fraser, 1992). The populations that have migrated from the Danubian principalities are known as the Vlach

Roma. A fourth major migration has occurred in response to the political upheaval in Eastern Europe during the early 1990s and continues to this day. During this period, Roma have again moved into Western Europe and beyond to North America and Australia. These four migrations provide a historical framework in which the gross demography of the Roma can be framed. However, they should be understood as amplifications of a continual process of population redistribution through migration.

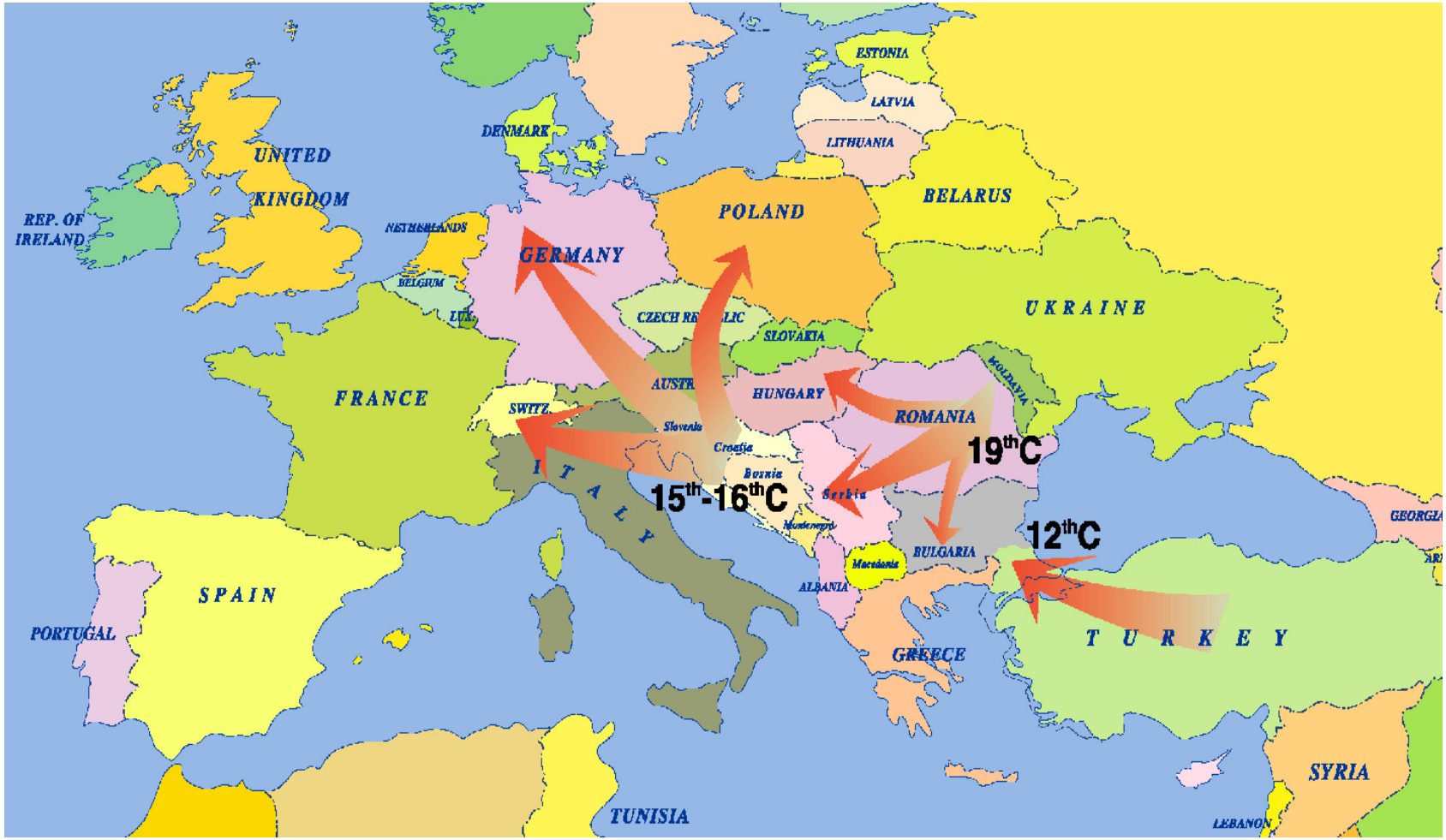

Figure 1-2 Major migrations of the Roma within Europe. These migrations have shaped the main population groupings of the Roma into Balkan, West European and Vlach Roma.

From at least as early as the 15th century penetrations into Western Europe, the Romani population has been undergoing a process of population fission. Some Romani groups sedenterised soon after their arrival in Europe and others settled at different stages thereafter. Many Romani groups, however, have led a nomadic existence for the majority of their time in Europe. The adoption of a mobile existence and the practice of specialised trades, such as metal-working or horse-trading, allowed them to coexist with the often hostile macro-society. This mode of existence, in which population size was kept small and mobile, is possibly a result of the persecution that the Roma faced throughout much of Europe beyond the Ottoman Empire. Indeed, their very "[s]urvival rested upon separating into small groups and living unobtrusively on the edges of the *gadzikano* (i.e. non-Roma) society" (Hancock, 1991). The period spanning 1550 to the late 1700s marked a distinct hardening in attitudes toward the Roma and the introduction of laws that would have resulted in their extermination if they had been successfully implemented (Fraser, 1992). During this time the Roma were variably interned, expelled, forced into servitude, forcibly assimilated or killed for the crime of being a Roma. It is an interesting historical footnote that in this period one of the policies adopted by both Spain and Portugal was the shipping of many of their resident Roma to their nascent colonies in South America and, in an ironic practice, to the Portuguese colony of Goa in Western India (Fraser, 1992).

In contrast to the maltreatment of the Roma in Western Europe during this time, Roma within much of the Ottoman Empire were left relatively unmolested (Fraser, 1992). The Ottomans remained largely uninvolved in the affairs of their subject states. Whilst this meant that there was no legislature discriminating against Roma in much of the Balkans, slavery in the vassal states of Wallachia and Moldavia continued unabated (Crowe, 1991). Romani slaves in these two principalities fell into three categories; slaves of the Crown, slaves of the monasteries and slaves of the estate owners (Fraser, 1992). The slaves of the Crown comprised numerous groups including the *Ursara* (bear-trainers), the *Lingurara* (wood-workers), the *Aurara*/*Rudara* (gold miners), and the *Laiesi* who had no fixed occupation and were able to roam the principalities (Fraser, 1992). Many of the contemporary divisions of Vlach Romani groups are derived from the names given to them during this period (Crowe, 1991). It is difficult to know how

many Roma were enslaved during this period, but one estimate from the 19th century reported 200,000 slaves in Wallachia and Moldavia (Fraser, 1992).

The oppression and marginalisation of the Roma reached its zenith in Europe during the Second World War when the Nazis selectively murdered people of Romani ethnicity in Germany and the occupied territories. This genocide, called the *Pojaramos* (the Great Devouring) by the Roma, was undertaken with the aid of "scientific" investigations headed by anthropologist Dr Robert Ritter of the "Research Centre for Racial Hygiene and Population Biology", who constructed detailed pedigrees demonstrating Romani heritage (Müller-Hill, 1998). In 1940 a report by Ritter stated that "[t]he Gypsy question can be considered solved when the main body of asocial and good-for-nothing Gypsy individuals of mixed blood is collected together in large *labour camps* and kept working there, and when further *breeding of this population of mixed blood* is stopped once and for all" (Müller-Hill, 1998, author's italics). This policy of population annihilation was pursued by the Third Reich, which forcibly sterilised Roma and sent them to their deaths in the concentration camps. The determination of one-eighth Romani ancestry (i.e. a single great-grandparent) was sufficient to condemn an individual to the death camps (Heuss, 1997). However, Ritter argued for the preservation of "pure Gypsies" for further studies (Müller-Hill, 1998). It is not known how many Roma were murdered during this time. Estimates range from 250,000–500,000 individuals (Fraser, 1992), to claims that up to two-thirds of the Romani population in Europe was killed (Patrin, 1999).

1.1.5 Historical Demography of the European Roma

Historical data on the number of Roma are scant and generally obscured by a lack of specificity. However, records exist that provide rough indications of the numbers of Roma in different regions. The 1522-1523 tax registry in Rumelia reported 10,294 Christian and Muslim Romani households, which equates to an approximate total population of 66,000 Roma in the Balkans (Marushiakova & Popov, 1997). In 1695 there were 45,000 tax-paying Romani males in the Ottoman Empire, which extended from Mesopotamia to the Balkans (Fraser, 1992). This would be equivalent to a population totalling 225,000 individuals in the Ottoman Empire, assuming an average

family size of three children. A 1780-1783 census in Hungarian territories (which included Croatia and Slovenia, but not Transylvania) placed the total number of Roma at 30,241-43,609 (Fraser, 1992). In 1785, there were 12,000 Roma in Spain with two-thirds of these resident in Andalusia (Fraser, 1992). A scholar in 1783 estimated that there were 700,000-800,000 Roma in Europe (Fraser, 1992). Thus, the current Romani population in Europe, totalling 7-8 million would represent a 10-fold increase over roughly 200 years. Extrapolating this rate of growth to earlier years, one would arrive at population sizes of 70,000-80,000 in 1583, 7,000-8,000 in 1383 and 700-800 in 1183, around the time of their arrival in Europe. The simplifying assumption of a continual rate of population growth is no doubt incorrect; however, alongside recorded population sizes it suggests that the Romani founding population in Europe was small in number.

1.2 Anthropology of the Roma

1.2.1 Anthropological Classifications of Extant Romani Populations

The populations subsumed under the ethnonym "Roma" are multiple and varied. Furthermore, numerous populations exist within Europe that are typically described as "Gypsies" but are distinguished from the Roma. In many cases, the historical and cultural bases of these ethnological distinctions are unclear. Therefore, attempts to classify populations and to identify relationships between them are obscured by semantic uncertainties. The anthropologists Marushiakova and Popov (1997) have described the Roma as "a disperse transboundary minority ethnic community" which represent a "constellation of groups". Classifications of these groups are based on criteria such as language and dialect, religion, traditional trade, self-appellation, rules of endogamy and lifestyle (i.e. nomadic or sedentary). Many of these criteria correlate with historical migrations and the subsequent different histories of the populations. Some broadly classificatory regional names are used: the *Sinti* of Germany, the *manouches* of France, the *Cale* of Spain, the *Ciganos* of Portugal and the *gitans* of southern France. However, in general, nationality is a poor criterion for classifications (Fraser, 1992).

Language is a major parameter used to classify Romani populations. On the basis of language it has been proposed that Romany can be classed into four metadialects. These include the Balkan dialects, the Vlach dialects of Rumanian influence, the Carpathian dialects of Hungarian and Serbian influences, and the Nordic dialects of mostly German influence (Tcherenkov & Laederich, unpublished manuscript). This classification by metadialect is compatible with population divisions dictated by major migrations. Religion, whilst generally considered an inaccurate parameter for defining Romani groups, is used to distinguish the Xoroxane Roma, who are Muslims, from Christian Roma.

Such classifications are necessarily generalising and, to a large degree, the historical relationships between populations subsumed under the groupings are unclear. It is apparent that the Roma have not existed as a monolithic population for over 500 years. In 1775-1776, a Hungarian scholar pointed out that "although all Gypsies have many features in common, there was no longer a homogeneous Gypsy nation or collective culture" (Fraser, 1992). This heterogeneity is apparent within individual European countries. In the former Yugoslavia at least twenty groups were identified (Fraser, 1992), and it is normal to come across three or more groups in one town (Boretzky, 1995). The meticulous chronicler of Bulgarian Romani groups, Gilliat-Smith[1], reported at least 19 distinct tribes in the north-east of Bulgaria around the time of the First World War ("Petulengro", 1915-1916), and current estimates suggest as many as 50 groups in Bulgaria (Marushiakova & Popov, 1997).

1.2.2 Salient Social and Cultural Features of Romani Populations

Numerous social and cultural features of the Roma distinguish them from other European peoples. Several of these traditional practices can be predicted to impact on the genetics of the population, particularly their marriage patterns and customs. A typical feature of Romani groups is the strict adherence to endogamic marriages (Marushiakova & Popov, 1997). For Romani groups, the maintenance of the "Group" is of primary importance and its purity is preserved through admission to the group only

[1] After spending a large proportion of his life studying the Roma, the British ethnographer, Gilliat-Smith adopted the Romany name, Petulengro.

through birth (Marushiakova & Popov, 1997). Thus, the selection of marriage partners from within the group ensures the preservation of a cohesive identity. This practice restricts marriages between members of different Romani groups and between Roma and non-Roma. It is generally asserted that first cousin marriages are forbidden in Romani culture (Rishi, 1976). However, consanguineous unions are common amongst many Indian populations (Bittles, Mason, Greene, & Rao, 1991). The stringent practice of endogamy can be expected to entail the marriage of close relatives, and studies that have examined marriage types within a Slovak Romani population report high coefficients of inbreeding (Ferák, Gençik, & Gençikova, 1982; Ferák, Sivaková, & Sieglova, 1987). Whether this reflects a preference for close kin marriage or a restricted choice of marriage partners is unclear.

In traditional groups, marriage contracts are often arranged and bride prices are paid in a similar manner to Indian practices (Marushiakova & Popov, 1997). Marriages generally occur at a very early age: between 13-16 years for females and between 15-21 years for males (Marushiakova & Popov, 1997). After the marriage it is customary for the couple to live with the family of the husband (Marushiakova & Popov, 1997). Belonging to large families is highly regarded, as is the bearing of a large number of children.

1.2.3 The Roma in Bulgaria

In Bulgaria, the Roma are known as the *Tsigani*, which is derived from the Greek, *Astinagoi*. Their initial arrival can be inferred as occurring sometime in the 13th or 14th century (Marushiakova & Popov, 1997). The first Ottoman tax register mentioning the Roma in Bulgarian lands was in 1475 and information including religion, occupation and areas occupied by the Roma has been recorded ever since (Marushiakova & Popov, 1997). Early records tell of Christian Roma in the Ottoman Empire, which indicates their presence prior to the Ottomans (Marushiakova & Popov, 1997). Tax registers record that many Roma were employed as town blacksmiths or musicians, however, over time more Roma became sedenterised and abandoned their traditional trades for new employment and for farming. Whilst the Roma were not

officially slaves in Bulgaria, in many cases they led a life of servitude (Marushiakova & Popov, 1997).

In Bulgaria, the 1522-1523 tax registry reported a total of 5,700 Roma and the 1881-1885 census data report 62,324 Roma (Marushiakova & Popov, 1997). Data from subsequent censuses report 99,004 Roma in 1905, 122,296 in 1910, 134,844 in 1926, 170,011 in 1946 and 197,805 in 1956 (Marushiakova & Popov, 1997). This suggests a two-fold increase in population over 50 years. The current Romani population of Bulgaria, believed to be 700,000-800,000 (Liégeois, 1994), would represent a four-fold increase in the following 50 years. However, a review of censuses from a three-year period between 1989-1992 reveals serious discrepancies in results due to the differing and sometimes arbitrary criteria used to define a person as a Roma (Marushiakova & Popov, 1997).

The demography and composition of the Bulgarian Romani population has been shaped to a large extent by the major migrations of Romani groups previously described. However, owing to the close geographical proximity to Rumania, there were migrations of Vlach Roma into Bulgaria prior to the end of slavery. These groups are generally characterised as speaking old-Vlach dialects (Marushiakova & Popov, 1997).

1.2.3.1 Social anthropology of the Bulgarian Roma

The Bulgarian Romani population was the subject of extensive ethnographic and linguistic investigations at the turn of this century by the British-born researcher B. J. Gilliat-Smith (Petulengro). Having spent four years in the Bulgarian city of Varna, Gilliat-Smith classified "Gypsy tribes inhabiting the Balkan Peninsula" based on "(1) the district, (2) the religion, (3) the mode of life – whether sedentary of nomadic, [and] (4) the occupation or trade" (Petulengro, 1915-1916). Using this taxonomic system, Gilliat-Smith (1915-1916) identified 19 individual groups in Northeast Bulgaria.

These fundamental criteria for classifying Romani groups have been used more recently by Marushiakova and Popov (1997). They have expanded the criteria to create a more comprehensive taxonomy that includes the preferred self-appellations of the group, the time of arrival in Bulgaria and endogamy rules as important differentiating characteristics. The authors have devised a classificatory system that delineates three

main Romani metagroups widely dispersed throughout Bulgaria. A metagroup "combines several groups with infringed boundaries" (Marushiakova & Popov, 1997) and is the overarching population structure. Groups within a metagroup are generally mutually endogamous, however, rules of endogamy within a metagroup can collapse when subjected to external forces (Marushiakova & Popov, 1997). The three Romani metagroups found in Bulgaria are:

1. Jerlii

The *Jerlii* are descendants of the first Roma to settle in Bulgaria. They are the most numerous and diverse metagroup in Bulgaria. The majority of the *Jerlii* abandoned a nomadic existence during the time of the Ottoman Empire. Linguistically, this group is characterised as speaking Turkish or Romany dialects; which are classed as Balkan or non-Vlach dialects. The *Jerlii* are subdivided into two main groups – the *Dassikane Roma* who are Christians and the *Xoroxane Roma* who are Muslims. These are still broad definitions and contain numerous well-preserved traditional groups, which practice strict endogamy and are clearly delineated. The *Jerlii* are scattered evenly throughout Bulgaria, with the bulk of the *Xoroxane* population in the northeast of Bulgaria.

2. Kalderash

The *Kalderash* (*Kaldarasi*) are the descendants of groups who left Wallachia, Moldavia and Transylvania during the *Great Kalderara Invasion* of the second half of the nineteenth century following their emancipation. They are subdivided into the *Lovari* and *Kelderari*. The *Kalderash* speak their own Vlach dialects, which are also known as Stratum III of the Balkan dialects. They are largely Orthodox Christians. They were nomadic until 1958 when they were forcibly settled by the Bulgarian government. Their adherence to traditional practices often leads them to assert that they are "Gypsy Gypsies". The *Kalderash* have adopted an extended endogamy that comprises the entire metagroup and which extends beyond the borders of Bulgaria. The *Kalderash* avoid co-residence with other Roma and do not form distinct neighbourhoods. They generally exist in small groups of 10 to 15 families, dispersed among the surrounding population.

3. Rudari/Ludari

The *Rudari* (or *Ludari*) represent the third major Romani group in Bulgaria. They also refer to themselves as *Wallachians* or *Wallachs* as they were enslaved in Wallachia and Moldavia. This group entered Bulgaria after the end of slavery in the 19th century. They speak an ancient form of Rumanian and adhere to Eastern Orthodox Christianity. This metagroup is further sub-divided into the *Lingurari* (spoon-makers) and *Ursari* (bear-trainers). The *Lingurari* prefer to live in villages by rivers or in the mountain foothills. These geographically differentiated groups are known as the *Intreni* and *Monteni* respectively. The *Ursari* are spread widely throughout Bulgaria.

A modified schematic outline of Marushiakova and Popov's stratification of Bulgarian Roma groups is presented in figure 1-3. Contact between members of these three metagroups is virtually non-existent and they have little to do with each other's affairs. Marriages between individuals from the different metagroups are discouraged and extremely rare. Within the three broadly defined metagroups there is a myriad of well-preserved subgroups. The classifications are based on practiced trades (both former and current), region of residence and kinship ties. Strict endogamy is observed within many of these groups; however, a complex system of regulations dictates permissible marital partners who may originate from outside the immediate group. In some groups it is more acceptable for non-Roma to marry into the population than Roma from other groups. It has been asserted that Romani "communities are even more exclusive than other ethnic communities" (Marushiakova & Popov, 1997). In Bulgarian Romani populations, adherence to strict endogamy appears to have been practiced for a long time, and thus can be expected to have impacted on the genetic structure and diversity within these populations.

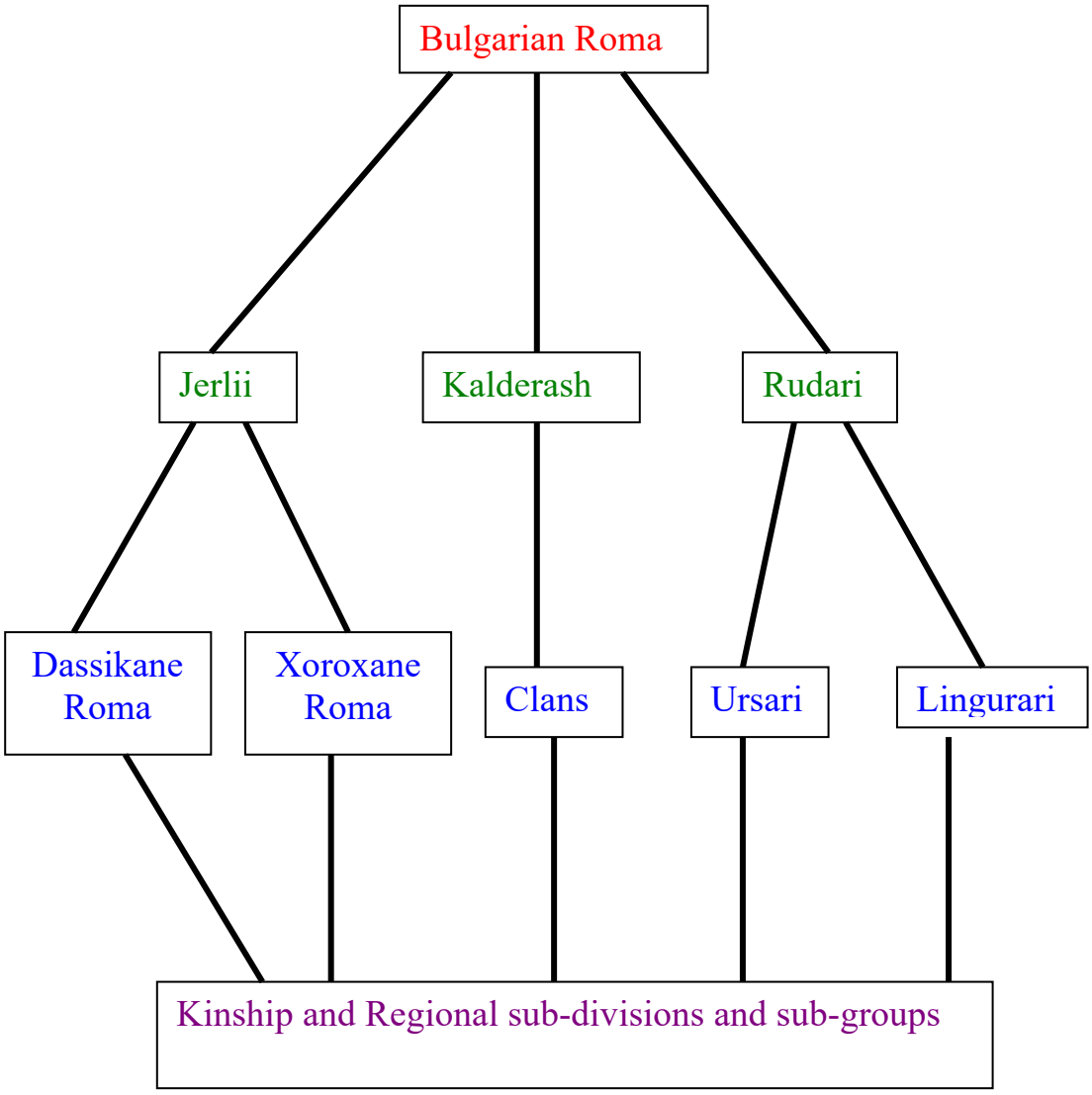

Figure 1-3 Anthropological classification of the Roma in Bulgaria (adapted from Marushiakova and Popov, 1997). The three metagroups (Jerlii, Kalderash and Rudari) are mutually endogamous. Complex social rules dictate marriage patterns within the metagroups.

1.3 Population Genetics of the Roma

1.3.1 Population Genetic Studies of the Roma

The investigation of polymorphic heritable markers in Romani populations has been undertaken for over eighty years. Researchers have examined Roma throughout Europe, although the majority of studies have examined Romani populations from Eastern Europe. A survey of relevant publications yields at least 30 independent studies of the genetics of Romani populations. Typically, investigators have used classical polymorphisms: blood group systems, enzyme polymorphisms and the human leukocyte antigen (HLA) system. The questions that these studies have endeavoured to answer invariably fall into three broad categories: the biological relationship between neighbouring Romani and non-Romani population, the relatedness of geographically separated Romani populations and the ethnic origins of the Roma. The issues explored are not only of interest to population biologists, but also to scholars of Romani history, anthropology and culture, and the Roma themselves. Indeed, many general texts and discussions on the Roma cite the evidence provided by genetic studies to support their own arguments (Rishi, 1976). Even amongst those authors who have eloquently criticised many aspects of genetic studies of Roma (Kohn, 1996), the overall conclusions have not been questioned. Thus, the assertions that the Roma are genetically distinct from other European populations and that genetic studies provide "scientific" proof of their Indian ethnicity have come to be widely accepted.

The vast majority of population studies on the Roma have employed the polymorphic blood group systems (serogenetic markers). Of these studies, the greatest body of comparable data has been collected on phenotypic variation in the ABO, Rhesus (Rh), MN and haptoglobin systems (table 1-1). Generally, the frequency distributions of phenotypic variants in each of these systems have been compared with other populations and conclusions drawn on the basis of observed differences and similarities. Using this piece-meal approach, many studies of Romany populations are characterised by confounding and seemingly contradictory results from the different systems. The cause of these discrepant results remains unclear. A possible explanation is the effect of genetic drift acting in small populations, which can significantly skew the distributions

of polymorphic traits. Moreover, drift will act independently on the different polymorphic systems, affecting the variant frequencies of different systems in markedly different ways. Different genetic heritages and different sources of admixture could also yield the observed diversity. It is likely that study designs that fail to address population history or social structure have incorrectly defined Romani populations. This would be expected to further confound findings and interpretations from these studies. Thus, an examination of previous population genetic studies requires careful consideration of study design and the methodologies employed. Nonetheless, examination of data generated from these studies provides some insights into Romani populations.

Table 1-1

Phenotypic frequencies of polymorphic variants in Romani populations as reported in relevant literature

Author	Population	Sample size	ABO blood group					Rhesus blood group		MN blood group			Haptoglobin		
			A_1	A_2	AB	B	O	Rh+	Rh-	MM	MN	NN	1-1	2-1	2-2
Rex-Kiss et al. 1973	Hungarian	507-600	30.8	3.5	10.7	28.0	27.0	89.1	10.9	31.0	49.8	19.2	5.8	29.2	64.6
Bartsocas et al. 1979	Greek	200	29.0		11.5	32.0	27.5	91	9	47.5	40.5	12.0			
Harper et al. 1977	Welsh	84	35.7	7.1	0	14.3	42.8	95.7	4.3	22.9	47.1	30.0	3.9	39.5	56.6
Bernasovsky et al. 1975	Slovakia	2935	32.9	2.4	9.3	25.2	30.2	89.5	10.5	27.2	51.6	21.2			
Sivakova et al. 1994	SE Slovakia (SGP1)	50	34	2	20	26	18	88	12	56	28	16	14	22	64
	SE Slovakia (SGP2)	51	41	2	16	21.6	20	94	6	31.4	47	24.6			
Bernavosky et al. 1994	Wallachians in Slovakia	119	20	4	0.8	6.7	68	80.7	19.3	44.5	51.4	5	10.1	40.3	49.6
Beckman and Takman 1965	Sweden	116	59.48		3.45	6.03	31.03	95.7	4.3	20.0	55.7	24.4	0	23.5	76.5
Clarke 1973	Britian	109	33.03	14.68	4.58	8.26	39.45	85.3	14.7	49.5	41.3	9.2	15.5	52.4	32.0
Avcin 1969	Slovenia	350	49.42		5.99	10.28	34.28	85.1	14.9	19.1	46.9	34.0			
Galikova 1969	Slovakia (east)	180											1.1	21.2	77.9
Galikova 1969	Slovakia (west)	180											2.2	32.2	65.6
Cazal et al. 1951[ψ]	France	113	23.0	2.70	14.1	38.1	22.1	85	15	36.0	46.5	17.4			
Ely 1961[ψ]	France (north)	47	21.0		6.0	19.0	53.0	97.6	2.4						
Ely 1966[ψ]	France (south)	41	41.0		10.0	27.0	22.0	87.8	12.2						
Nicoli and Sermet 1965[ψ]	France (south)	92	40.0		7.6	9.7	42.0	87.6	12.4						
Schmidt 1930[ψ]	Yugoslavia (pop 1)	299	23.0		7.7	24.4	45.0								
	Yugoslavia (pop 2)	126	40.0		7.1	19.8	33.0								
Verzar and Werszecky 1921[ψ]	Hungary	385	21.0		5.7	39	34.2								
Libman 1930[ψ]	Uzbekistan	104	42.3		9.6	28.8	19.2								
Hesch 1930[ψ]	Romania	102	27.4		8.8	37.2	26.5								

[ψ]values calculated from data in Harper *et al*., 1977

1.3.1.1 A critique of sampling methodologies used in population genetic studies of Roma

Poor sampling techniques have adversely affected the majority of genetic studies of Romani populations. The fundamental flaw of many of these studies is the application of inappropriate classificatory criteria to the study population, most notably nationality. In cases where researchers have defined the population, they have relied either on folk-taxonomic classifications of Romani populations (eg Clarke, 1973 classifies British Roma into Romanies, Posh-rats, Didakis and Travellers according to the amount of "Romany blood" in the individual) or crude regional classifications (eg. Bernasovsky, Suchy, Bernasovska, & Vargova, (1976) studied "East Slovak Gypsies"). In some cases, where anthropological distinctions between Romani groups have been made it has been under the notion that the degree of social assimilation with the macro-society (eg. Galikova, Vilimova, Ferák, & Mayerova, [1969] define assimilated, semi-assimilated and unassimilated Romani groups) and choice of residency (Rex-Kiss, Szabo, Szabo, & Hartmann, 1973) is of biological relevance. Notions of racial purity in the selection of individuals for investigation have been even more disturbing. Researchers have been insistent that they have sampled only "genuine Gypsies" (Beckman & Takman, 1965) or individuals of Romani lineage (Clarke, 1973; Harper, Williams, & Sunderland, 1977) although the criteria used for these distinctions are unclear. These inaccurate and inappropriate classifications used by population geneticists stand in marked contrast to the great complexity of Romani population structure as demonstrated by historians, linguists and anthropologists.

Authors have generally omitted details of the circumstances under which biological samples were obtained. However, two papers state that samples were obtained from Romani males in jail (Rex-Kiss et al., 1973; Sivaková, 1983). The use of prisoners as research subjects introduces a sampling bias to the study, whilst implying the prejudiced notion that a jail would be the logical place to obtain samples from Roma. Furthermore, it is unclear whether these studies had been undertaken with the informed consent of subjects.

1.3.1.2 Genetic evidence for the relatedness of Roma with other European populations

A number of studies have aimed at comparing the genetic composition of Romani populations with that of indigenous European populations. Generally, these comparisons have involved an examination of each independent polymorphic system for differences and similarities observed between populations. Typically, autochthonous European populations are characterised by a relatively low frequency of the B blood group (Mourant, Kopec, & Domaniewska-Sobczak, 1976). Thus, the ABO blood group system has been the primary means of comparing the Roma with other European populations. The conclusions offered by these results have not always been clear. Moreover, when other polymorphic systems are examined they often suggest contradictory conclusions.

In the Welsh Roma, the frequency of the B blood type is similar to that found in the non-Romani Welsh population (Harper et al., 1977). However, the two populations have different phenotypic distributions of the Rh, MN and haptoglobin systems (Harper et al., 1977). Alternatively, in a large Slovak Roma sample, the MN system shows no significant differences between the Romani and non-Romani populations, whereas statistically significant differences are seen using the ABO and Rh system (Bernasovsky et al., 1976). Similar discordance in the results using different polymorphic systems has been observed in the Hungarian Roma (Rex-Kiss et al., 1973) and a study of Slovak Roma (Sivaková, Sieglova, Lubyova, & Novakova, 1994). In the Greek Romani population, a very high frequency of B blood type was observed which differs significantly from the autochthonous population (Bartsocas et al., 1979). Differences between the two populations were also observed using the Rh and MN systems, however the distribution of phenotypes using the Kell and Duffy system was identical in the Romani and non-Romani populations (Bartsocas et al., 1979). Similarly, a high frequency of the B blood type distinguishes another sample of Hungarian Roma from the autochthonous population (Tauszik et al., 1985). Differences between the two populations are observed in the Rh and MN systems but not in the P and Kidd systems (Tauszik et al., 1985). Whereas these studies have found a much higher frequency of the B blood group than that found in the local population, a study of Slovenian Roma

determined that the B blood type frequency was some 4-8% less than in the non-Romani Slovenians (Avcin, 1969). Significant differences between these two populations were also apparent using the MN and Rh systems (Avcin, 1969).

Using the HLA system, Gyodi et al., (1981) found that the most frequent haplotypes in the Roma occur at low frequency in Hungarians and conversely the most frequent haplotype in Hungarians is absent in Roma. A study of Spanish Roma by de Pablo et al., (1992) based on HLA data showed genetic distances between the Roma and autochthonous European populations were large compared to those observed between autochthonous populations. Genetic distance and principal component analyses of blood group data from various ethnic groups in Hungary demonstrated the genetic separateness of Hungarian Roma from all other populations in the country (Guglielmino & Beres, 1996).

Thus, it is clear that throughout Europe, Romani populations are consistently found to be genetically distinct from populations alongside whom they have resided for many hundreds of years. This indicates that the Roma are genetically isolated from other European populations. Whether this is due to social isolation and genetic drift or to different ethnic origins is unknown.

1.3.1.3 Relationships between the Romani populations of Europe as revealed by population genetic studies

Given the large degree of socio-cultural heterogeneity exhibited by the European Roma, it is reasonable to expect significant differences at a biological level. Indeed, a comparison of the data collected from geographically separated groups points to significant biological heterogeneity (table 1-1). Attempts to identify genetic affinities between different Romani populations have generally used the methodological approach by which Romani populations were compared to autochthonous populations; that is, a comparison of similarities and differences observed in the marker frequencies of different systems in different populations. As such, these studies have encountered the same problems in terms of lack of consistency using different systems.

The majority of authors have utilised the ABO blood group system as a means of inferring genetic affinities between Romani groups from different countries, usually

emphasising the frequency of the B blood type. The frequency of this phenotypic variant can be seen to range from 6-38% (table 1-1). The somewhat arbitrary division between Roma groups that exhibit a B blood type frequency in the range of 20-40% and those groups in which the B blood type frequency is markedly lower (5-15%) has prompted one author to surmise that the Roma comprise two different populations (Clarke, 1973). Harper et al., (1977) determined a frequency of the B blood type of 10% in Welsh Roma, a result comparable to that found in French (Nicoli and Sermet, 1965), Swedish (Beckman & Takman, 1965), British (Clarke, 1973) and Yugoslavian (Avcin, 1969) Roma populations. A later study of Wallachian Roma in Slovakia, relatively recent immigrants from Rumania, reported a B blood type frequency of 6.7% (Bernasovsky, Halko, Biros, Sivakova, & Jurickova, 1994). Thus, populations with a low frequency of blood group B have been observed throughout Europe and do not conform to any geographic structuring. A comparison of the frequency distributions of other polymorphic systems in these populations do not support the relationships implied by the ABO system.

In contrast, a number of Romani populations are characterised by elevated B blood group frequencies. The Greek Roma studied by Bartsocas et al. (1979), have a B phenotype frequency of 32%, one of the highest values of any Romani population in Europe. However, the frequency of the M phenotype in the same population is most similar to that of the British Roma, who are characterised by a very low B type frequency. A similarly high B blood group frequency was found in Hungarian Roma (Rex-Kiss et al., 1973), who also have a comparable Rh distribution to the Greek Roma; however, the two populations differ dramatically in the distribution of MN variants. The large study of East Slovak Roma by Bernasovsky et al., (1976) found that phenotypic frequencies at the ABO, Rh and MN loci all corresponded closely with those found in Hungarian Roma by Rex-Kiss et al., (1973). These two populations provide the only example of complete concordance in results from different systems.

A limited number of studies have examined genetic markers in more than one Romani population within the same country. Galikova et al., (1969) investigated the frequency distributions of haptoglobin types in Romani populations from East and West Slovakia and found significant differences between the two. Sivakova et al., (1994)

investigated two different populations of Roma from southeast Slovakia and found that the two populations were genetically differentiable. The study of Wallachian Roma in Slovakia reported different phenotypic frequencies across all systems compared to those found in other Slovak Roma (Bernasovsky et al., 1994). Gyodi et al., (1981) used the HLA system to investigate two Romani groups in Hungary and found significant differences between them. These data were compared to HLA data from Spanish Roma in a later study by de Pablo et al., (1992). Genetic distances between the two Romani populations were greater than the distances between the Roma and all other European populations examined (de Pablo et al., 1992).

Mastana & Papiha, (1992) attempted to synthesise published data of blood group polymorphisms in Romani populations. They concluded, on the basis of genetic distance and principal component analyses, that significant heterogeneity exists in the European Roma and argued for a significant differentiation between Eastern European and Western European Roma (Mastana & Papiha, 1992). A reanalysis of data from classical markers by Kalaydjieva, Gresham, & Calafell, (2001) illustrates the large genetic distances between Romani populations. However, no clinal structuring is apparent. Thus, the vast variation in polymorphic systems observed in Romani populations throughout Europe indicates that they are best described as a conglomerate of genetic isolates. Attempts to identify genetic affinities between geographically separated groups by comparing the distribution of polymorphic traits have been inconclusive due to the approaches taken and the limited resolving power of the polymorphic systems used.

1.3.1.4 Relationships between Romani and Indian populations as revealed by genetic studies

Verzar & Weszeczky, (1921) were the first to point out that the high frequency of the B blood group in the Roma resembled the elevated frequency of the B blood group found in populations of the Indian subcontinent. A number of studies following this work have confirmed that a high frequency of the B blood group is characteristic of the majority of Romani populations (13 out of the 19 populations listed in table 1-1 have a B blood group frequency ≥ 15%). Mourant, Kopec, & Domaniewska-Sobczak, (1976)

in an attempt to summarise previous findings, declared that the "mean A and B gene frequencies of nearly 5,000 European Gypsies are each approximately 22 per cent, figures which are closely comparable to those of West Pakistan". Thus, the observation of a high B blood group frequency has been the primary basis for the assertion that genetic data confirm the Indian origins of the Roma. This conclusion has been supported by findings of a high M frequency in the MN system, low Rh (*d*) frequencies in the Rhesus system, and elevated Hp1 frequencies in the haptoglobin system, all of which are characteristic of northern Indian populations (Mourant, Kopec, & Domaniewska-Sobczak, 1976)

However, the summary of blood group frequencies in Romani populations points to striking genetic heterogeneity (table 1-1). The extremes of between-population phenotypic variation in B blood group frequencies are evident in the very low values seen in the Wallachian Roma population in Slovakia (Bernasovsky et al., 1994) and the Swedish Roma (Beckman & Takman, 1965). This is in direct contrast to those reported in French (Cazal, Graafland, & Mathieu, 1951) and Hungarian (Verzar & Weszeczky, 1921) Roma where values close to 40% have been observed for the B blood group frequency. Analyses of the MN system further illustrate the heterogeneity of Romani populations across Europe. Frequencies of the M phenotype have been found to be as low as 19% (Avcin, 1969) and as high as 56% (Sivakova et al., 1994). Similar variation between populations in the frequency of the Rhesus blood group phenotypes is evident, with Rh (*d*) frequencies ranging from 2% (Ely, 1961) to 19% (Bernasovsky et al., 1994) in Romani populations across Europe. The frequency of haptoglobin polymorphisms in Roma populations across Europe also exhibit significant variability, with the frequency of the Hp1 phenotype ranging from 15.5% in British Roma (Clarke, 1973) to its complete absence in Swedish Roma (Beckman & Takman, 1965).

A comparison of data from multiple polymorphic loci within Romani and Indian subcontinental populations illustrates a lack of consistency in the conclusions offered by the evidence. Within the Welsh Roma, the low frequencies of the B and M blood groups suggest no genetic affinities with Indian populations, and contrast with the Rh (*d*) and Hp_1 frequencies, which are of Indian magnitude (Harper et al., 1977). Investigations of British Roma provide no indication of Indian origins on the basis of ABO and

haptoglobin frequencies, however the high level of the M phenotype is of Indian magnitude (Clarke, 1973). In a study of the Swedish Roma, the frequency of the B blood group was found to be well below Indian values, whereas Rh and haptoglobin distributions were reported as being compatible with Indian origins (Beckman & Takman, 1965). Rex-Kiss et al., (1973) reported that the distribution of variants of the different blood group systems in Hungarian Roma was almost identical to those found in Pakistani populations, with the exception of the MN system. Bartsocas et al., (1979) claimed that frequencies in the blood group systems demonstrated similarities with the inhabitants of the Punjab and Western Pakistan, "especially with regards to the ABO, Rh and Duffy blood groups". Sivakova (1983) examined the distributions of red blood cell acid phosphatase (ACP), phosphoglucomutase (PGM_1) and adenylate kinase (AK) iso-enzymes in samples from Slovak Roma. Phenotypic distributions in the ACP and PGM_1 systems were compatible with data from North Indian populations, however the results from the AK system appeared to contradict the conclusion of a North Indian ethnogenesis. Others have claimed evidence for Indian origins of the Roma based on a single locus. Galikova et al. (1969) investigated the haptoglobin polymorphism distribution among Slovak Roma and concluded its distribution supported an Indian origin. Tauszik, Forrai, & Hollan, (1987) investigated the percentage of tasters of phenylthiocarbamide among Hungarian Roma and reported the high proportion of tasters provided further confirmation of Indian origins. An investigation of immunoglobulin allotypes in a Hungarian Roma population claimed that Indian and Pakistani populations were the only ones to contain all variants found in the Roma, thereby supporting claims of Indian origins (van Loghem, Tauszik, Hollan, & Nijenhuis, 1985).

Thus, claims that the frequency distributions of genetic polymorphisms provide evidence supporting Indian origins of the Roma are not credible when the entire body of data is considered. The genetic heterogeneity observed amongst populations and polymorphic systems within populations prevent any general conclusions. When genetic distances between Romani populations and Indian populations have been calculated on the basis of blood group polymorphisms, the analysis has been conducted without the inclusion of European populations (Mastana & Papiha, 1992), thus preventing a

comparison of the relative relatedness of Roma to Europeans with that of Roma to Indians. When data from the HLA system have been used to calculate genetic distances, distances between Indians and some autochthonous European populations are less than, or similar to, those between Indian and Romani populations (de Pablo et al., 1992). A population tree constructed from mtDNA, places Romani populations distant to all other populations, including Europeans and Indians (Kalaydjieva et al., 2001). Therefore, a survey of the relevant literature illustrates that the genetic evidence for the Indian origins of the Roma is weak.

1.3.2 Mendelian Genetic Disorders in the Roma

A number of mendelian disorders have been identified in the Roma. Of these, three are novel disorders and thus far appear to occur uniquely in the Roma. In addition, a number of known rare genetic disorders have been identified. In many cases, a single mutation has been identified as the predominant cause of the disorder. This suggests that founder effects have inflated the frequency of these disease alleles. In addition to these disorders, disease-causing mutations found in other populations have been identified in the Roma. The small number of studies into genetic disorders in different Romani populations offers some insights into the distribution of the apparently private mutations.

The three novel disorders found in the Roma have been initially described in Romani populations in Bulgaria. Hereditary motor and sensory neuropathy type Lom (HMSNL) was the first novel genetic disorder identified in the Roma. Kalaydjieva et al., (1996) reported the mapping of the disease locus to 8q24. The homogeneity of disease haplotypes led the authors to propose a putative founder mutation present in three socially separated Romani populations. Subsequently, a second novel disorder, the congenital cataracts and facial dysmorphism neuropathy syndrome (CCFDN) was reported (Angelicheva et al., 1999). The gene for CCFDN was mapped to 18qter and a conserved common disease haplotype indicated a founder mutation (Angelicheva et al., 1999). Recently, a third disorder, termed hereditary motor and sensory neuropathy type Russe (HMSNR), has been identified (Rogers et al., 2000). Haplotype analysis suggests a founder mutation that maps to 10q23 (Rogers et al., 2000). HMSNR, which is

phenotypically similar to HMSML but with a unique genetic aetiology, was found to segregate within some HMSNL families (Rogers et al., 2000).

Founder mutations have also been identified as causing genetic disorders originally identified in other populations. A C283Y mutation in the γ-sarcoglycan gene (*SGCG*) was reported as the cause of limb girdle muscular dystrophy type 2C (LGMD2C) in seven unrelated Spanish, French and Italian Romani families (Piccolo et al., 1996). Closely related disease haplotypes confirmed that this was a single founder mutation (Piccolo et al., 1996). The C283Y mutation was subsequently reported in Portuguese Roma (Lasa et al., 1998). This mutation has also been found in Bulgarian Romani LGMD2C patients (Tournev et al., 1998), and carriers identified in a sample of Romani neonates from Northeast Bulgaria (Todorova, Ashikov, Beltcheva, Tournev, & Kremensky, 1999). It is likely that this is the identical founder mutation, although haplotype analysis is required to prove this. Galactosemia in the Bulgarian Roma has been found to result from a P28T founder mutation in the galactokinase gene (*GK1*) (Kalaydjieva et al., 1999). Subsequent to this study, the mutation has been identified in Spanish and Hungarian Roma (Hunter, 2000). A 1267delG founder mutation in the acetylcholine receptor ε subunit gene (*AChRε*) has been identified as the cause of congenital myasthenia in Roma from Hungary, Serbia and Macedonia (Abicht et al., 1999). In addition, a founder E387K mutation in the cytochrome P4501B1 (*CYP1B1*) has been identified as the cause of congenital glaucoma in Slovakian Roma (Plasilova et al., 1999). Private mutations have also been identified as causing Type 3 von Willebrand disease (Casana et al., 2000), autosomal dominant polycystic kidney disease (Forrai et al., 1989; Veldhuisen et al., 1997) and Glanzmann thrombasthenia (Schegel et al., 1994). Disease-causing mutations identified in the Roma that are found in other populations include mutations causing phenylketonuria (Desviat, Pérez, & Ugarte, 1997; Kalanin et al., 1994; Kalaydjieva et al., 1992) and medium-chain acyl coA dehydrogenase (MCAD) deficiency (Kremensky et al., 1998; Martinez et al., 1998).

A limited number of studies have examined the population frequency of founder mutations. These studies indicate that some private mutations occur at gene frequencies in the range of 0.01-0.025 in the Roma (Kalaydjieva et al., 1999; Plasilova et al., 1999; Todorova et al., 1999). These values correspond with carrier frequencies ranging from

2-5%. Preliminary evidence exists that some Romani populations may be at increased risk for particular disorders, as illustrated by a Slovakian Romani group in which the carrier frequency of the *CYP1B1* E387K allele is 11% (Plasilova et al., 1999). In this population the high carrier frequency was reported to result in cases of pseudodominant inheritance of congenital glaucoma.

1.3.3 Summary of Genetic Studies of Roma

Investigations of polymorphic genetic markers and disease genes point to the unique genetic heritage of the Roma. Characterisation of Romani populations using polymorphic markers has demonstrated that these populations are genetically distinct from other European populations. The identification of private disease-causing founder mutations supports this conclusion. However, the failure to identify these mutations in other European populations cannot be equated with their absence in these populations. Therefore, the historical basis of the unique genetic composition of the Roma is unclear. Clearly, the occurrence of identical founder mutations in different Romani populations provides presumptive evidence for common origins or admixture. However, population genetic studies have shed little light on the genetic relatedness of populations. Small populations with possibly long and restrictive bottleneck effects and limited gene flow may rapidly diverge from each other, thereby obscuring evidence of common origins. It is apparent that the use of genetic markers with increased resolution, and study designs that account for social history and structure of the Roma are required in order to gain insights into the genetic origins and structure of the Roma.

CHAPTER 2

REVIEW OF LITERATURE ON MOLECULAR GENETIC STUDIES OF POPULATIONS AND DISEASE

2.1 On the Application of Molecular Genetics to the Study of Human Populations

2.1.1 Introduction

Investigations of biological variation in human populations have enjoyed renewed interest with the advent of molecular genetics and the ability to identify genetic variation at the genotypic level. This field of investigation, termed population genetics, attempts to combine mendelism, darwinism and biometry to determine how the gene could explain the creation, maintenance and distribution of phenotypes in populations (Chakravarti, 2001). As well as providing new insights into population history, the characterisation of human populations has emerged as an essential component of many genetic investigations. Furthermore, the growing concept of an "anthropology of genetic disease" (Weiss, 1998) has broad applications to many aspects of genetics, from methodological approaches of gene identification to rational approaches for disease diagnosis and treatment.

The first study of genetic variation in human populations – examining ABO blood group frequencies - was published in 1919 (Stoneking, 2001). These protein polymorphisms and others provided the first means of investigating molecular genetic variation. However, these genetic variants are phenotypic variants and thus possibly subject to selective pressures. The identification of variation in DNA sequence has provided researchers with an abundance of new polymorphic loci for population genetic studies (Cavalli-Sforza, 1998). DNA polymorphisms are numerous and include variation of single nucleotides, insertions and deletions (indels), mini- and micro-satellite DNA and a multitude of complex repetitive sequences. Variation is found

throughout the human genome on the autosomes, sex chromosomes and mitochondrial DNA. Of these, the Y chromosome and mtDNA have become most widely used in population genetic studies, as they are uniparentally inherited and exist in the haploid state thereby escaping the potentially scrambling effects of recombination.

2.1.2 On the Use of Mitochondrial DNA for the Study of Human Populations

The mitochondria are cytosolic organelles responsible for cellular respiration in almost all eukaryotes (Voet & Voet, 1995). Many hundreds or thousands of mitochondria are contained within each cell. Each mitochondrion contains multiple copies of a small and unique genome consisting of 16,569 base pairs of highly conserved and economically organised DNA sequence (Anderson et al., 1981). The sequence of the human mitochondrial genome encodes 13 proteins involved in oxidative phosphorylation. In addition, two ribosomal RNAs (rRNAs) and 22 transfer RNAs (tRNAs) are encoded in the genome. The mitochondrion-specific tRNAs facilitate the use of a unique genetic code in mitochondrial translation (Voet & Voet, 1995). Mitochondria are transmitted alongside the other cytosolic components from the mother to the oocyte. Thus, mitochondrial DNA follows a strictly maternal inheritance.

The mitochondrial genome exhibits a number of unique features that make it useful for population genetic studies and related disciplines (e.g. forensic science). These features include its maternal inheritance, an apparent lack of recombination and a relatively high degree of mutability. The haploid maternal inheritance of mtDNA means that it is particularly sensitive to reductions and expansions in population size (Parsons et al., 1997). Furthermore, the high mutation rate and apparent lack of selection result in the rapid differentiation of maternal lineages between populations (Parsons et al., 1997). The mitochondrial genome has long been considered not to undergo homologous recombination. Thus, variation in the genome arises exclusively from DNA mutation. An increased rate of mutation has been observed in the displacement loop (D-loop), an 1,122bp region between the $tRNA^{PRO}$ and $tRNA^{PHE}$ genes in which no genes are encoded (Anderson et al., 1981), relative to that observed in the coding regions of the genome. The difference in mutational rates in different regions of the genome provides varying degrees of temporal resolution for population genetic studies.

Although some studies have sequenced the entire mitochondrial genome to examine genetic variation (Finnilä, Lehtonen, & Majamaa, 2001; Horai, Hayasaka, Kondo, Tsugane, & Takahata, 1995), two approaches that are less labour-intensive are generally utilised to estimate the total mitochondrial DNA variation. In the coding portion of the genome, variation is typically characterised using restriction fragment length polymorphism (RFLP) analysis. Within the D-loop, the Hypervariable Segments 1 and 2 (HVS1 and HVS2 respectively) are characterised using direct sequencing. Using these approaches to characterise genomic variation, a standard nomenclature and phylogenetic relationship for mtDNA types have been developed. Superhaplogroups and haplogroups are broad classes of mtDNA types defined on the basis of variation in the coding region. The assignment of mitochondrial genomes to defined haplogroups is generally performed using diagnostic RFLP analysis. In addition, characteristic sequence variants in the HVS1 have been found to be associated with specific haplogroups (Macaulay et al., 1999; Richards et al., 2000; Richards, Macaulay, Bandelt, & Sykes, 1998; Simoni, Calafell, Pettener, Bertranpetit, & Barbujani, 2000a). However, use of this association as a means of inferring haplogroups is contentious (see the debate between Torroni et al., (2000) and Simoni, Calafell, Pettener, Bertranpetit, & Barbujani, [2000b]). Within each haplogroup, sequence variation in the HVS1 and HVS2 provides additional degrees of resolution.

Mitochondrial DNA analysis has been used to address questions ranging in time-scale from evolutionary to recent population history, and to assess current population structure and variation. Since the initial study on mtDNA variation in continental populations (Cann, Stoneking, & Wilson, 1987), research has consistently shown greater mtDNA variation in African populations than other worldwide populations (eg. Vigilant et al., 1991; Chen, Torroni, Excoffier, Santachiara-Benerecetti, & Wallace, 1995). This has provided support for a common African origin of *Homo sapiens sapiens* some 150,000 years ago (Stoneking, 2001). Mitochondrial analysis has been used to address other major prehistoric demographic events such as the Neolithic expansion in Europe (Bertranpetit, Calafell, Comas, Pérez-Lezaun, & Mateu, 1998; Comas et al., 1997; Richards et al., 1996; Sykes, Corte-Real, & Richards, 1998) and the peopling of the Americas (Ward, 1998).

The structure of populations and relationships between linguistically and/or historically related populations have been investigated using mtDNA analysis. Mateu et al., (1997) used mtDNA analysis to compare the peopling of two islands off the coast of Africa, and showed that the impact of different population histories was reflected in the mtDNA of present-day inhabitants. A study of mtDNA in Australian Aborigines (van Holst Pellekaan, Frommer, Sved, & Boettcher, 1998) showed that there was population substructuring at the tribal level, whereas a study of two Indian populations (Mountain et al., 1995) found that the cultural identification of individuals was inconsistent with genetic grouping. In a study of mtDNA sequences of Bulgarians and Turks, it was found that the physical boundary between Europe and Asia corresponded with significant differences between European and West Asian maternal lineages (Calafell, Underhill, Tolun, Angelicheva, & Kalaydjieva, 1996). Similarly, investigation of mtDNA in the Saami, an indigenous nomadic people of northern Scandinavia, demonstrated that this population is genetically distinct from the rest of Europe (Delghandi, Utsi, & Krauss, 1998).

With the proliferation in studies of mtDNA lineages, particularly in European populations, researchers have begun to synthesise the existing data (Macaulay et al., 1999; Richards et al., 1998). Investigations of other global populations are rapidly clarifying the continental origins of different mitochondrial types (Quintana-Murci et al., 1999; Richards et al., 2000). The emerging picture is that some specific mtDNAs are restricted to regional populations. These data serve as a reference from which the ethnogenesis of populations, such as those of Central Asia (Comas et al., 1998), Iceland (Helgason, Sigurethardottir, Gulcher, Ward, & Stefansson, 2000), Brazil (Alves-Silva et al., 2000) and Colombia (Mesa et al., 2000) might be reconstructed.

Although mtDNA analysis has been widely applied to the study of human populations, some of the unique features that make it appropriate for this purpose have been questioned. Namely, it is possible that a strictly maternal inheritance may not be the case in humans, given that paternal inheritance of mtDNA has been observed in other species, including mussels (Zouros, Freeman, Ball, & Pogson, 1992) and fungi (Yang & Griffiths, 1993). Furthermore, paternal leakage has been observed in *Drosophila melanogastor* (Kondo, Matsuura, & Chigusa, 1992) and mice (Gyllensten,

Wharton, Josefsson, & Wilson, 1991). Recently, the commonly accepted lack of recombination in the mitochondrial genome has been questioned on the basis of a negative correlation between linkage and distance (Awadalla, Eyre-Walker, & Smith, 1999; Eyre-Walker, 2000). However, alternative mechanisms have been proposed that might explain this finding, including nonindependent mutation mechanisms and parallel sequencing protocols that may introduce systematic errors causing some covariation with distance (Hey, 2000), errors in the data (Kivisild & Villems, 2000) and inappropriate methodological approaches (Jorde & Bamshad, 2000; Kumar, Hedrick, Dowling, & Stoneking, 2000; Parsons & Irwin, 2000). Moreover, another study has failed to replicate the findings of the initial report (Elson et al., 2001).

The application of mtDNA analysis to the study of populations is based on the premise that variation is due to an accumulation of selectively neutral mutations. However, the mechanisms and forces acting on this process in mitochondria remain unresolved. It has been shown that mutational rates within the D-loop vary between nucleotide sites (Excoffier & Yang, 1999; Meyer, Weiss, & von Haeseler, 1999). Hypervariable sites have been shown to be mutational hotspots, although the reason for this hypervariabilty remains unknown (Stoneking, 2000). Some evidence suggests that the mutability of specific sites is dependent on the sequence context (Howell & Smejkal, 2000; Malyarchuk & Derenko, 1999).

Estimations of mutational rates vary depending on the method by which they have been determined. The largest discrepancy in rates occurs between results obtained using phylogenetic studies and those obtained via pedigree studies. In the non-coding region, Parsons et al., (1997) obtained an empirically observed mutation rate of 2.5/site/million years (Myr), some twenty times greater than with values inferred from phylogenetic studies (Horai et al., 1995). A larger pedigree study provided a result intermediate to these two values [0.32/site/Myr] (Sigurgardottir, Helgason, Gulcher, Stefansson, & Donnelly, 2000). The discrepancy between rates determined from pedigree and evolutionary studies is a consistent finding that suggests the rate and pattern of mutations observed between generations differ from those observed over longer periods of time (Parsons et al., 1997). A possible reason is the presence of multiple copies of the genome within each mitochondrion, each of which can potentially

differ at nucleotide sites. This phenomenon, referred to as heteroplasmy, varies between individuals and has been shown to vary with tissue type and with age (Calloway, Reynolds, Herrin, & Anderson, 2000). It is reasonable to assume that, given the many billions of copies of mtDNA in an individual, everyone is heteroplasmic to some degree (Tully et al., 2000). Detecting heteroplasmy requires techniques that are sufficiently sensitive to detect alternative nucleotides occurring at low frequency. Using fluorescent sequencing techniques one report claimed the ability to detect heteroplasmic mutations occurring at a frequency of 20% (Cavelier, Jazin, Jalonen, & Gyllensten, 2000) whilst a group using denaturing gradient-gel electrophoresis assay estimated the detection of heteroplasmic variants occurring at levels of 5% and greater (Tully et al., 2000). The latter study suggested that approximately 14% of the population is heteroplasmic within this detection level (Tully et al., 2000). Therefore, an important consideration for population geneticists is that mutations must "segregate within a larger mtDNA pool at the organellar, cellular, intergenerational and developmental levels" before they can be detected as substitutions (Parsons et al., 1997). The forces acting on the segregation of mitochondrial types during oogenesis are debatable, with conflicting evidence reported for preferential transmission of mutant genomes (Chinnery et al., 2000), or random genetic drift being the principal determinant (Brown, Samuels, Michael, Turnbull, & Chinnery, 2001).

2.1.3 The Use of Y Chromosome Analyses to Study Human Populations

The Y chromosome exhibits exclusive male inheritance, making it analogous to mtDNA in its uni-parental mode of transmission. The haploid state of the Y chromosome means that it escapes recombination, with the exception of a small region known as the pseudoautosomal region [PAR] (Jobling & Tyler-Smith, 1995). The entire 60-megabase chromosome can therefore be considered as a single locus. The exclusive father-to-son transmission of the Y chromosome provides a means of investigating male-specific histories in human populations. The history of males in a population is likely to differ from female history and will reflect cultural practices governing mating patterns, migrations, wars and colonisation (Jobling & Tyler-Smith, 1995). Following the initial reports of Y chromosome polymorphisms (Casanova et al., 1985; Lucotte & Ngo, 1985)

the number of known polymorphic variants has rapidly increased to over 400 (de Knijff, 2000; Underhill et al., 2000; Underhill et al., 2001). Y chromosome polymorphisms currently being exploited for population genetic studies fall into three classes: unique mutation events (UMEs), microsatellite and minisatellite loci.

Loci that have undergone a mutation once on a single Y chromosome provide a means of differentiating deep-rooted male lineages. UMEs that have been identified on the Y chromosome include ALU polymorphisms (Hammer, 1994; Spurdle, Hammer, & Jenkins, 1994), single nucleotide polymorphisms [SNPs] (Underhill et al., 1997; Underhill et al., 2000), and long interspersed nucleotide elements [LINES] (Santos et al., 2000). Amongst these polymorphisms, SNPs are by far the most numerous, with recent publications bringing the number of Y chromosome SNPs to over 200 (Underhill et al., 2000; Underhill et al., 2001). The singularity of these mutational events allows the construction of a most parsimonious gene tree, thus simplifying the reconstruction of the evolutionary history of the chromosome. Distinct Y chromosomes that are defined solely on the basis of UMEs are designated "haplogroups" (de Knijff, 2000). The antiquity of Y chromosomes defined by UMEs makes haplogroup analysis appropriate for addressing questions regarding evolutionary events (Hammer & Horai, 1995; Underhill et al., 2000) and the peopling and relatedness of regional populations (Semino et al., 2000; Su et al., 1999; Underhill, Jin, Zemans, Oefner, & Cavalli-Sforza, 1996; Zerjal et al., 1997).

In contrast to SNPs, Y chromosome microsatellites or short tandem repeats (Y STRs) demonstrate moderate mutability (Jobling & Tyler-Smith, 1995). Microsatellite DNA generally consists of di-, tri-, tetra- and penta-nucleotide repeat sequence motifs. Y chromosomes that are defined using microsatellites are denoted "haplotypes" (de Knijff, 2000). The number of microsatellite loci in use on the Y chromosome is small, with the initial 14 loci (de Knijff et al., 1997; Kayser et al., 1997) augmented with an additional 6 loci (White, Tatum, Deaven, & Longmire, 1999). Mutations at these simple repeats loci are thought to occur due to DNA polymerase slippage during DNA replication, resulting in the sequence increasing or decreasing by one or two repeat units (Goldstein & Pollock, 1997). Two pedigree-based studies of Y STRs have reported mutational rates of 3.2×10^{-3} mutations/generation (Kayser et al., 1997) and $21\% \times 10^{-3}$

mutations/generation (Heyer, Puymirat, Dieltjes, Bakker, & de Knijff, 1997). An expanded study of almost 5,000 observed meioses reported an average mutational rate over 15 Y STR loci of 2.8 x 10^{-3} mutations/generation (Kayser et al., 2000), closely matching the results from studies of autosomal STRs (Weber & Wong, 1993). In contrast to these findings, mutation rates inferred from evolutionary data are an order of magnitude smaller (Forster et al., 2000). Furthermore, significant variation in mutation rates at different Y STR loci has been proposed by some authors (Carvalho-Silva, Santos, Hutz, Salzano, & Pena, 1999; Thomas et al., 2000; Thomas et al., 1998) and empirical observations show locus-specific variation ranging from 0-8.58 x 10^{-3} mutations/generation (Kayser et al., 2000). The directionality of microsatellite mutation has been shown to be dependent on allele size (Ellegren, 2000; Xu, Peng, & Fang, 2000) and this directional bias has been observed in Y STRs (Kayser et al., 2000), indicating the need for this variable to be included in analyses of Y STRs. Analysis of the different loci also has shown that many microsatellites are compound repeats (de Knijff et al., 1997; Kayser et al., 1997; Kayser et al., 2000), and since not all mutations are necessarily observed using the original methods, protocols have been adjusted accordingly (Forster et al., 1998; Forster et al., 2000; Rolf, Meyer, Brinkmann, & de Knijff, 1998).

Given the rapid rate of diversification of Y STR haplotypes through the processes of mutation and genetic drift, it is possible to draw conclusions about male history and social behaviour on the basis of the relationships between different haplotypes found in populations. Conclusions based on Y STR haplotypes are further justified by the observation that identical microsatellite haplotypes are seldom independently generated along different lineages (Malaspina et al., 1998). Thus, the observation that a single haplotype is frequent in both Ashkenazi and Sephardic priests suggests a common origin of the religious leaders from these two historically separated populations (Thomas et al., 1998). In the Finns, the identification of two predominant Y chromosome haplotypes, separated by more than ten mutational steps, provides evidence for the dual origins of the male population (Kittles et al., 1998). Meanwhile, within India, a study has demonstrated that the majority of Y haplotypes occur uniquely in ethnic groups (Bhattacharyya et al., 1999).

Minisatellite DNA are tandem arrays of short repeats (6-12bp). A single minisatellite has been identified on the Y chromosome, termed MSY1 (Jobling, Bouzekri, & Taylor, 1998). The MSY1 locus has been shown to have a complex structure that is, however, amenable to high-throughput analysis (Bouzekri, Taylor, Hammer, & Jobling, 1998; Jobling, Bouzekri & Taylor., 1998). Although the mutational mechanism of minisatellites is poorly understood they are known to be rapidly mutating systems with mutation rates estimated for the MSY1 locus from 0.02-0.11 mutations/generation (Bouzekri et al., 1998; de Knijff, 2000; Jobling, Bouzekri & Taylor, 1998). The high mutability of MSY1 provides an additional degree of resolution for the Y chromosome. This is particularly useful for paternity test cases (Jobling, Bouzekri & Taylor., 1998), but should also be useful for looking at short-term population history.

All three mutable systems on the Y chromosome can be applied to investigations of human history, ranging from evolutionary questions to migrations, genetic affiliations between linguistic groups, and the history of population admixture and ethnogenesis. As the mutational rates of the three systems differ by orders of magnitude, the applicability of each system is dependent on the research questions. Clearly, the creation of completely characterised lineages using combinations of, or all three mutable systems, provides the most comprehensive means of investigating male history.

Y chromosomes that are characterised using UEPs and microsatellites, termed "lineages" by de Knijff, (2000), have been found to be restricted to single populations. Thus, Y chromosome lineages in more than one population suggest common origins or male-mediated gene flow. Based on this premise, examination of the genetic composition of a population that is the product of admixture can disentangle the origins of the paternal lineages. Thus, the study of an older population, such as the Lemba from southern Africa, demonstrated possible Semitic admixture (Thomas et al., 2000), and investigation of the composition of the Icelandic population suggests 20-25% Gaelic male founders with the remainder of Norse ancestry (Helgason, Sigurethardottir, Nicholson et al., 2000). This same approach has been applied to the characterisation of populations resulting from recent colonial history. In both a "white" Brazilian (Carvalho-Silva, Santos, Rocha, & Pena, 2001) and Colombian "settler" (Carvajal-

Carmona et al., 2000) population, the vast majority of Y chromosomes have been shown to be of European ancestry, with minimal but discernible African and Amerindian male contributions.

The use of a male-specific system for investigating genetic diversity and composition complements the study of mtDNA variation that pertains exclusively to females. A number of studies have utilised the two systems in concert, to reveal significantly different male and female genetic histories within the same population. An investigation of the two genetic systems in Ethiopians revealed that one quarter of the Y chromosomes in the population had a possible Caucasoid origin, whereas only 10% of the mtDNA were Caucasoid, indicating that Caucasoid gene flow was primarily through males (Passarino et al., 1998). Similarly, a comparative study of Indian castes suggested that mtDNA distances reflected social rank and were the result of female gene flow between castes, while a lack of male gene flow resulted in no correlation between social rank and Y chromosome distances (Bamshad et al., 1998). In the aforementioned Latin American populations, which are composed of mainly European patrilineages, 90% of the Colombian matrilineages were shown to be of Amerindian origin (Carvajal-Carmona et al., 2000; Mesa et al., 2000) whilst at least 60% of the Brazilian matrilineages were African or Amerindian (Alves-Silva et al., 2000). More general conclusions have also been made on the basis of complementary studies, such as the assertion that world-wide female migration rates have been eight times higher throughout history than those of males (Seielstad, Minch, & Cavalli-Sforza, 1998).

Although the Y chromosome is a potent tool for investigating population history, there are a number of unresolved issues regarding its application. Y chromosomes have a smaller effective population size than autosomal chromosomes which accounts for reduced diversity observed at polymorphic loci. This smaller effective population size results in genetic drift having a more dramatic effect on Y chromosomal variation than on autosomes (Pérez-Lezaun et al., 1997). This may potentially make it a more sensitive index of population history however, conversely, such sensitivity may provide results that are not representative of the entire population's history (de Knijff, 2000). Within a population, assortative mating may result in genetic profiles that do not accurately represent population history. A further consideration is the possible lack of neutrality of

the Y chromosome. The Y chromosome contains a number of genes which have homologues on the X chromosome, whilst others are involved in testes development and function (Lahn & Page, 1997). Jobling et al., (1998) demonstrated an instance of selection acting on a particular Y chromosome haplotype associated with infertile males. Additional evidence for selection acting on the Y chromosome has been reviewed (Jobling & Tyler-Smith, 2000) and, whilst inconclusive, it suggests that such a possibility should not be discounted. If an advantageous mutation was to occur on a Y chromosome at some point in a population's history, such a chromosome (and its neutral variants) could rapidly increase in frequency. This type of "selective sweep" has been discounted on a global scale due to the concordance in autosomal and Y chromosomal F_{ST} values (Bertranpetit, 2000; Pérez-Lezaun et al., 1997); however, these findings do not preclude the occurrence of localised selective advantages. Whilst these unresolved issues await developments in the field, a more pressing problem should be addressed. That is in the nomenclature of Y chromosomes. Currently, with each new report on the Y chromosome, there is a new study-specific nomenclature offered. This makes interpretations and comparisons of data sets from publications very difficult. To facilitate unimpeded comparisons of Y chromosome data, it is essential that terminologies and methodologies are standardised.

2.1.4 The Application of Disease Allele Haplotype Analyses to the Study of Populations

Each mutation at a disease locus originates on a chromosome, with its polymorphic characteristics giving rise to a distinct marker haplotype footprint (Guo & Xiong, 1997). Perturbations in this "footprint" can then occur through marker mutation and recombination. Through the examination of variation generated at associated neutral sites, one can attempt to reconstruct the history of the mutation in the population. Dating of a disease-causing mutation can aid in the understanding of the origin, evolution, and dispersion of the disease (Guo & Xiong, 1997). This can provide insights into population history and structure. This undertaking is based on the premise that the number of different haplotypes that have evolved from the ancestral chromosome is proportional to the time since the mutation occurred (Morral et al., 1994).

By convention, the most common haplotype associated with the disease-causing mutation is designated the ancestral haplotype. The oldest disease haplotype in the population is expected to have given rise to the highest degree of variation (Morral et al., 1994). A model must be employed that explains the variation generated at the different polymorphic loci and determines the amount of time required for the observed variation to occur. Estimations of the rate at which changes accumulate pose the greatest difficulty to researchers. Recombination is one means by which haplotype variation is produced. The rate of recombination is roughly correlated with the physical distance between genetic elements (Terwilliger & Ott, 1994). However, large deviations from this general correlation and inaccuracies in published genetic maps, such as errors in estimates of genetic distance between markers and in the physical order of those markers, greatly affect calculations. Variation also occurs at loci through marker mutation. Attempts have been made to empirically determine mutation rates at autosomal microsatellite loci (Brinkmann, Klintschar, Neuhuber, Huhne, & Rolf, 1998; Weber & Wong, 1993). However, there appears to be a large degree of variation in mutation rates at different loci. Therefore, average mutation rates, which are generally used in calculations, may be vastly different to those at specific loci. In the development of methods and algorithms for determining the age of mutations, various studies have accounted for either recombination or mutation, whilst disregarding the other variable. A minority of studies have attempted to include both phenomena in their calculations.

Morral et al., (1994) used three intragenic markers in the cystic fibrosis gene, cystic fibrosis transmembrane conductance regulator (*CFTR*), to date the ΔF508 mutation in European populations. Intragenic markers were used to justify discounting recombination and calculation of the age of the mutation was based solely on microsatellite mutation. However, this simplifying assumption is not entirely valid as there is no reason to suppose that intragenic recombination cannot occur. By constructing a most parsimonious tree, relating all haplotypes to the ancestral haplotype, the mean number of mutations to the root haplotype was used to calculate the age of the mutation (Bertranpetit & Calafell, 1996). The limitations of this method are exemplified by the fact that the authors concluded a possible age range of the mutational event between 52,000 – 173,000 years before present.

Attempts to date mutations by considering only recombination have similarly been limited by uncertainty regarding recombination rates. Studies on the history of a founder mutation causing Infantile Onset Spinocerebellar Ataxia (IOSCA) in the Finnish population (Varilo et al., 1996), and Idiopathic Torsion Dystonia (ITD) in the Ashkenazi Jewish population (Risch et al., 1995), used linkage disequilibrium to date disease-causing mutations. In these cases, linkage disequilibrium could still be observed over large genetic distances, which indicated a relatively recent mutational event in these founder populations. Risch et al., (1995) argue that the degree of linkage disequilibrium represents an estimate of the proportion of disease chromosomes bearing the original associated marker allele, and that differences in this value across marker loci should primarily result from the effects of recombination. Using this logic, calculation of the number of generations required to generate the observed diversity becomes a function of linkage disequilibrium and genetic distance.

Stephens et al., (1998) developed an algorithm which incorporates variation at marker loci due to recombination and mutation. In dating the origin of the CCR5-Δ32 allele conferring resistance to AIDS, they estimated a rate of change at microsatellite loci that accounted for both means of generating new alleles at the loci investigated. Reconstruction of the most parsimonious phylogenetic history and the present haplotype frequencies was used to calculate the time required to produce the extant distribution of haplotypes (Stephens et al., 1998). This methods assumes that the proportion of haplotypes that show no change from the ancestral haplotype can be used to estimate the age of origin of the allele. This approach has the benefit of providing estimates that are independent of gene tree topology.

The efforts to date mutations have necessarily been based on simplifying assumptions that are likely to have a detrimental effect on the result. The ability to determine accurately the age of the mutational event requires improved understanding of the biological mechanisms of microsatellite mutation and recombination and their rates of occurrence. In addition, factors such as population size, mating patterns, genetic drift and carrier selection can be expected to profoundly affect gene frequencies and thereby further confound this endeavour. Population modelling that examined these factors would provide important information about the fate of disease alleles and their haplotype

backgrounds. Nonetheless, the dates that are determined using current methods can serve to provide a timeframe that allows some insight into the population being studied. It would seem that the incorporation of additional resources, such as genealogies and known historical events, would provide a useful resource for complementing this undertaking. Furthermore, the integration of this analysis with information on variation elsewhere in the genome would provide a broader approach to studying the history of disease genes and the populations in which they occur.

2.1.5 Summary of Molecular Genetic Tools for Studying Populations

Genetic studies provide a unique insight into the origins, history and structure of human populations. These studies can serve to support conclusions from historical and socioanthropological data. However, as the field of population molecular genetics develops, it is becoming increasingly possible to address questions for which evidence from other sources is virtually absent. The analysis of mtDNA and Y chromosomes affords a complementary approach to the study of human populations. They tell sex-specific histories of populations, which often differ. Though there are numerous uncertainties with regard to the biology of these two genetic elements, their haploid state simplifies interpretations. Analysis of the evolution of a disease haplotype within a population provides an additional means of investigating population history, which is not sex-specific. Using these three approaches in parallel, bearing in mind the limitations of each, allows independent and complementary means of addressing questions about the genetic structure and history of a population.

2.2 The Identification of Disease Genes and the Role of Population Structure

Determination of the ultimate cause of an inherited disease entails the identification of the mutation at the genome level. The enormity of this task is evident when one considers the 3 billion nucleotides of the human genome, of which only 1-2% encode functional genes (Lander et al., 2001). Furthermore, as the majority of genes are not yet characterised, this task often entails not only the identification of a gene defect but of the gene itself. The identification of disease genes has numerous implications for medicine. These include the development of DNA diagnostics and the detection of

disease carriers facilitating presymptomatic or prenatal counselling, and the possible development of gene therapies (Collins, 1992). In addition, the identification of disease genes is an important step in unravelling the aetiology of a genetic disorder. The determination of the primary gene defect directs the next stage of research, in which the protein malfunction may be studied. Although the majority of single gene disorders are extremely rare, the new understanding of cellular physiology should aid in understanding and treating more common disorders.

2.2.1 General Approaches to Identifying Disease Genes

Up until the early 1990s, the majority of disease genes were identified using functional cloning (Collins, 1992). This approach entails the prior determination of the protein defect. Knowledge of the protein that is defective in the disorder means that the amino acid sequence can be used to probe and identify the DNA sequence. Therefore, functional cloning of disease genes is limited to disorders whose biochemical basis is known (Collins, 1992). Furthermore, purification of the protein and determination of the peptide sequence is a necessity. This method is of limited applicability for the majority of genetic disorders, as functional knowledge is scant or non-existent for all but a few diseases.

To overcome this problem, the strategy that has superseded functional cloning is known as positional cloning. This approach to gene identification entails mapping the gene defect to a chromosomal location without any prior knowledge of the gene function. The gene is mapped based on a genetic model and a known mode of inheritance. Polymorphic marker loci are used to analyse the DNA of affected individuals in relation to unaffected family members or a sample from a control population. A statistically significant relationship between the disease and known polymorphic loci enables the chromosomal localisation of the disease gene locus and eventual determination of the genetic defect. Investigations into the function of the disease gene are only commenced subsequent to its identification. This method of identifying disease genes has commonly been referred to as “reverse genetics” but it has been claimed that it is in fact “genetics in purest form, unadulterated by any influences of biochemistry, cell biology or physiology” (Collins, 1992).

In reality, many disease gene identification projects do not adopt a strict positional cloning approach. Mapping efforts will often localise the disease gene to a large chromosomal region, which may contain many hundreds of genes and transcripts. In an effort to overcome the daunting task of analysing every gene in the region, reasonable functional candidates, as determined from knowledge of expression patterns and homology to genes of known function, are selected and preferentially investigated.

2.2.2 Gene Mapping Strategies and the Role of Population Structure

The identification of a disease gene using the positional cloning approach entails an initial stage of mapping the locus to a chromosomal segment. If one is using pedigree data, gene mapping is performed using linkage analysis (Ott & Hoh, 2000). This is the process whereby the unknown gene defect is found to be statistically associated with known loci. The localisation of disease loci using this method is entirely dependent on the biological phenomenon of recombination (Ott, 1991). Statistical analysis of the transmission of alleles within a pedigree and their association with a particular phenotype produces a lod score, which quantifies the likelihood that the locus is associated with the inheritance of the phenotype (Morton, 1955; Botstein, White, Skolnick, & Davis, 1980). A statistically significant relationship provides evidence that the polymorphic locus is physically linked to the disease-causing locus. As lod scores are additive, it is possible to use numerous unrelated families to reach the canonical threshold value of 3 (this log value indicates a locus is 1,000 times more likely to be linked to the disease gene than not linked).

A number of alternative methods to classical linkage analysis have been employed for mapping disease genes. These methods offer the benefit of having high statistical power without requiring large sample sizes. In addition, some of the methods do not necessarily require extended multigenerational pedigrees and can use single affected individuals. At the same time, a number of essential criteria, such as a low disease gene frequency and a young age of the mutation, limit their application. The methods are commonly employed as a means of rapidly localising the disease-gene locus to a gross chromosomal region, and are followed by saturation of the candidate

regions with polymorphic markers and the employment of additional methods to confirm and refine the locus.

Lander and Botstein (1987) proposed a method termed homozygosity mapping to map disease genes in the offspring of consanguineous unions. The method is based on the expectation that one-sixteenth of the genome will be homozygous by descent (HBD) in the offspring of first cousin matings. The homozygous regions are expected to be randomly distributed between different offspring of these matings, except at a common disease locus (Lander & Botstein, 1987). The method is particularly useful as it requires only singletons from consanguineous marriages rather than families with multiple affected individuals (Kruglyak, Daly, & Lander, 1995). Consanguineous unions which are more distant than first cousin yield more information about linkage but this is countered by the decreased region of homozygosity (Lander & Botstein, 1987). Samples from different populations can be used, as exemplified by the use of homozygosity mapping to identify the ataxia-oculomotor apraxia locus in consanguineous families from Japan and Portugal (Ceu Moreira et al., 2001). However, heterogeneous genetic aetiologies of similar phenotypes are more likely in disparate populations. Within population isolates, unexpected allelic heterogeneity and unrelated homozygous segments can impede this approach (Miano et al., 2000).

A similar approach to the identification of candidate disease-gene regions is referred to as segment sharing (Houwen et al., 1994). This method is appropriate for recently founded populations in which the disease allele can be either be demonstrated or inferred as originating from a common ancestor. In the initial study by Houwen et al., (1994), the authors performed a genome scan using just 250 markers in an extended pedigree of 10 individuals, four of whom were affected. An essentially empirical approach was taken in constructing haplotypes using two adjacent markers spaced about 10cM apart and searching for shared genomic segments between affected individuals. If a sufficient number of meioses have occurred, the disease locus can be expected to segregate with those segments shared only by affected individuals. This approach can use a very small sample of patients and unaffected relatives that would not produce a significant result in traditional linkage.

In population isolates one can reasonably expect increased allelic homogeneity of disorders, due to founder effect. A founder mutation is observed as an identical mutation that occurs on closely related haplotypes. Following the initial mutational event, the original haplotype begins to decay through the process of recombination (Guo & Xiong, 1997). The extent of the decay varies greatly depending on the number of meioses since the mutation – which is in turn dependent on the age of the mutation, demographic expansions and drift – and on the physical characteristics of the chromosomal segment. In both homozygosity mapping and segment sharing, an essential assumption is that the deleterious allele is rare and that the mutation is not too old. Both of these factors translate to the requirement that sufficiently few meioses have occurred, so that the haplotype is preserved over detectable regions. In homozygosity mapping, underestimating the frequency of a disease-allele will lead to overestimation of the lod score (Kruglyak et al., 1995). Similarly, mutations that have occurred in the distant past, and therefore have undergone numerous meioses, will have a greatly reduced region of homozygosity. In populations that were founded and expanded in the distant past, one can expect that identical by descent (IBD) segments associated with a disease locus will be less than 1cM, whereas in more recently founded populations IBD regions should be 5-20cM (Houwen et al., 1994). The impact of these parameters on the mapping approach that is adopted highlights the importance of understanding the history and structure of the population from which the study subjects are selected.

An alternative approach to the search for IBD chromosomal segments in affected individuals, is to assess linkage disequilibrium within a population in a genome-wide scan. This is based on the hypothesis that a particular disorder in a population is due to founder effect and that identical by state (IBS) alleles are in fact IBD. This approach, commonly referred to as linkage disequilibrium mapping or an association study, looks for significant non-random association between a specific allele at a polymorphic locus and the occurrence of the disease. Linkage disequilibrium mapping differs from classical linkage in that it attempts to gain information from parental and historical recombinations rather than just from those observed in existing families (Peltonen, 2000). In undertaking such a gene mapping strategy two assumptions are made; namely, that all affected individuals have a common ancestor and therefore there is strong

linkage disequilibrium with markers close to the disease locus, and that the mutation is young enough for linkage disequilibrium to be detected with a reasonable marker density (Visapaa et al., 1998). This approach to gene mapping facilitates the novel method of pooling the DNA of affected and unaffected individuals. Linkage disequilibrium at polymorphic loci can be observed in the different intensities of peaks or bands in an electrophoretic gel (Sheffield, Nishimura, & Stone, 1995). Although the approach would require careful DNA quantification, it was successfully used in the mapping of Hirschsprung's disease to chromosome 13q22 in an extensive Mennonite kindred (Puffenberger et al., 1994), and an axonal form of CMT in a large Tunisian family (Barhoumi et al., 2001).

Genome-wide scans are useful for localising a disease locus to a broadly defined chromosomal region; however, determination of the disease gene within these regions can be a laborious and time-consuming undertaking. In population isolates, although linkage disequilibrium can be identified over large genetic intervals, identical haplotypes in affected persons from different families are found only across highly restricted DNA regions (Peltonen, 2000). Therefore, fine scale genetic mapping can be used to narrow the region of investigation to one that is amenable to physical characterisation. This is largely an empirical exercise, which entails constructing dense marker haplotypes of the region in affected individuals. Both historical and parental recombinations can be observed, and used to refine the region to one of complete homozygosity. This has proved extremely useful in reducing the region of interest to an interval measured in kilobases (Peltonen, 2000). The application of the method is valid only in populations fulfilling certain criteria; namely, that the majority of disease alleles descend from a single ancestral mutation that now has a relatively high frequency, and that the disease allele has had sufficient time to undergo recombinations in the population, thus reducing the region of strongest linkage disequilibrium (Hastbacka et al., 1992; Lehesjoki et al., 1993). These criteria are almost exactly opposite to those required to initially map a gene in a genome scan using linkage disequilibrium. Thus, younger populations are useful for identifying linkage disequilibrium over large distances, whereas older populations are more amenable to refining the region (Jorde, Watkins, Kere, Nyman, & Eriksson, 2000).

It is apparent that well characterised population isolates can provide an efficient means of localising a disease gene locus and narrowing the chromosomal region of interest. However, a major limitation to the use of population isolates for identifying genetic defects occurs in the final stage of gene identification. Once the conserved haplotype has been narrowed to the smallest possible region using all available meioses, all genes within the region are candidate disease genes. In analysing the sequence of these genes, distinguishing disease causing mutations from non-disease causing mutations can be problematic. This is because any base substitution found within the candidate genes can be argued to be unique polymorphism occurring in an isolated founder population (Bonné-Tamir et al., 1997). Furthermore, in the absence of allelic heterogeneity, the search for a single disease causing mutation can be painstaking. This is highlighted by the search for the diastrophic dysplasia gene which was initially mapped in 1990 (Hastbacka, Kaitila, Sistonen, & de la Chapelle, 1990) and successfully identified in 1994 (Hastbacka et al., 1994) based on mutations in non-Finnish individuals. However, the Finnish founder mutation, a splice donor site mutation, was not reported until nine years after the initial report (Hastbacka et al., 1999).

2.2.3 Construction of Integrated Physical and Genetic maps

An essential component of a positional cloning project is the construction of a comprehensive physical map of the chromosomal region. This is usually undertaken in parallel with refined genetic mapping and provides the correct marker order, which is essential for mapping recombination breakpoints. In addition, physical mapping involves the identification of known genes and transcripts that map to the chromosomal segment. This clarifies the genomic content of the region and serves to identify the positional candidate genes.

The first step in physical mapping is the construction of a map of contiguous genomic clones (i.e. a contig) covering the region. This is performed by screening genomic libraries such as yeast artificial chromosomes (YACs), bacterial artificial chromosomes (BACs) and P1 artificial chromosomes (PACs). The library is initially probed with polymorphic markers, which define the critical region. Each clone is then used as a probe to identify successive overlapping clones in a process known as

chromosome walking (Voet & Voet, 1995). This piecemeal process can be pursued in both directions until complete coverage of the critical region is obtained.

Clonal coverage of the genomic region of interest allows the mapping of sequence tagged sites (STSs) to the region. These can include anonymous STSs (genomic fragments of no known function), ESTs (expressed sequence tags, which are derived from mRNAs), polymorphic markers and genes. The relative order of these STSs can often be determined by comparative mapping in the genomic clones or using radiation hybrid mapping. Genomic clones can also be used to identify novel genes in the region using techniques such as exon trapping, cDNA selection (Hastbacka et al., 1994), zoo blots and northern blots (Collins, 1992).

More recently, large scale sequencing and the advances of the human genome project (Collins et al., 1998; Lander et al., 2001) have enabled the analysis of the complete genome sequence in the region of interest. This serves to provide integrated maps which are much more comprehensive and reliable (Collins, 2000) and has resulted in a shift away from laboratory-based physical mapping to the primacy of computer databases and exon prediction programs, commonly referred to as cloning *in silico*.

2.2.4 Well Characterised Population Isolates: Population structure and examples of mapped genes

Exploitation of the increased genetic homogeneity of mendelian disorders in population isolates has proved fruitful in yielding disease genes. In many cases, a predominant disease haplotype is observed which accounts for the majority of disease alleles in the population and is found to bear a single mutation. However, it should be noted that minor haplotypes are found in these populations and allelic homogeneity is rarely complete. In many of the molecular genetic studies performed in these populations, historical and genealogical records have been of great benefit. In addition, the application of population genetic techniques has enhanced the understanding of the populations and thereby the search for disease genes. A brief summary of three of the best characterised genetic isolates follows.

2.2.4.1 The Ashkenazi Jews

The Jews have a history that extends well beyond 3,000 years (Blaney, 2000). Apart from Biblical sources, the initial size of the Jewish population is unknown. They are a population of Middle Eastern origins who have been dispersed throughout the world. The demographic history of the Jews is one of numerous expansions and contractions over the past 2,000-3,000 years. Amongst the most notable is the division of the Jews into the Sephardim and Ashkenazim. Today, the Ashkenazim mainly inhabit Europe and North America whilst the Sephardim are resident in the Middle East and North Africa. In many of these countries, Jews have maintained endogamous practices and admixture with the surrounding society has been uncommon (Rosenberg et al., 2001).

Population genetic studies of the Jews have served to illuminate the structure and history of these dispersed subpopulations. Studies of the Y chromosome in Jewish populations support a common origin of Ashkenazi and Sephardic males (Santachiara Benerecetti et al., 1993). Analysis of the Y chromosomes of priests in both populations indicated their descent from a common ancestor possibly 3,000 years before present (Thomas et al., 1998). Later studies examined Jewish populations dispersed throughout North Africa, the Middle East and Europe using Y chromosome loci (Hammer et al., 2000) and Libya, Ethiopia and Yemen using autosomal loci (Rosenberg et al., 2001), and clearly demonstrated their common origins and isolation from neighbouring communities. Genetic research into an idiopathic torsional dystonia (ITD) founder mutation prompted the suggestion that the current Ashkenazi Jewish populations descend from one thousand or fewer individuals as recently as 500 years ago (Risch et al., 1995).

As a result of long-term cultural isolation, the Jews have a unique spectrum of genetic disorders (Goodman, 1978). Many of these disorders have been extensively studied and the genetic basis for a number has been elucidated (Motulsky, 1995). In many cases, gene-mapping strategies have been employed that make use of the population structuring. For example, linkage disequilibrium mapping was used to localise the idiopathic torsional dystonia gene [DYT1] (Risch et al., 1995). Where founder effects have been demonstrated, invariably a high gene frequency has been

observed. This corresponds to high carrier frequencies such as 1/29 for Tay-Sachs disease, 1/40 for Canavan disease, 1/17 for Gaucher disease, and 1/100 for Bloom syndrome (summarised in Gilbert, 1998). The cause of these inflated gene frequencies in the Ashkenazi Jewish population has been the source of continued debate with some researchers arguing the action of natural selection (Zlotogora, 1994; 1998; Zlotogora, Zeigler, & Bach, 1988), recent genetic drift (Risch et al., 1995) or high gene frequencies in the original population (Goldstein et al., 1999).

2.2.4.2 The Finnish Population

The present-day Finnish population is believed to be derived from a migration of Uralic speakers some 4,000 years ago (Peltonen, Jalanko, & Varilo, 1999). This was followed by a major population expansion 2,000 to 2,500 year ago which coincided with increased migration from Baltic and Germanic regions (de la Chapelle, 1993). Early settlement in Finland was mainly centred in the south of the country, with the north not occupied until the 16th century (Peltonen et al., 1999). Hence, areas settled during this later period represent a subset of the general Finnish population (de la Chapelle, 1993), and/or the entrance of new migrants (Kittles et al., 1999; Kittles et al., 1998). This population is sometimes referred to as the "New Finnish" population. The size of the Finnish population has expanded rapidly from around 250,000 individuals in the beginning of the 18th century to the present number of 5.1 million (Peltonen et al., 1999).

Molecular genetic studies of the population history of the Finns suggest a population bottleneck some 2,000 years ago (Kittles et al., 1999). This is reflected in the relatively limited genetic diversity observed in the Finns, a result of the small founding population and minimal admixture. Further analysis of the Y chromosome has provided evidence that there were two predominant founding populations (Kittles et al., 1999; Kittles et al., 1998). In addition, the Finnish population has undergone numerous minor contractions and expansions, which represent multiple minor bottlenecks.

The occurrence of unique genetic disorders in the Finnish population was first noted by doctors in the 1960s and termed the "Finnish disease heritage" (Norio, Nevanlinna, & Perheentupa, 1973). Many of the causative disease alleles occur at high frequencies. Molecular genetic studies of these disorders have proved remarkably

successful in positioning and identifying the disease genes. This is largely due to the genetic homogeneity of many of the disorders. Initial localisation of the genes has utilised traditional linkage methods, as well as searches for shared genomic segments and linkage disequilibrium mapping (Peltonen, Palotie, & Lange, 2000). The fact that for all the cloned Finnish disease genes one major mutation has accounted for ≥70% of disease chromosomes (Peltonen et al., 1999), has meant that fine structure genetic mapping has been powerful in narrowing the genomic region and eventually identifying the disease gene. There are numerous examples of the successful identification of disease genes in the Finnish disease heritage in which the two stage approach has been applied. Peltonen et al., (1999) lists 32 disease loci that have been identified and 15 in which the mutated gene has been identified. For example, the locus for the disorder presenile frontal lobe dementia with bone cysts (PLO-SL) was mapped to a 9cM region that was then reduced to 150kb through refined mapping (Pekkarinen, Hovatta et al., 1998; Pekkarinen, Kestila et al., 1998). Similarly, the progressive myoclonus epilepsy (EPM1) locus was mapped to a 7cM region and reduced to 176kb through refined mapping (Lehesjoki et al., 1993; Lehesjoki et al., 1991).

In Finland, regional clustering of disorders and mutations has been noted (de la Chapelle, 1993). This reflects substructuring in the Finnish population with differing histories of the disorders and mutations in each subpopulation. Indeed, in the haplotype analysis of the diastrophic dysplasia gene, the major disease haplotype was found to be distributed evenly across Finland whilst the minor haplotypes showed east and west geographic clustering (Hastbacka et al., 1992). Regional clustering of HLA haplotypes has also been observed in the Finnish population (Siren, Sareneva, Lokki, & Koskimies, 1996). Population substructure provides a useful resource, as gene mapping is facilitated in the younger population and refinement of the regions is aided by the increased diversity in the "Old Finnish" population (Jorde et al., 2000).

2.2.4.3 The French Canadians

French colonisation of Canada began in the early 17th century and continued until the capture of Quebec by the British in 1759, at which point it virtually ceased altogether (Scriver, 1992). During this time it is estimated that only 8,000-10,000 people

permanently established themselves in the colony (Charbonneau et al., 1987). Limited subsequent migration and emigration means that the current 5 million Francophones in Quebec are mainly descended from these colonists. During the almost 400 year history of the French population in Canada a number of migrations into different regional areas, such as Charlevoix, along the Saguenay River and the Lac-Saint-Jean area (often referred to as the CSLSJ populations) have occurred (Labuda et al., 1996). Thus, the French Canadians represent a much more recently founded population isolate than the Jews or Finns, with internal migrations representing subsequent population bottlenecks. Molecular genetic studies (Heyer, Tremblay, & Desjardins, 1997), and genealogical and demographic studies (Heyer & Tremblay, 1995), have shown that a small number of founders have contributed to the majority of the present day gene pool.

In Quebec, a number of known genetic disorders occur at increased frequencies (Scriver & Fujiwara, 1992), and a number of previously unknown disorders have been identified in this population (Labuda et al., 1996). Many of these genetic disorders have carrier frequencies in the range 1.5-5% including spastic ataxia, 1/21; polyneuropathy, 1/25; histidinemia, 1/32 and pyruvate kinase deficiency, 1/64 (summarised in Labuda et al., 1996). Theoretical studies have shown that these frequencies can be explained by their introduction by a single founder (Heyer, 1999). In many cases, genetic homogeneity and founder effect have facilitated the application of linkage disequilibrium mapping to localise and refine disease loci. An example is the autosomal recessive spastic ataxia of Charlevoix-Saguenay (ARSACS) locus, which was initially localised using a small number of polymorphic loci examined for excess of homozygosity under the assumption of a founder mutation (Bouchard et al., 1998). An apparent increase in homozygosity at one locus on chromosome 13 focused attention to this region, and linkage analysis was used to map the gene to an 11.1cM segment on chromosome 13q11 which was refined to 5.5cM on the basis of a recombination (Bouchard et al., 1998). Haplotype analysis refined this to a 1.6cM region and location score analysis predicted that the disease locus was 0.42cM distal from one of the polymorphic loci (Richter et al., 1999). It is interesting to note that the authors were not able to further refine the region using haplotype analysis. Linkage disequilibrium around the ARSACS locus was shown to be much greater than around the DTD locus in

the Finns, an older founder population (Engert et al., 2000). The fact that the candidate region could only be reduced to 1.6cM highlights the possible pitfalls of a founder population that is too young to allow further refinement of the disease locus.

2.3 Diseases Under Investigation

Two autosomal recessive disorders were investigated in this thesis. A summary of the literature on clinical and genetic aspects of the disorders is provided below.

2.3.1 Hereditary Motor and Sensory Neuropathy – Type Lom (HMSNL)

Hereditary motor and sensory neuropathy type – Lom (HMSNL) is an autosomal recessive disorder that was first identified in Roma resident in Bulgaria (Kalaydjieva et al., 1996). It is named after the Bulgarian town in which affected individuals were initially identified. However, in addition to the Lom population, the disorder was also identified in the Monteni and Kalderash (Kalaydjieva et al., 1996). Since the initial report of the disorder several more affected Romani individuals have been identified in different European countries including Italy (Merlini et al., 1998), Spain (Colomer et al., 2000), Slovenia (Butinar et al., 1999), France and Rumania. Furthermore, the clinical and neuropathological characterisation of this novel disorder has been refined.

2.3.1.1 Clinical Features of HMSNL

HMSNL is an inherited disorder affecting the peripheral nervous system (OMIM #601455). HMSNL is classified as an autosomal recessive form of Charcot-Marie-Tooth disease type 1 (CMT4D[2]). This heterogeneous class of disorders comprises demyelinating neuropathies, as opposed to axonal neuropathies which are classified as CMT2[3] (Scherer, 1999) and spinal CMT (Timmerman, Nelis, de Jonge, Martin, & van Broeckhoven, 1998). Clinical features common to CMT1 include distal muscle weakness, diminished tendon reflexes, and often foot deformities such as pes cavus (Timmerman et al., 1998).

[2] The nomenclature currently in use distinguishes recessive forms of CMT1 by denoting them as CMT4. However, it has been argued that this confusing classification should be abandoned (Thomas, 2000).

HMSNL is a severe form of CMT1. Onset typically occurs during the first decade of life and it initially manifests as difficulty in walking (Kalaydjieva et al., 1998). Skeletal deformities frequently occur in patients and all sensory modalities are impaired (Kalaydjieva et al., 1998). Electrophysiological studies have revealed greatly reduced nerve conduction velocities (Kalaydjieva et al., 1998). A distinguishing feature of HMSNL is its association with sensorineural deafness (Butinar et al., 1999; Kalaydjieva et al., 1996; Kalaydjieva et al., 1998).

2.3.1.2 Neuropathology of HMSNL

The histopathological hallmarks of CMT1 include extensive de- and remyelination, resulting in hypertrophic changes and onion bulb formations observed in peripheral nerve biopsies (Timmerman et al., 1998). Sural nerve biopsies of HMSNL patients have shown a dramatic reduction in myelinated nerve fibres and fibre density (Kalaydjieva et al., 1998) and the myelin is uncompacted (King et al., 1999). The occurrence of poorly developed atypical onion bulb formations has been reported from a number of studies of HMSNL patients (Baethmann, Gohlich-Ratmann, Schroder, Kalaydjieva, & Voit, 1998; Kalaydjieva et al., 1998; Merlini et al., 1998). In addition, severe and early axonal loss is a pathological feature of HMSNL (Baethmann et al., 1998). A distinguishing feature of HMSNL is the intra-axonal accumulation of irregularly arranged curvilinear profiles, which has previously been observed in cultured cells under experimental vitamin E deficiency (Baethmann et al., 1998; King et al., 1999).

2.3.1.3 Genetic aetiology of HMSNL

HMSNL was initially reported as occurring in three Romani populations in Bulgaria (Kalaydjieva et al., 1996). The disorder has an autosomal recessive pattern of inheritance. The authors mapped the gene using a two-stage strategy. Initially, three affected individuals connected by an average of 9.5 meioses in an extended pedigree were examined in a genome scan for segment sharing. Two shared segments were

[3] Current nomenclature specifies dominant forms of axonal neuropathies as CMT2 and recessive forms as CMT3.

identified on chromosome 8. The investigation was then expanded to include all individuals in the pedigrees and the gene was mapped distal to these segments using 2-point and multi-point lod scores. The highest lod score of 7.7 was obtained for marker D8S378 at a recombination fraction of $\theta = 0$. Saturation of the region with known polymorphic markers identified a core disease haplotype spanning approximately 3cM with evidence of three possible recombinations. All patients were homozygous for a two-marker haplotype that spanned 1.6cM. Then observed allelic homogeneity led the authors to hypothesise that HMSNL in the Roma is caused by a single ancestral founder mutation that occurred more than 800 years ago.

Myelinopathies consist of a spectrum of clinical disorders, which in addition to CMT1, include hereditary neuropathy with liability to pressure palsy (HNPP), Dejerine-Sottas syndrome (DSS), and congenital hypomyelinating neuropathy (Lupski, 2000). Thus far, molecular genetic studies have revealed a number of defects in different genes causing these disorders. CMT1A, CMT1B and CMTX are caused by mutations in peripheral myelin protein 22kd (*PMP22*), myelin protein zero (*MPZ*), and connexin 32 (*CX32*) respectively (Suter & Snipes, 1995). In all cases, the mutations have occurred in the heterozygous state and exhibit a dominant or sex-linked form of inheritance. Recessive inheritance of CMT1 is rare but it is believed to be clinically more severe than dominant forms. A number of autosomal recessive loci for CMT1 have been identified using gene mapping approaches in population isolates. These include a locus on 8q13-q22.1 identified in Tunisian families (Ben Othmane et al., 1993), a locus on 5q23.33 in an Algerian kindred (LeGuern et al., 1996) and a locus on 11q23 in a large Italian kindred (Bolino et al., 1996). Prior to the end of 1999, no genes had been identified that cause recessive CMT1 leaving the molecular basis of this heterogeneous class of disorders unknown. The report by Bolino et al., (2000) of mutations in *myotubularin-related protein-2* causing CMT4B represented the identification of the first gene causing autosomal recessive CMT1.

2.3.2 Limb Girdle Muscular Dystrophy Type 2C (LGMD2C)

Autosomal recessive muscular dystrophies resembling X-linked Duchenne muscular dystrophy form a wide spectrum of clinical severities. To date, at least seven

types of autosomal recessive limb girdle muscular dystrophies (LGMD) have been identified with the probability of further examples (OMIM, 2001). The underlying genetic aetiology of many of these disorders has been elucidated and found to involve defects in different components of the dystro-sarcoglycan complex, which spans the sarcolemma to provide a link between the subsarcolemmal cytoskeleton and the extracellular matrix component, laminin (OMIM, 2001). Limb Girdle Muscular Dystrophy type 2C (OMIM *253700) is caused by mutations in the γ-sarcoglycan gene (*SGCG*) and is one of the most severe autosomal recessive sarcoglycanopathies.

2.3.2.1 Clinical Features of LGMD2C

Onset of the disorder typically occurs by 5 years of age, with most patients wheelchair-bound by 12 years of age. Wasting primarily affects the striated muscles of the limb girdle and truncal muscles (OMIM, 2001). Life expectancy is greatly reduced and patients rarely live to be more than 25 years of age. Investigation of clinical severity in LGMD2C has shown that there is considerable phenotypic variation in individuals with the same mutation (McNally et al., 1996). Merlini et al., (2000) have studied the phenotype of patients homozygous for the C283Y mutation and determined that 49% were severely affected and 51% had a more moderate phenotype.

2.3.2.2 Neuropathology of LGMD2C

Gamma sarcoglycan is an essential subunit of the dystroglycan complex (DGC). It is a transmembrane protein with a postulated structural function in the linking of dystrophin and laminin across the muscle cell membrane (Vainzof et al., 1996). Absence of the γ-sarcoglycan subunit, either due to primary mutations in the gene or to mutations in other subunits of the DGC, invariably results in a severe phenotype (Vainzof et al., 1996). It has been noted that γ-sarcoglycan deficiency is sufficient to cause the dystrophic process independently of dystrophin, which is normal in LGMD2C (Hack et al., 1998). Furthermore, the deficiency of dystrophin in Duchenne Muscular Dsytrophy causes a secondary deficiency of γ-sarcoglycan, leading some to suggest that Duchenne Muscular Dystrophy and LGMD2C share a common pathogenesis related to the deficiency of gamma sarcoglycan (Li, Dickson, & Spiro, 1998).

2.3.2.3 Genetic Aetiology of LGMD2C

Cosegregation of chromosome 13 markers with one form of Severe Childhood Autosomal Recessive Muscular Dystrophy (SCARMD)[4] was first reported in a Tunisian kindred by Ben Othmane et al., (1995). Soon afterwards, Noguchi et al., (1995) reported the cloning of a novel gene encoding a 35 kDa Dystrophin-associated protein, termed γ sarcoglycan (*SGCG*). The paper demonstrated that mutations in the *SGCG* result in LGMD2C and a founder mutation in this gene was responsible for the SCARMD in the Tunisian patients. In 1996, Picccolo et al., reported a C283Y mutation in *SGCG* of LGMD2C affected individuals of Romani ethnicity. Closely related disease haplotypes suggested a common founder mutation at least 1,200 years old. Following these initiaal studies, numerous mutations in *SGCG* causing LGMD2C have been reported (OMIM, 2001).

2.4 Community Genetics

Community genetics is a field that marries predictive testing for inherited disorders with public health. The establishment in 1998 of a dedicated academic journal, *Community Genetics*, which addresses developments in this field, is testament to its growing importance. The editor of *Community Genetics,* Leo P. ten Kate (1998), has defined community genetics as "encompassing all activities to enable the identification of people in a community with increased genetic risks who want to acquire this knowledge in order to make informed decisions". In the same inaugural issue of the journal, Modell and Kuliev (1998) define community genetics as embracing "all approaches for the early identification and prevention of genetic risk that can be applied to whole populations". As with all public health programs, the ultimate aim of community genetics is to improve the health of the population. This is achieved through the central roles of genetic epidemiology, education, audit, development of infrastructure and collaboration with support associations (Modell & Kuliev, 1998). The fundamental

[4] Prior to identification of the genes that encode proteins in the dystroglycan complex a number of disorders with similar phenotype were referred to collectively as severe childhood autosomal recessive muscular dystrophies (SCARMDs). Identification of the causative genes has facilitated refined classification of the disorders on the basis of phenotype and genotype.

distinction between community genetics and clinical genetics is in the approach. Whereas clinical genetics entails waiting for people to request a consultation, community genetics is the process of approaching members of the community who may be at risk but have not yet been identified or helped (ten Kate, 1998).

2.4.1 Carrier Screening

Carrier screening is the identification of healthy individuals who possess a single copy of a deleterious disease allele that may lead to an inherited disease in their offspring. With the development of molecular genetic techniques, it is now possible to directly and definitively ascertain an individual's status with regard to a particular disease-causing mutation. Thus, carrier testing can be offered to help individuals make more informed reproductive decisions (Wilfond & Fost, 1990).

The British Medical Association (1998) has stated that the following criteria should be met in order to justify a genetic screening program:

- The problem must be important, that is it must affect a high proportion of the population or it must be sufficiently severe.
- A suitable screening test should be available in terms of reliability, sensitivity and predictive value.
- It must provide useful information for the management and reproductive decisions.
- The benefits must outweigh the risks.
- Adequate provision must be made for information, counselling and privacy.

Genetic screening programs are often targeted at all members of a community or population. However, alternatives to a general population-based screening program have been offered by a number of researchers. Wald (1991) and Wald, George, Wald, & Mackenzie (1993) proposed treating the reproductive couple as the screening unit and categorising them as at risk or no risk, thereby avoiding potential psychological strains caused by knowledge of an individual's carrier status. However, this approach denies the right of the individual to genetic knowledge. Super, Schwarz, & Malone (1992) proposed a "cascade testing" method in which screening is focused on the relatives of index cases. Other authors have stated that the identification of a carrier or affected

family member provides the impetus for screening the extended family (Kolodny, 1992). In communities that are deemed to be at high risk for particular disorders, such an approach may unfairly discriminate against unrelated at-risk individuals.

In undertaking a carrier testing program, Brock (1994, 1995) identified four potential targets; neonates, high school children, young adults, and pregnant women. A number of programs have detected carriers in neonates. This information is, however, of little use until the person reaches reproductive age, and this does not allow informed consent of the individual. Scriver et al., (1984) have demonstrated the success of heterozygote screening in high school students. However, there is general consensus that young adults represent the most logical target groups (Brock, 1994). In a statement from the American Society of Human Genetics (ASHG/ACMG, 1995) it was stated that "[i]f the medical or psychosocial benefits of a genetic test will not accrue until adulthood, as in the case of carrier status or adult-onset diseases, genetic testing generally should be deferred". However, this assertion may not be cross-culturally applicable. Obviously, carrier-testing should precede conception, thus the appropriate age for genetic testing should account for marriage practices within a community.

Regardless of the approach taken, there appears to be a large variation in success in terms of uptake, depending on the immediacy of the service offered. Watson, Mayall, Lamb, Chapple, & Williamson, (1992) found that when an offer of screening was made opportunistically by members of the research team, 66-87% agreed to the test, but when offered by written invitation, only 10% submitted to the test. Similarly, Bekker et al., (1993) found that the most important variable determining participation rates in screening programs was the personal approach by a professional and the offer of immediate testing.

A measurement of the validity and reliability of the actual test will aid in justifying its use. Holtzman, (1989) outlines three parameters by which a genetic test for disease should be judged. These parameters can be adapted for application to genetic screening for carriers as follows:

- Sensitivity: What proportion of carriers for the disorder will be detected?
- Specificity: What proportion of people who are not carriers will have normal (negative) results?

- Predictive value: What proportion of positive tests are true positives in the population?

It is essential that the limitations of genetic testing be adequately addressed and conveyed to an individual. Tests have been introduced into health care prematurely, for example, the rapid introduction of cystic fibrosis testing prior to confirmation of its suitability (Holtzman, Murphy, Watson, & Barr, 1997). Currently, even for many monogenic traits, the detection rate of genetic mutations is only 60-90% (van Ommen, Bakker, & den Dunnen, 1999). Therefore, careful consideration of the criteria outlined by Holtzman, (1989) is essential prior to the implementation of carrier testing programs.

2.4.2 Genetic Counselling

Genetic counselling has been described as the provision of genetic education coupled with psychosocial counselling (Bowles Biesecker & Marteau, 1999). They define the goals of genetic counselling as:

- Facilitation of autonomous and informed decision making.
- Engendering an appreciation of the inheritance of the genetic condition.
- The integration of genetic information into a useful framework.
- Improvement of the emotional well-being of those affected or their family members.

There are two fundamental concepts that must be conveyed to an individual who has undergone carrier testing. The least subjective concept is that of risk in terms of the likelihood of giving birth to a child with the disease. For a simple autosomal recessive disorder, this risk for an individual carrier is simply a product of the chance of transmitting the disease allele to an offspring and the frequency of the disease allele in the population. If a reproductive couple is tested and both parents are found to be carriers of an autosomal recessive disorder, the risk of producing a child with the disease is one in four, based on the principle of mendelian inheritance.

The concept of risk is distinct from the subjective notion of burden. Burden can be considered as morbidity and the possibility of early mortality, together with the physical, emotional, and financial load for parents (Leonard, Chase, & Childs, 1972). A determination of burden therefore results from a consideration of many factors and can

be heavily influenced by socioeconomic status and culture. However, it is burden more than risk that has a large role in decisions about future childbearing (Leonard et al., 1972).

Once knowledge of carrier status is gained, there are a number of options available to the individual or couple. These are largely determined by the stage at which the relevant information is obtained and the autonomous decisions of those identified to be at risk. The possible outcomes or actions include:

- Selecting a partner who is not a carrier for that disorder.
- Remaining childless.
- Fertility treatment to avoid the disorder, such as oocyte or sperm donation, or preimplantation diagnosis.
- Prenatal diagnosis and termination or continuation of pregnancy.
- "leaving the outcome to fate".

(BMA, 1998, p106)

It has long been held that the fundamental ethos of genetic counselling is that of non-directive counselling. This stems from the concept of an individual's right to make reproductive decisions unencumbered by the opinions of the health practitioner. There has been continued widespread support amongst genetic counsellors and medical geneticists for a policy of non-directiveness (Wertz & Fletcher, 1988). However, the value and practicality of this approach have been questioned (Clarke, 1991). A study that explicitly quantified directiveness in the clinical setting found that all consultations were characterised by some degree of directiveness (Michie, Bron, Bobrow, & Marteau, 1997). This finding has been welcomed by some who claim that nondirective genetic counselling is an impossibility and the focus on nondirectiveness diverts attention from other important goals of genetic counselling (Bernhardt, 1997). Moreover, others assert that there exists no evidence in the literature that the nondirective approach benefits the individual (Wolff & Jung, 1995).

2.4.3 Genetic Screening in Population Isolates

The relative importance of genetic disorders as a health concern is a function of the overall health situation of the individual, family or population. Clearly, in many

countries health problems exist that are of far greater concern than genetic disorders. The importance of genetic disorders tends to be recognised when infant mortality falls below about 40/1,000 (Modell & Kuliev, 1998). Therefore, it can be expected that as general health improves in a population, genetic disorders will become an increasing health concern.

In a population isolate, the frequency of particular genetic diseases is often high and a small number of disease causing mutations is to be expected (Chiba-Falek et al., 1998). The reduction in allelic heterogeneity in such populations translates into increased sensitivity for carrier testing. Thus, population screening programs have proved successful in populations such as the Ashkenazi Jews in which a greater proportion of carriers can be successfully identified (Kaplan, 1998). However, population-based carrier screening has a history of both spectacular successes and failures. By 1992, nearly one million young adults had been tested for Tay-Sachs disease (Kaback et al., 1993). This had resulted in the detection of 36,000 heterozygotes and 1,056 couples deemed to be at risk. A total of 2,516 pregnancies had been monitored and of the 469 affected foetuses identified, 451 had been aborted (Kaback et al., 1993). As a result of this program, the incidence of children born with Tay-Sachs disease has diminished markedly. In contrast, the screening of African-Americans for carriers of sickle-cell anaemia has demonstrated how genetic screening should not be performed. During the 1970s, seventeen U.S. states passed laws on sickle cell screening. In seven of these states screening was mandatory and, in five, marriage licenses and school attendance was denied to those who chose not to be tested (Holtzman, 1989). The draconian measures employed by the states resulted in the eventual abandonment of carrier testing for sickle-cell anaemia.

Carrier testing raises a myriad of issues that must be approached sensitively and thoughtfully. Many of these issues, such as informed consent, non-directive counselling, and post-test alternatives are common to all testing programs. However, carrier testing in small identifiable populations presents novel issues many of which have only recently begun to be addressed in the literature. Paramount among these issues is that of group identity and fears of collective stigmatisation. It is likely that a group's own sense of self-worth and solidarity may be undermined by the finding that they have a greater

genetic propensity for certain inherited diseases (Juengst, 1998). Indeed, Jews in the U.S.A. have expressed concern at the numerous genetic studies that have focused on their population (Foster, Bernsten, & Carter, 1998). Such concern is warranted, given the history of the popular and political misappropriation of genetic studies to support racist agendas (Kohn, 1996). This leagcy dictates that community genetics programs should be implemented in consultation with community members, and in accordance with their desires and concerns.

2.5 Summary of the Literature Review and Research Aims of this PhD Thesis

Increasingly genetics, the science of heredity, is focusing on populations as the unit of study. This represents the fusion of medical genetics and evolutionary population genetics. The convergence of these fields has yielded advances that would not have occurred if they had remained mutually isolated. Thus far, the most fruitful studies have been of population isolates – those populations in which social or geographic constraints restrict the gene pool. The study of genetic structure, diversity and variation has provided insights into the origins and histories of these populations. The incorporation of these findings with knowledge from the humanities informs hypotheses regarding the genetic basis of disease for both the clinician and the researcher. Knowledge of prevalent mutations within a population provides a rational starting point for molecular diagnoses and predictive testing. Meanwhile, the assumption that occurrence of a disease within a population is due to the population-wide segregation of a disease allele derived from a common ancestor dictates the appropriate strategy for gene identification.

Genetic studies of the Roma indicate that they have some of the characteristics of a population isolate. Mendelian disorders have been shown to be the result of private founder mutations. The homogeneity and frequency of disease alleles are consistent with founder effects in genetically restricted populations. Related disease genes identified in different populations point to genetic affinities between some Romani groups. However, the studies of non-disease loci provide contradictory and confounding evidence. A review of the genetic evidence suggests that the Roma comprise a

genetically heterogeneous conglomerate of populations, and their relatedness and origins remain unclear.

The purpose of this thesis is to address some of the unresolved questions regarding the genetic architecture of the Roma – its composition, the relatedness of populations and diversity within each population. Four broad aims have been developed to address unresolved questions of the genetics of the Roma. Specific questions developed within each aim are stated at the commencement of the four sections, which are focused on one of the following objectives:

- To study the genetic composition and structure of Romani populations through the investigation of neutral genetic variation.
- To apply knowledge of the unique structure of the Romani population to refined genetic mapping and positional cloning of the HMSNL gene.
- To investigate the distribution and history of two founder mutations, causing LGMD2C and HMSNL, in Romani populations.
- To assess a pilot genetic screening program for a founder mutation causing LGMD2C in a high-risk Romani community.

Section I

A POPULATION GENETIC STUDY OF THE ROMA

CHAPTER 3

SUBJECTS AND METHODS

3.1 Introduction and Study Design

3.1.1 Summary of Previous Findings

Genetic investigations of the Roma have been undertaken for over eighty years. The initial study by Verzar & Weszeczky (1921) and the majority of subsequent studies have focused on the origins of the Roma and the relatedness of Romani populations. Many of these studies have been impoverished by flaws in study design, inaccurate definition of populations, and inappropriate sampling approaches and methodologies. This is primarily due to a disregard of the social history of the Roma in attempting to interpret genetic data.

Preliminary evidence for genetic relationships between some of these populations is suggested by shared disease mutations (Kalaydjieva et al., 1996; Piccolo et al., 1996; Abicht et al., 1999). These private mutations, found in diverse Romani populations, suggest either common origins or gene flow. Evidence for genetic relatedness of three socially separated Romani populations resident in Bulgaria, in which HMSNL occurs, is provided by a common predominant paternal lineage and shared maternal lineages (Kalaydjieva et al., 2001). However, genetic relationships between a wide range of Romani groups have not been investigated using uniparentally inherited markers.

Social anthropological studies invariably describe strict endogamous marriage practices in the Roma (Marushiakova & Popov, 1997). Virtually every genetic study of Romani populations indicates that the population differs from autochthonous populations and from other Romani populations. This is possibly due to the rapid divergence of allele frequencies in genetically restricted populations. Limited male-specific diversity has been noted in three Vlach Romani populations by Kalaydjieva et al., (2001), and in a forensic study of Hungarian Roma (Füredi, Woller, Padar, & Angyal, 1999). Founder mutations also provide possible evidence of limited genetic

diversity. However, knowledge of the relationship between social practices and genetic diversity in the Roma is poorly understood.

3.1.2 Research Questions

This study aims to address three broad issues pertaining to the population history and structure of the Roma. A number of questions are posed within these aims:

1. To investigate the origins of the Roma and their relationship to other populations.
   A. What does the composition of the Romani gene pool indicate about the parental population(s) from which the present day Roma are derived?
   B. How are the Roma related to other worldwide populations, including autochthonous Asian, European and African populations?
   C. Are the Roma descended from an ethnically diverse or homogeneous people?
2. To investigate the relatedness of Romani populations.
   A. Do socially and geographically separated Romani populations share common biological origins?
   B. How are geographically and socially separated Romani populations related to each other?
   C. Can population histories be inferred from the genetic data?
3. To investigate the biological impact of social practices.
   A. What impact has the practice of endogamy had on the genetic diversity?
   B. Can the impact of social history be observed in the genetic composition of the populations?
   C. Are sex-specific histories discernible from the genetic data?

3.1.3 Design of the Study

Data were generated for five Romani populations for the first time in this study. These are the Turgovzi, Feredjelli, Intreni, Spanish and Lithuanian Roma. Data from three Romani populations investigated by Kalaydjieva et al., (2001) were included in data analyses: the Lom, Monteni and Kalderash. The populations studied are geographically dispersed throughout Bulgaria and Europe (figure 3-1). Romani populations sampled in Bulgaria provide a representation of the three metagroups

outlined by Marushiakova & Popov, (1997); namely, the Jerlii, Kalderash and Rudari. The Spanish and Lithuanian Roma provide examples of Western and Northern European Romani groups. Thus, the eight populations studied provided representation of the three major migrations that have occurred within Europe – migration into the Balkans, the migration into Western and Northern Europe (henceforth referred to as the West European migration), and the migration from Moldavia and Wallachia. Table 3-1 provides a summary of demographic, anthropological and historical information for the Romani populations investigated in this study.

The investigation of variation in the female-specific mtDNA and male-specific Y-chromosome provide a complementary approach to the study of these populations. Both systems were examined using two classes of polymorphic loci with differing rates of mutation, which allows different degrees of temporal resolution. Characterisation of Y chromosomal variation was performed by genotyping unique event polymorphisms (UEPs) and microsatellite DNA/short tandem repeats (STRs). Analysis of mtDNA was performed using RFLP analysis and direct sequencing of the HVS1. Both maternal and paternal lineages were examined in the same male subjects. Statistical analyses were used to quantify genetic composition and to assess interpopulation and intrapopulation structure.

Figure 3-1. Geographic locations of the Romani populations included in the population genetic study

Table 3-1

Cultural and historical summary of populations included in population genetic study

Population	Country of residence	Major migrational grouping	Meta-group	Religion	Language	Estimated size	Lifestyle and social practices
Turgovzi	Bulgaria	Balkan	Jelii	Muslim	Turkish Romany (Balkan dialect)	5,000	Settled Regional endogamy
Feredjelli	Bulgaria	Balkan	Jerlii	Muslim	Turkish	3,000	Settled Group endogamy
Spanish Roma	Spain	West European	Caló	Roman Catholic	Caló	800,000	Settled Regional endogamy
Lithuanian Roma	Lithuania	West European	Russian	Roman Catholic	Baltic Romany	10,000	Ex-nomads Strict group endogamy
Intreni	Bulgaria	Vlach	Rudari	Eastern Orthodox Christian	Rumanian	NA	Ex-nomads (1920s) Endogamous within metagroup
Monteni[Ψ]	Bulgaria	Vlach	Rudari	Eastern Orthodox Christian	Rumanian	NA	Ex-nomads (1920s) Endogamous within metagroup
Lom[Ψ]	Bulgaria	Vlach	Rudari	Baptist	Romany (Old Vlach)	7,000	Ex-nomads (1890s) Regional endogamy
Kalderash[Ψ]	Bulgaria	Vlach	Kalderash	Eastern Orthodox Christian	Romany (New Vlach)	40,000-80,000	Ex-nomads (1950-1960s) Strict inter-group endogamy (complex trans-national interclan rules)

[Ψ]Populations studied by Kalaydjieva et al., (2001)

3.1.4 Sample Collection

Biological samples from Romani populations resident in Bulgaria, Spain, and Lithuania were obtained by various senior researchers (table 3-2). Blood samples were obtained from individuals with informed oral consent, and all studies were performed in accordance with the ethical guidelines stipulated by Edith Cowan University. DNA samples of twenty Lithuanian Romani males and one hundred sex-anonymous Spanish Roma were provided by collaborating researchers.

Table 3-2
Information on population sampling programs

Population	Place of residence	Sampling performed by	Number of unrelated males investigated
Turgovzi	Omurtag, Bulgaria	Drs Kalaydjieva, Angelicheva and Tournev	36
Feredjelli	Omurtag, Bulgaria	Drs Kalaydjieva, Angelicheva and Tournev	21
Intreni	Liaskovetz, Tantra Valley, Bulgaria	Drs Kalaydjieva, Angelicheva and Tournev	18
Monteni[Ψ]	Balkan Mountains, Bulgaria	Drs Kalaydjieva, Angelicheva and Tournev	42
Lom[Ψ]	Lom, Bulgaria	Drs Kalaydjieva, Angelicheva	19
Kalderash[Ψ]	Various locations, Bulgaria	Drs Kalaydjieva, Angelicheva and Tournev	15
Spanish Roma	Madrid, Spain	Dr de Pablo	35
Lithuanian Roma	Vilnius, Lithuania	Dr Kucinskas	20
Total			**206**

[Ψ] Population studied by Kalaydjieva et al., (2001).

3.2 Preparation and Quantification of DNA Samples

3.2.1 Isolation of DNA Samples

Most blood samples from the Turgovzi, Feredjelli and Intreni individuals were collected on FTA™ Genecards (Invitrogen™ Life Technologies). This system is designed to lyse the blood cells and permanently bind DNA upon contact with the paper matrix. Purification of FTA™ Genecards removes all proteins, carbohydrates, lipids and organic matter whilst leaving the DNA immobilised on the paper. The washing protocol recommended by the manufacturers was followed with minor amendments that were found to enhance the purity of the end product.

A small disc of approximately 4mm diameter was excised from each FTA™ Genecard using a punch. The disc was placed in a 250µL Eppendorf™ tube and 200µL of FTA™ Purification Reagent (Invitrogen™ Life Technologies) was added. At this point, incubation of the sample at 4°C for twelve hours greatly expedited the cleaning process through a reduction in total labour time and with no apparent detriment to the end-product. Following the initial incubation, the FTA™ Purification Reagent was removed and the disc was subjected to two additional washes using 200µl of FTA™ Purification Reagent with incubation for five minutes at room temperature on each occasion. During each wash, the mixture was briefly vortexed at t = 0 minutes, t = 2.5 minutes and t = 5 minutes. After a total of three washes with the FTA™ Purification Reagent, the sample was rinsed twice with Tris-EDTA solution. The sample was then dried at room temperature for one hour.

Samples collected on 3MM Whatman filter paper (Whatman) required conventional DNA extraction using a salting out protocol. A $1cm^2$ blood spot was cut from the filter paper and placed in a 1.5mL-Eppendorf™ tube along with 250µL of 0.1% Triton X-100 and 15µL of 10mg/mL proteinase K. The mixture was vortexed for one minute and incubated at 50˚C for 30 minutes. This step was repeated once and 25µL of 10x SET buffer was added. A 1:1 chloroform/phenol extraction was performed using 250µL of chloroform and 250µL of phenol. The contents of the tube were mixed by inversion and then centrifuged at 13000rpm for 10 minutes. The supernatant was

removed to a fresh 250µL Eppendorf™ tube and 1/10 the volume of 3M sodium acetate (pH 4.9) added. To precipitate the DNA, an equal volume of 100% isopropanol was added and the mixture was left overnight at –20˚C. The next day, the mixture was centrifuged at 13000rpm for 30 minutes. The resulting pellet was washed once with ice cold 70% ethanol and centrifuged for 15 minutes at 13000rpm. All liquid was removed and the pellet was air dried for 1 hour. The dried pellet was then resuspended in 50µL of dH_2O.

3.2.2 Quantification of DNA Samples

Liquid DNA samples were quantified using a Beckman DU 640 UV spectrophotometer (Beckman Coulter Inc.). A 1:50 dilution of the DNA specimen in water was prepared and the absorbance determined at wavelengths of 260nm (A_{260nm}) and 280nm (A_{280nm}). The spectrophotometer was blanked using dH_2O. Absorbance by nucleic acids is read at 260nm and by proteins at 280nm. A ratio of 1.8±0.2:1 of A_{260nm}: A_{280nm} was indicative of suitable product purity for enzymatic manipulation of DNA. DNA concentration was determined using Beer's law, $A_{260nm} = b \cdot \varepsilon \cdot c$ where c is the DNA concentration, b is the path length of the light and ε is the molar absorbance coefficient. Working solutions of 10ng/µL were prepared from stock DNA.

DNA samples immobilised on FTA Genecards™ cannot be quantified.

3.2.3 Sex Determination of Anonymous DNA Samples

Samples of unrelated Spanish Roma were sex-anonymous. Therefore, sex identification was performed using the amelogenin locus PCR assay. The amelogenin gene is present in the pseudoautosomal region (PAR) of the X and Y chromosomes; however, a small deletion in the former allows determination of sex-chromosome genotype.

The amelogenin locus was amplified as described by Nakahori, Hamano, Iwaya & Nakagome (1991) using the primers AMXY-F (5'CTG-ATG-GTT-GGC-CTC-AAG-CCT-GTG-3') and AMXY-R (5'TAA-AGA-GAT-TCA-TTA-ACT-TGA-CTG-3'). PCR reagents were 1µL of 10x PCR buffer (Qiagen), 0.4µL of 15mM Mg^{2+} (final concentration of 6mM), 0.4µL of 5mM dNTPs, 0.1µL of each primer (at 5µM

concentration), 0.1μL of T*aq* polymerase (ie. 0.1U) [Qiagen], 6.6μL of dH_2O and 1μL of 10ng/μL DNA. Touchdown PCR was performed as follows: initial denaturation of 10 minutes at 94˚C; 20 cycles of 1 minute at 94˚C, 2 minutes at 68˚C (Δ-0.5˚C /cycle), 2 minutes at 72˚C; 15 cycles of 1 minute at 94˚C, 2 minutes at 55˚C, 2 minutes at 72˚C; final extension period of 10 minutes at 72˚C; cooling to 4˚C. PCR products were then loaded on to a 2% agarose containing 2.3μL of 100% ethidium bromide and electrophoresed in 1x TAE at 90V for 30 minutes.

XX genotypes yield a single PCR fragment of 190bp. XY genotypes produce the 190bp fragment and a fragment of 320bp. DNA fragments were visualised using a Mighty Bright™ UV transilluminator (Hoeffer Scientific Instruments) and photographed using the Kodak® DC120 Electrophoresis Documentation and Analysis System™ (Eastman Kodak Company), which includes the Kodak® DC Zoom™ Digital Camera and 1D Image Analysis Software™.

3.3 Analysis of Y Chromosome Variation

3.3.1 Introduction

Y-chromosomes were genotyped in unrelated males at 31 UEP loci and 8 microsatellite loci. All loci are situated on the non-recombining portion of the Y-chromosome.

3.3.2 Y Chromosome Unique Event Polymorphism (UEP) Genotyping

Y chromosome UEPs were genotyped by Dr Peter Underhill at the Department of Genetics, Stanford University. UEPs were genotyped using a combination of denaturing high-pressure liquid chromatography (DHPLC) and direct sequencing. Raw data were forwarded to myself for analysis. The UEP loci genotyped were M1, M145, M40, M96, M174, M33, M75, M2, M35, M15, M55, M9, M45, M89, M172, M170, M173, M67, M124, M52, M69, M82, M92, M17, M12, M73, M201, PN2, M168, M216 and M217. These loci are SNPs, except M1, which is an *Alu* insertion (Spurdle, Hammer, & Jenkins, 1994) and PN2 and M82, which are 2-bp deletions. Information on

UEP loci, including primer sequences, PCR protocols and allelic is in Underhill et al., (2000) and Underhill et al., (2001).

3.3.3 Y Chromosome Microsatellite Genotyping

Seven Y chromosome microsatellite loci were characterised using methodologies outlined in Kayser et al., (1997). Y-chromosome microsatellite loci included the tetranucleotide repeat loci DYS19, DYS390, DYS391, DY393, DYS389I and DYS389II; and the trinucleotide repeat locus DYS392. Six primer pairs amplify these loci as the DYS389 forward primer anneals twice yielding two polymorphic PCR products. Information regarding PCR primer sequences and amplification conditions, standardised nomenclature, consensus sequence of the loci and PCR fragment size and repeat number correlations are available from the website http://ruly70.medfac.leidenuniv.nl/~fldo/(.) Allelic ladders for microsatellite loci (excluding DYS389AB and CD) were provided by Dr P de Knijff of the Forensic Laboratory for DNA Research, Department of Human Genetics, Leiden University.

PCR primers were commercially prepared (Geneworks Pty Ltd), with a fluorescent label chemically attached to one primer of each pair. Markers DYS393, DYS390 and DYS391 were labelled with 4, 7, 2, 7-tetrachloro-6-carboxyfluorescein (TET) which appears as green when using Filter set B on the ABI 373A DNA Analyser. Primers for DYS389 were labelled with 6-carboxyfluorescein (FAM) which appears as blue using Filter set B on the ABI 373A DNA Analyser. DYS19 and DYS392 were labelled with 4, 7, 2, 4, 5, 7-hexachloro-6-carboxyfluorescein (HEX) which appears as yellow using Filter set B on the ABI 373A DNA Analyser.

3.3.3.1 PCR protocols for Y chromosome microsatellites

For the seven microsatellite loci, two optimised PCR mixtures and four temperature cycling protocols were found to provide the best results. PCR mixtures were a total of 10μL and their composition varied only in the final $MgCl_2$ concentration. PCR mixtures for markers DYS390, DYS391, DYS392, DYS393 and DYS19 contained 1μL of 10x PCR buffer (Qiagen), 1μL of a mixture containing 2.5mM of each dNTP (Australian Biotech Inc.), 0.4μL of 25μM forward and reverse primers, 0.1U of *Taq*

DNA Polymerase (Quiagen), 0.8µL of 25mM $MgCl_2$ (Qiagen), 3.3µL of dH_2O and 3µL of 10ng/µL sample DNA. The PCR mixture for DYS389 was near-identical, except it required 0.6µL of 25mM $MgCl_2$ and 3.5µL of dH_2O. The four optimised thermocycling programs for the GeneAmp PCR System 9600 (Applied Biosystems) are outlined in table 3-3

3-3
Optimised PCR cycling conditions for Y chromosome microsatellite loci

Cycle Stage	DYS389	DYS-19,391,392 & 393	DYS390
Denaturation	5min@94˚C	5min@94˚C	5min@94˚C
Amplification	14 cycles: 20s@94˚C 30s@63˚C(Δ-0.5˚C/cycle) 30s@72˚C 20 cycles: 20s@94˚C 30s@56˚C 30s@72˚C	16 cycles: 20s@94˚C 1min@63˚C(Δ-0.5˚C/cycle) 1min@72˚C 25 cycles: 20s@94˚C 45s@55˚C 1min@72˚C	10 cycles: 30s@94˚C 30s@60˚C(Δ-0.5˚C/cycle) 30s@72˚C 25 cycles: 20s@94˚C 30s@55˚C 30s@72˚C
Final Extension	5min@72˚C	5min@72˚C	5min@72˚C
Cooling	4˚C	4˚C	4˚C

3.3.4 Sample Preparation for the 373A DNA Analyser

PCR products from DYS393 were diluted 1:100 with dH_2O. This prevented signal peaks from being too large on the 373A ABI DNA Analyser (Applied Biosystems). Samples for the ABI 373A DNA Analyser were prepared in a 250µL Eppendorf tube with 2.5µL of PCR product mixed with 2.5µL of formamide, 0.5µL of loading buffer and 0.6 µL of N, N, N', N'-tetramethyl-6-carboxyrhodiamine (TAMRA)-labelled internal size standard (Applied Biosystems). Multiplexing of samples was also performed. The optimal results were obtained when PCR products labelled with the same fluorescent tag were run in the same lane, that is PCR products from the loci DYS390, DYS391 and DYS393; and DYS19 and DYS392 were combined. When multiplexing samples, the total volume of PCR product remained 2.5µL and comprised equal volumes of each sample. Samples were vortexed and centrifuged to collect all liquid. They were placed on a hot plate at 94˚C for 4 minutes to denature PCR products and stored on ice prior to being loaded on the gel.

3.3.5 Size Separation of DNA Fragments Using the 373A DNA Analyser

Polyacrylamide gels were prepared using 50mL of 6% acrylamide/bisacrylamide at a ratio of 19:1. Polymerisation of the acrylamide was catalysed by the addition of 350μL of 10% ammonium persulphate and 35μL of N,N,N',N'-tetramethylethylenediamine (TEMED). Gels were formed between 36cm gel plates using 0.4mm spacers, and a 50 lane square tooth well former was inserted. The gel was left to set for a period of at least two hours. It was then loaded into the machine and pre-run for 5 minutes using 1x TBE as the running buffer.

The 373A DNA Analyser (Applied Biosystems) run parameters were set as filter set B, a PMT of 645, a voltage of 1200V, laser power of 40V, a current of 800 milliamps and 10 scans per minute. Approximately 1.5μL of denatured sample was loaded into each well, and samples were electrophoresed for 8 hours. Sample lanes were tracked manually on the gel image and extracted using the GENESCAN™ program (Applied Biosystems). The TAMRA-500 size standard (Applied Biosystems) was defined according to the values provided by the manufacturer and manually verified for each lane. The data were exported to the GENOTYPER™ program (Applied Biosystems), which assigns allele sizes to peaks based on the internal size standard. Peaks were manually checked to ensure data quality, and split or ambiguous peaks were rejected. Allelic ladders were run on each gel to account for any gel-specific variation. This ensured that the correct repeat number for each microsatellite was determined.

3.4 Analysis of Mitochondrial DNA

3.4.1 Introduction

DNA variation in the coding region and hyper-variable sequence 1 (HVS1) of the noncoding D-loop of the mitochondrial genome was assessed using RFLP analysis and direct sequencing respectively.

3.4.2 RFLP Genotyping

DNA samples were analysed for diagnostic coding region RFLPs by Dr Guiseppe Passarino of the Department of Genetics, Stanford University. Standard restriction endonuclease digests were performed according to protocols of Passarino et al., (1996) and Richards, Macaulay, Bandelt, & Sykes, (1998). Raw data from genotyping results were forwarded to myself.

3.4.3 Analysis of the mtDNA HVS1

3.4.3.1 PCR amplification of the HVS1

The HVS1 of the mitochondrial DNA D-loop was amplified using composite primers, which consist of phage M13 and human mtDNA-specific sequences. The segment of the mitochondrial genome spanning nucleotide positions 15,997 to 16,400 was amplified using a "hot start" PCR technique. This was found to be essential to ensure a high quality PCR product amenable to sequencing. PCR was performed in a total volume of 50μL. An initial mixture was made containing 3μL of 10x PCR buffer (Qiagen), 3μL of 25mM $MgCl_2$, 0.5μL of 5mM dNTPs, 0.5μL of 10μM forward primer 37.5μL of dH_2O and 3μL of 10ng/μL genomic DNA sample. This mixture was heated at 94˚C in a thermocycler for 5 minutes to ensure complete denaturation of the target DNA. After this period, the remaining reagents consisting of 2μL of 10x PCR buffer, 0.5μL of the reverse primer and 0.25μL of T*aq* polymerase (Qiagen) were added to the reaction. Thermocycling proceeded with 30 cycles of 94˚C for 45 seconds, 66˚C for 60 seconds and 72˚C for 60 seconds. This was followed by a final extension time of 10 minutes at 72˚C and subsequent cooling to 4˚C.

3.4.3.2 Confirmation and Cleanup of HVS1 PCR Product

A sample of 5μL of PCR product was electrophoresed on an ethidium bromide stained 2% agarose gel for 30 minutes at 80 volts. A product of 460bp indicated a positive result. The PCR products were purified by gel filtration chromatography using QIAquick PCR Purification Kit (Qiagen) following the manufacturer's protocol.

3.4.3.3 Sequencing Reaction for HVS1

Direct cycle sequencing of PCR products was performed using the M13 Ready Reaction Dye Primer™ kit (Applied Biosystems). Thermocycling conditions entailed 25 cycles of 94˚C for 5 seconds, 60˚C for 5 seconds and 72˚C for 60 seconds. PCR products were sequenced in both directions using the M13 -21 and reverse primers in order to confirm sequence variants.

Sequencing products were precipitated at -20˚C for 1 hour with the addition of 45µL of 100% ethanol and 2µL of 4.9M NaOH. This mixture was centrifuged at 13,000 rpm for 30 minutes. The pelleted sequencing products were dried at room temperature and resuspended in a 5:1 formamide to loading buffer mixture. Samples were then denatured at 94˚C for 5 minutes and placed on ice until loaded on to the 373A DNA Analyser (Applied Biosystems).

3.4.3.4 Sequence Determination Using 373A DNA Analyser

Sequencing gels were formed from freshly prepared 4% polyacrylamide mixes as follows: 30g of urea were dissolved in 22mL of dH_2O to which was added 6mL of 40% 19:1 acrylamide/bisacrylamide solution. The mixture was deionised using 1g of deionising resin beads and filtered and degassed under vacuum. To the filtrate was added 6mL of filtered 10x TBE and additional dH_2O to a total volume of 60mL. Polymerisation was catalysed with 300µL of 10% ammonium persulphate and 30µL of TEMED and the solution was poured between 48cm well-to-read glass plates separated by 4mm spacers. The gel was allowed to set for 2 hours at room temperature. Wells were formed using a plastic shark tooth comb. The gel was loaded into the 373A DNA Analyser and pre-run for 5 minutes.

A denatured sample of 1.5µL was loaded into each well. Filter set A was used with a PMT of 640, a voltage of 1200V, a laser power of 40W and a run time of 12 hours. Gel images were captured using Sequence Analyser (Applied Biosystems). Sample lanes were manually tacked and exported to Sequence Navigator (Applied Biosystems) for analysis.

3.5 Statistical Analyses

3.5.1 Processing of Y chromosome Data

PCR fragment sizes were entered into a Microsoft Excel spreadsheet and manually converted to repeat numbers based on the sequenced allelic ladders. Amplification of DYS389 and subsequent gel electrophoresis results in two easily distinguishable products, denoted DYS389I and DYS389II. The sequence structure of this locus is well-characterised and contains 2 repetitive segments, denoted DYS389AB and DYS389CD, each of which is composed of two different types of repeat units, interrupted with an invariant segment (Rolf, Meyer, Brinkmann, & de Knijff, 1998). For analytic purposes, authors have either rejected one of the fragments (Bhattacharyya et al., 1999; Bosch et al., 1999), analysed all four fragments separately using a nested PCR protocol (Forster et al., 2000), subtracted the smaller fragment from the larger (Hurles et al., 1999) or used raw PCR fragment sizes (Black, 1999). For the purposes of this study, the DYS389II fragment was considered as being composed of repeat blocks ABCD (referred to as m, n, p and q by Forster et al., 2000). The PCR fragment DYS389I includes blocks C and D and is thus referred to as DYS389CD. Subtracting DYS389I from DYS389II leaves blocks A and B and is therefore referred to as DYS389AB. This approach ensures that the two loci are treated independently and variation is neither missed nor counted twice.

3.5.2 Processing of Mitochondrial DNA Data

Raw sequence data were analysed using Sequence Navigator (Applied Biosystems). Sequences were edited manually by examination of the electropherogram and consensus sequences were determined from forward and reverse sequences. In order to compare results with previous studies, a 360bp segment from positions 16,024-16,383 (Anderson et al., 1981) was analysed. Edited sequences were exported as ASCII2 text files for analysis using computer programs.

3.5.3 Computer Applications

Statistical analyses of Y-chromosomes and mtDNA data were performed using Arlequin 2.000 (Schneider, Kueffer, Roessli, & Excoffier, 1996). In addition, MitDesc (F. Calafell, pers comm) was used for some mtDNA analysis. Neighbour-joining trees were constructed from distance matrices using the NEIGHBOR program included in the PHYLIP package (Felsenstein, 1989) and drawn with DRAWTREE and TREEVIEW (Page, 1996). Phylogenetic analyses of Y chromosomes and mtDNA were performed using Network 2.00 (Bandelt, Forster, Sykes, & Richards, 1995).

3.5.4 Intrapopulation/Genetic Diversity Analyses

Estimated allele frequencies and haplotype frequencies were calculated for each population using the equation:

$$\hat{p}_i = \frac{x_i}{n}$$

where allele/haplotype *i* is observed *x* times in a sample containing *n* gene copies (Weir, 1990; Weir, 1996). This statistic provides a value for comparisons between frequency estimates of discrete heritable traits in populations.

Gene diversity was estimated by summing the squares of allele frequencies as follows:

$$D = 1 - \left[\frac{n}{n-1} \left(\sum f_i^2 \right) \right]$$

(Weir, 1996)

where n is the sample size and f_i is the frequency of allele i. This measure of variation is sometimes referred to as the average heterozygosity and is applicable in inbred populations in which there may be few heterozygotes but numerous different homozygous types (Weir, 1996). It is also an appropriate measure of variability in haploid systems such as the Y chromosome and mtDNA.

Haplotype diversity, which is equivalent to expected heterozygosity for diploid data or the probability that two randomly chosen haplotypes are different in a sample, was calculated using:

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

where n is the number of gene copies in the sample, k is the number of haplotypes and p_i is the sample frequency of the i-th haplotype. The sampling variance of this statistic was calculated using the equation:

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

(Nei, 1987).

Nucleotide diversity, the probability that two randomly chosen homologous nucleotides are different, was estimated as:

$$\pi = \frac{n}{n-1}\left(\frac{\sum_{j=1}^{L}\left(1-\sum_{i=1}^{4}\chi_{ij}^2\right)}{L}\right)$$

(Nei, 1987; Tajima, 1983).

Where n is the sample size, L is the number of nucleotides in the sequence and x_{ij} is the frequency of the ith nucleotide at position j.

Sequence diversity is analogous to haplotype diversity and was estimated as:

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

(Calafell, Underhill, Tolun, Angelicheva, & Kalaydjieva, 1996)

Where p_i the frequency of each of the k unique sequences in the sample.

The mean pairwise difference in number of repeats across seven Y chromosome microsatellite loci was computed, which provides a relative value for the relatedness of haplotypes within a population (Pérez-Lezaun et al., 1999) and within Y chromosome haplogroups (Bosch et al., 1999). The mean number of pairwise differences between mitochondrial sequences was also calculated. Mean pairwise differences were calculated using the equation:

$$\pi = \sum_{i=1}^{k} \sum_{j \leq 1} p_i p_j \hat{d}_{ij}$$

Where d_{ij} is an estimate of the number of mutations that occurred since the divergence of haplotype/sequence i and j, k is the number of haplotypes/sequences and p_i and p_j are the frequencies of haplotypes/sequences i and j.

Sampling variance of this statistic was calculated using the equation:

$$V\left(\hat{\pi}\right) = \frac{3n(n+1)\hat{\pi} + 2(n^2 + n + 3)\hat{\pi}^2}{11(n^2 - 7n + 6)}$$

(Tajima, 1983).

3.5.5 Interpopulation Analyses

Interpopulation analyses of Y microsatellite haplotypes were performed by calculating genetic distances between populations. Slatkin (1995) developed a genetic distance that assumes single step mutation model (SMM) in microsatellites. This value, referred to as R_{ST}, is analogous to F_{ST} (Wright, 1965) which was derived under the assumption of the infinite alleles model (IAM).

Population pairwise R_{ST} (Slatkin, 1995) values were calculated using the equation:

$$R_{ST} = \frac{(S_{bar} - S_w)}{S_{bar}}$$

Where, S_w is the sum over all loci of twice the weighted mean of the within-population variances V (A) and V (B); and S_{bar} is the sum over all loci of twice the variance V (A+B) of the combined population (Slatkin, 1995).

Using mtDNA data, genetic distances between populations were determined using the intermatch-mismatch distance. This distance is calculated using the equation:

$$d = d_{ij} - \frac{(d_i - d_j)}{2}$$

where d_{ij} is the average number of nucleotide differences between populations i and j and d_i and d_j are the average pairwise differences within populations i and j (Di Rienzo et al., 1994; Mountain et al., 1995).

The statistical significance of population pairwise distances were determined through 1000 iterations using the bootstrapping resampling method (Efron, 1982).

From population pairwise genetic distance matrices, unrooted neighbour joining trees (Saitou & Nei, 1987) were generated using PHYLIP 3.5c (Felsenstein, 1989). This program was also used to test the robustness of tree branches using the statistical resampling process of bootstrapping for 1000 iterations (Efron, 1982).

To examine further population structure in the European Roma, an analysis of molecular variance (AMOVA) was performed (Excoffier, Smouse, & Quattro, 1992). This analysis estimates variance components reflecting different levels of hierarchical subdivisions – those due to genetic differences (i) between groups, (ii) between populations within groups, and (iii) between individuals within populations. Genetic variance of Y chromosome microsatellites was determined using the "sum of size differences" option in Arlequin 2.000. Genetic variance of mtDNA sequences was performed using the "pairwise differences" option in Arlequin 2.000.

3.5.6 Method for Determining Coalescent Age of Y Chromosome Lineages

The age of Y chromosome haplogroups was determined based on haplotype variation using the method described by Stephens et al., (1998). This method assumes that the probability P a haplotype does not change from its ancestor *G* generations ago is

$P = (1\text{-}r)^G$

In an expanded population, *P* is the proportion of haplotypes which are ancestral (Risch et al., 1995). Therefore, an estimate of *G* can be determined by

$G = -\ln(P)/r$

Where *r* is the estimated rate at which variation is accumulated in the haplotype, based on the recombination and mutation rate. In the case of Y chromosomes, variation is only attributed to mutation. Two different values of *r* were determined using mutation rates for YSTR loci determined from a pedigree based study, 2.8×10^{-3} (95% CI 1.72-4.27×10^{-3}) [Kayser et al., 2000] and an evolutionary study, 2.6×10^{-4} (95% CI 2.33-2.87×10^{-4}) [Forster et al., 2000] multiplied by the number of polymorphic loci. An average of the values determined using the two rates was calculated. Ninety-five percent confidence intervals were calculated using the 95%CI for mutation rates.

The number of generations, *G*, was converted to number of years considering a generation age of 20 years.

CHAPTER 4

RESULTS

4.1 Genetic Composition of the Roma

4.1.1 Identification of Male Lineages in the Roma

For the purposes of this study, it was assumed that identical Y chromosome microsatellite haplotypes are the result of a recent common ancestor and therefore must occur on the same ancestral Y chromosome defined by UEPs (i.e. the same haplogroup). This assumption is based on the high mutation rate of Y chromosome microsatellites (Heyer, Puymirat, Dieltjes, Bakker, & de Knijff, 1997; Jobling, Heyer, Dieltjes, & de Knijff, 1999), and the near absence of homoplasmy within haplogroups (Bosch et al., 1999). Several Y chromosomes in the sample had identical microsatellite haplotypes. Therefore, to minimise redundancy, UEP genotyping was performed on a subset of 94 chromosomes from 169 unrelated males. Haplogroups were inferred for Y chromosomes bearing identical microsatellite haplotypes. This assumption was validated by genotyping some redundant Y chromosomes which, were invariably found to belong to the predicted haplogroup.

4.1.1.1 Y chromosome haplogroups identified in the Roma

Ten haplogroups were identified in the sub-sample of 94 Y chromosomes (table 4-1). Fifty-nine Y chromosomes were assigned to one of these haplogroups on the basis of their identical haplotypes resulting in haplogroup assignments for a total of 153 Y chromosomes. The additional sixteen Y chromosomes bore unique haplotypes and could not be assigned to any known haplogroup. The haplogroup nomenclature is in accordance with that proposed by Underhill et al., (2000); except for haplogroup V-52 defined by the loci M216 and M217, which was first described by Underhill et al., (2001).

Table 4-1
Y chromosome haplogroups identified in the Roma. Number of chromosomes observed in a sample of 94 males and inferred in a sample of 169 males.

Haplogroup name	Mutations defining haplogroup	No. of observed chr.	No. of inferred chr.
VI-68	M89, 52, 69, 82	24	67
VI-52	M89, 170	25	35
VI-56	M89, 172, 67	15	18
IX-104	M89, 9, 45, 173	12	14
VI-71	M89, 168	8	8
III-36	M1, 145, 40, 96, PN2, 35	3	4
V-52	M216, 217	4	4
VI-58	M89, 172	1	1
VI-57	M89, 172, 67, 92	1	1
IX-108	M89, 9, 45, 173, 17	1	1
Unknown	-	-	16
Total		94	169

4.1.1.2 Distribution of Y chromosome haplogroups in the Roma

Haplogroup VI-68 is the most frequently occurring Y chromosome haplogroup in the Roma, representing 39.6% of the sample (figure 4-1). Other frequently occurring haplogroups include VI-52 (20.7%), VI-56 (10.7%) and IX-104 (8.3%).

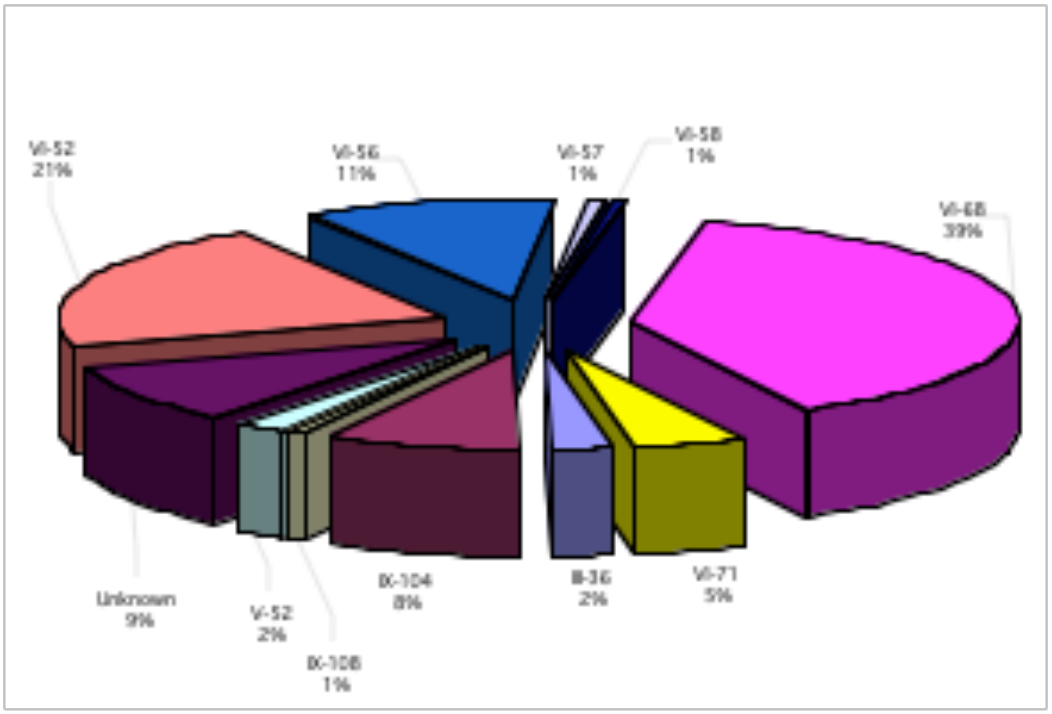

Figure 4-1 Proportion of Y chromosome haplogroups in Romani males

All the Y chromosome haplogroups identified in the Roma have been previously reported as occurring in other worldwide populations (table 4-2).

Table 4-2
Distribution and frequency in global populations of Y chromosome haplogroups identified in the Roma.

Haplogroup	Frequency in sample of 169 Roma	Global populations in which haplogroup is found[1,3]
VI-68	39.6%	Hunza (5.3%)[1], Pakistan & India (4.5%)[1], C. Asia (0.5%)[1]
VI-52	20.7%	Europe (13.3%)[1], Basque (2.2%)[1] W. Europe (10%)[3], E. Europe (20%)[3], Middle East (3%)[3]
VI-56	10.7%	Middle East (8.3%)[1], Sardinia (4.5%)[1], C. Asia (1.1%)[1]
IX-104	8.3%	World-wide[2], W Europe (56.2%)[3], E. Europe (19.8%)[3], Middle East (8.2%)[3]
VI-71	4.7%	World-wide[2]
III-36	2.4%	Khoisan (10.2%)[1], Ethiopia (6.8%)[1], S. Africa (1.9%)[1]
V-52	2.4%	-
IX-108	0.6%	Pakistan & India (31.8%)[1], Hunza (28.9%)[1], C. Asia (16.3%)[1], Europe (5%)[1], W. Europe (1.1.%)[3], E. Europe (33.8%)[3], Middle East (13.2%)[3]
VI-57	0.6%	India & Pakistan (4.5%)[1], C. Asia (0.5%)[1]
VI-58	0.6%	Morocco (10.7%)[1], Middle East (8.3%)[1], C. Asia (6.5%)[1], Pakistan & India (3.4%)[1], Europe (3.3%)[1], Hunza (2.6%)[1]

[1]Summarised from Underhill et al., (2000).
[2]Worldwide distribution denotes its occurrence in Europe, Asia, Africa and America.
[3]Summarised from Semino et al., (2000) (NB in this paper VI-52=Eu7, IX-104=Eu18, IX-108=Eu19)

The most frequently occurring Y chromosome haplogroup in the Roma, VI-68, has not been reported in European populations before. Previous studies suggest it is an infrequently occurring haplogroup in the Indian subcontinent and Central Asia (Underhill et al., 2000). Haplogroup VI-56, which represents over 10% of Romani Y chromosomes, is found mainly in the Middle East and only in Sardinia within Europe (Underhill et al., 2000). In contrast to these two haplogroups, the second most frequently occurring haplogroup, VI-52, is most frequent in Eastern Europe, and haplogroup IX-104 is the most common haplogroup in Western Europe (Semino et al., 2000). Haplogroup VI-52 has not been reported in India, whereas haplogroup IX-104 is globally dispersed and found in India. All other haplogroups identified in the Roma are found in either West Asian or European populations or both. The exception is haplogroup III-36, which is found only in African populations.

4.1.1.3 Results of Y chromosome microsatellite genotyping

Haplotypes were constructed for 122 Y chromosomes using seven microsatellite loci. Data were collated with those of Kalaydjieva et al., (2001). Four Lom samples typed for UEPs were not genotyped for Y STRs by Kalaydjieva et al., (2001). Thus, complete haplotypes were available from 164 samples from eight populations. A single Lithuanian Roma sample was found to be biallelic at locus DYS19. This is probably due to duplication of the Y chromosome segment that includes this locus (de Knijff et al., 1997). Duplications at this locus, as revealed by observable differences in allele size, occur at an estimated frequency of 0.12% (Kayser et al., 2000). The possibility of contamination with another sample was considered; however, the unambiguous results from other Y chromosome loci and mitochondrial DNA suggest that this was unlikely. Therefore, for the purposes of analyses, this sample was resolved into two haplotypes that were identical at the other six Y STR loci but differed at DYS19. Thus, 21 Y chromosome haplotypes were obtained from the samples of 20 Lithuanian males, and the total sample size was considered to be 165.

A total of 52 unique haplotypes were identified in the sample (table 4-3). Haplotype names are derived from the haplogroup assignment of the Y chromosome plus a unique letter suffix. Twenty-two of the chromosomes occur more than once in the sample whilst the other 31 are singletons. Haplotype VI-68-a represents 27.3% of Y chromosomes in the entire Romani sample. Other frequent haplotypes are VI-52-a (12.1%), VI-68-c (7.8%) and VI-56-b (7.3%). Eleven haplotypes were not assigned to a haplogroup and therefore given the generic prefix “Ht”.

Table 4-3

Y chromosome haplotypes identified in the Roma

Haplotype	DYS19	DYS390	DYS391	DYS392	DYS393	DYS389AB	DYS389CD	No. of Chr.
VI-68-a	15	22	10	11	12	16	14	45
VI-68-b	14	22	9	11	12	16	14	2
VI-68-c	14	22	10	11	12	16	14	13
VI-68-d	15	23	10	11	12	16	14	2
VI-68-e	15	21	10	11	12	16	14	1
VI-52-a	14	22	10	11	13	16	12	20
VI-52-b	15	25	11	11	13	18	13	3
VI-52-c	17	24	10	11	13	17	13	3
VI-52-d	16	22	10	11	12	19	13	2
VI-52-e	16	24	11	11	13	17	13	1
VI-52-f	13	23	10	12	12	18	14	1
VI-52-g	14	23	10	11	13	17	13	1
VI-52-h	15	23	9	12	14	18	14	1
VI-52-I	14	22	10	11	13	16	13	1
VI-52-j	17	23	10	11	13	16	14	1
VI-52-k	16	24	11	11	13	18	13	1
VI-56-a	14	23	11	11	12	17	14	2
VI-56-b	14	23	10	11	12	17	14	12
VI-56-c	14	23	10	11	12	16	14	2
VI-56-d	15	23	10	11	12	17	14	1
VI-56-e	14	23	10	11	12	17	15	1
IX-104-a	14	24	11	13	12	16	13	2
IX-104-b	14	24	11	13	12	17	13	1
IX-104-c	14	25	10	13	13	16	13	2
IX-104-d	14	24	11	13	13	16	14	2
IX-104-e	14	24	11	13	13	16	13	3
IX-104-f	14	23	11	13	13	16	13	1
IX-104-g	14	24	11	11	13	16	14	1
IX-104-h	14	24	10	14	13	15	13	1
IX-104-I	15	24	11	13	13	16	13	1
VI-71-a	14	23	10	11	13	16	12	1
VI-71-b	14	25	10	11	13	17	14	3
VI-71-c	15	23	10	11	12	16	13	1
VI-71-d	14	21	10	11	13	16	12	3
III-36-a	13	24	10	11	14	19	14	1
III-36-b	13	24	10	11	13	17	13	3
V-52-a	15	24	11	11	12	16	13	1
V-52-b	15	24	10	11	13	16	13	2
V-52-c	15	25	10	11	13	16	13	1
VI-58-a	14	23	11	11	12	18	12	1
VI-57-a	16	22	10	11	12	18	13	1
IX-108-a	14	24	11	11	13	17	13	1
Ht-a	14	25	10	11	13	16	14	2
Ht-b	16	22	11	11	12	19	13	1
Ht-c	14	25	10	11	13	15	14	1
Ht-d	13	24	10	11	13	19	13	1
Ht-e	14	22	10	12	14	16	14	1
Ht-f	15	22	10	12	14	16	14	1
Ht-g	16	23	10	11	13	16	13	1
Ht-h	15	22	10	11	12	16	15	1
Ht-j	14	22	10	11	12	17	14	4
Ht-k	15	22	10	11	13	16	14	2
Ht-l	17	24	10	11	13	16	13	1

4.1.1.4 Analysis of Y chromosome haplogroups

Haplogroups were assessed for internal diversity by the analysis of haplotypes occurring within the same haplogroup. Network analysis was used to examine the history of haplogroups within the population.

4.1.1.4.1 Diversity within haplogroups.

Diversity was determined within each haplogroup (table 4-4). Of the seven haplogroups that are represented more than once in the sample of Romani males, VI-68 and VI-56 are by far the least diverse with average haplotype diversities below 0.56, average pairwise differences between haplotypes that are well below 1 and average gene diversities less than 0.1. In contrast, haplogroups VI-52, IX-104 and VI-71 have haplotype diversities above 0.66, average pairwise difference values over 2 and average gene diversities over 0.3.

Table 4-4
Summary statistics of Y chromosome haplogroup diversities

Haplogroup	N	No. of haplotypes	Mean pairwise differences	Haplotype diversity	Average gene diversity
VI-68	63	5	0.52 +/- 0.44	0.458 +/- 0.004	0.075 +/- 0.070
VI-52	35	11	2.94 +/- 1.58	0.671 +/- 0.008	0.419 +/- 0.250
VI-56	18	5	0.64 +/- 0.52	0.551 +/- 0.018	0.092 +/- 0.083
IX-104	14	9	2.16 +/- 1.28	0.934 +/- 0.002	0.309 +/- 0.205
VI-71	8	4	2.46 +/- 1.49	0.786 +/- 0.013	0.352 +/- 0.242
III-36	4	2	1.50 +/- 1.12	0.500 +/- 0.086	0.214 +/- 0.191
V-52	4	3	1.50 +/- 1.12	0.833 +/- 0.055	0.214 +/- 0.191
VI-58	1	1	.	.	.
VI-57	1	1	.	.	.
IX-108	1	1	.	.	.

4.1.1.4.2 Network Analysis of Haplogroups

Y chromosomes belonging to identical UEP defined haplogroups share a close evolutionary ancestry. Homoplasmy is minimised thereby, allowing a meaningful reconstruction of phylogenetic relationships between microsatellite haplotypes. To examine these relationships, median-joining networks were constructed for the four most frequent haplogroups (Figure 4-2). It is apparent that the five haplotypes identified in the

most frequently occurring haplogroup, VI-68, are closely related with all haplotypes related by single step mutations. Similarly, the five haplotypes contained within VI-56 are closely related within the network. Moreover, in both these haplogroup networks no inferred nodes are necessary. In contrast, haplotypes within haplogroup VI-52 show a much more complex relationship with multiple inferred nodes representing unobserved haplotypes. Haplogroup IX-104 displays a network with some inferred nodes and an overall intermediate complexity in comparison to haplogroups VI-68, VI-56 and VI-52.

.

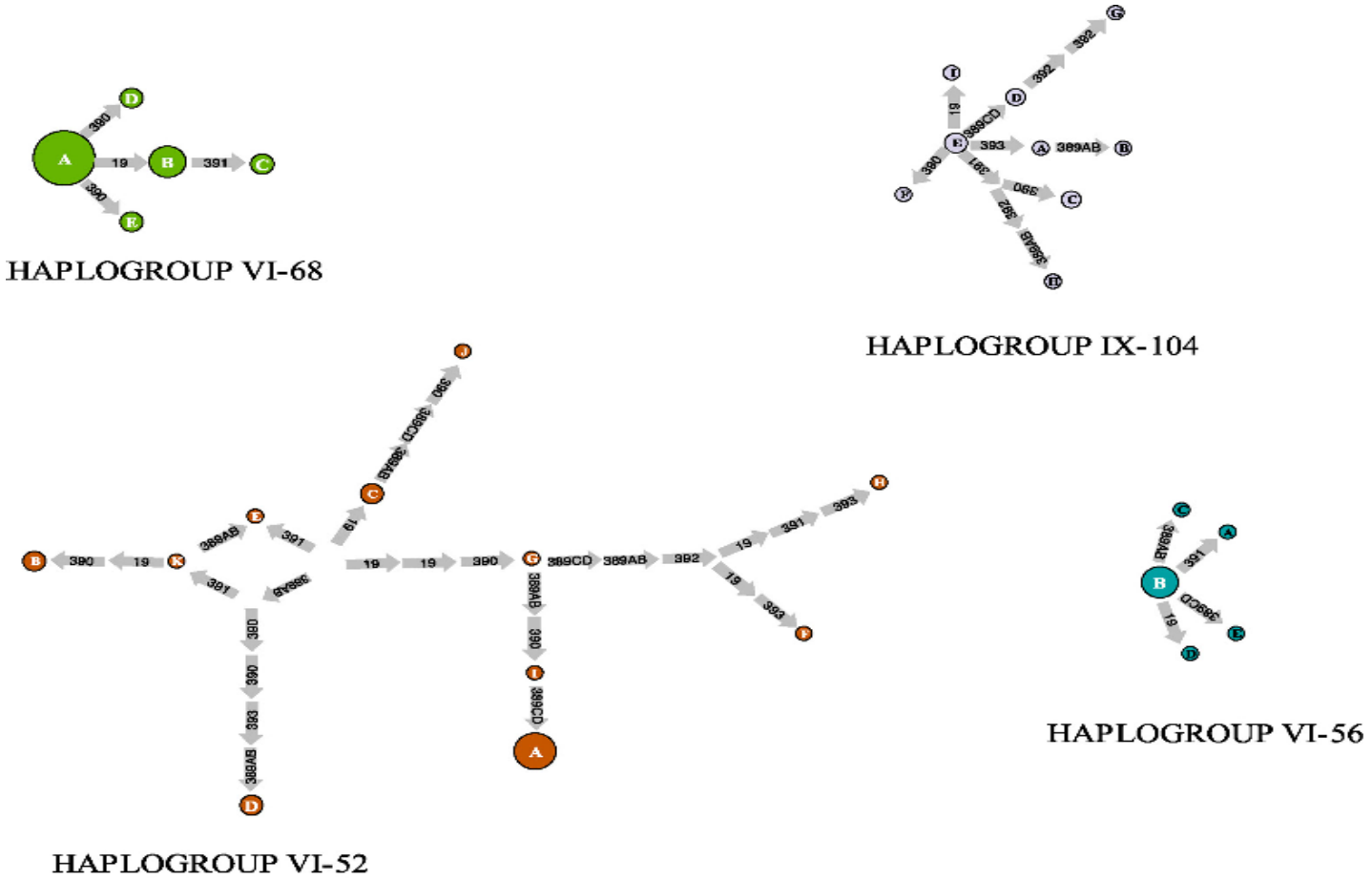

Figure 4-2 Networks displaying relationships between Y chromosome microsatellite haplotypes within the most frequently occurring haplogroups in the Roma

4.1.1.4.3 Age of founding Y chromosome haplogroups in the Roma

For a Y chromosome haplogroup that has been introduced into the population once, the determined age of the haplogroup will roughly correspond with that historical event. Multiple introductions of a haplogroup by different males invalidates this premise, and the coalescent age of a haplogroup will precede its introduction into the population. Network analysis indicates that haplogroup VI-68 and VI-56 have been introduced by a limited number of related males, whereas the diversity in haplogroups VI-52 and IX-104 suggests multiple admixture events. Therefore, the ages of the founding haplogroups VI-68 and VI-56 in the Roma were determined based on the frequency of the ancestral haplotype using coalescent theory. The rate *r*, at which change accumulates within each haplogroup due to mutation at the seven loci was estimated as $r_{pedigree}$ = 0.0196 (95% CI 0.0120-0.0299) and $r_{evolutionary}$ = 0.00182 (95% CI 0.00163-0.00201). Using $r_{pedigree}$, the coalescence of haplogroup VI-68 was estimated to be 343 years BP (95% CI 225-560), and for haplogroup VI-56 it was estimated to be 414 years BP (95% CI 271-676). Calculation of these values using $r_{evolutionary}$ produced estimates of 3697 years (95% CI 3,347-4,128) for haplogroup VI-68 and 4,455 years (95%CI 4,034-4,975) for haplogroup VI-56. An average of these estimates, obtained using different mutation rates, suggests the diversity within haplogroup VI-68 is 2,020 (95% CI 1,789-2,344) years old, and within VI-56 it is 2,435 (95% CI 2,153-2,826) years old

4.1.2 Female Lineages in the Roma

The mtDNA HVS1 of 102 unrelated males from five Romani populations were sequenced. These data were collated with HVS1 sequences from 83 Roma from three populations previously studied (Kalaydjieva et al., 2001). RFLP analysis was performed for 169 of the 185 samples (i.e. RFLP analysis was not performed on Intreni samples).

4.1.2.1 Results of RFLP genotyping

Mitochondrial haplogroups were designated based on RFLP motifs and the sequence status of nucleotide position 00073 (Passarino et al., 1996; Macaulay et al.,

1999; Richards, Macaulay, Bandelt, & Sykes, 1998). Congruence between haplogroups defined by RFLPs and characteristic HVS1 variants was examined and any discrepancies reanalysed (table 4-5). Haplogroup assignment for mtDNA from Intreni samples was based solely on HVS1 variants.

Table 4-5
Definitions of mtDNA haplogroups identified in Romani individuals

Haplogroup	HVS1 Variants	00073 Status	RFLP motif
H	…	A	-7025*Alu*I, -14766*Mse*I
I	16223, 16129	G	-4529*Hae*II, +8249*Ava*II/-8250*Hae*III +10032*Alu*I
J	16126, 16069	G	+4216*Nla*III, +10394*Dde*I, -13704*Bst*0I
M	16223	G	+10394*Dde*I, +10397*Alu*I
Pre V/H	…	A	-14766*Mse*I
T	16126, 16294	G	+4216*Nla*III, +4914*Bfa*I, +13366*Bam*HI, 15606*Alu*I, -15925*Msp*I
U(K)	16224, 16311	G	+12308*Hinf*I, -9052*Hae*II/-9053*Hha*I, +10394*Dde*I
U1	16189, 16249	G	+12308*Hinf*I, -4490*Alu*I, -13103*Hinf*I/+13104*Mbo*I
U3	16343	G	+12308*Hinf*I
U5	16270	G	+12308*Hinf*I
W	16223, 16292	G	+8249*Ava*II/-8250*Hae*III, -8994*Hae*III
X	16223, 16278	G	+14465*Acc*I
N1b	16145, 16176G, 16223	G	

Thirteen mtDNA haplogroups were found in the sample of 185 Romani individuals. Haplogroup H is the most prevalent haplogroup at a frequency of 29.2% (figure 4-3). Haplogroup M accounts for one-quarter of all mtDNA haplogroups. Other common haplogroups include haplogroup U3 (13%), haplogroup J (10.1%) and haplogroup X (8.6%).

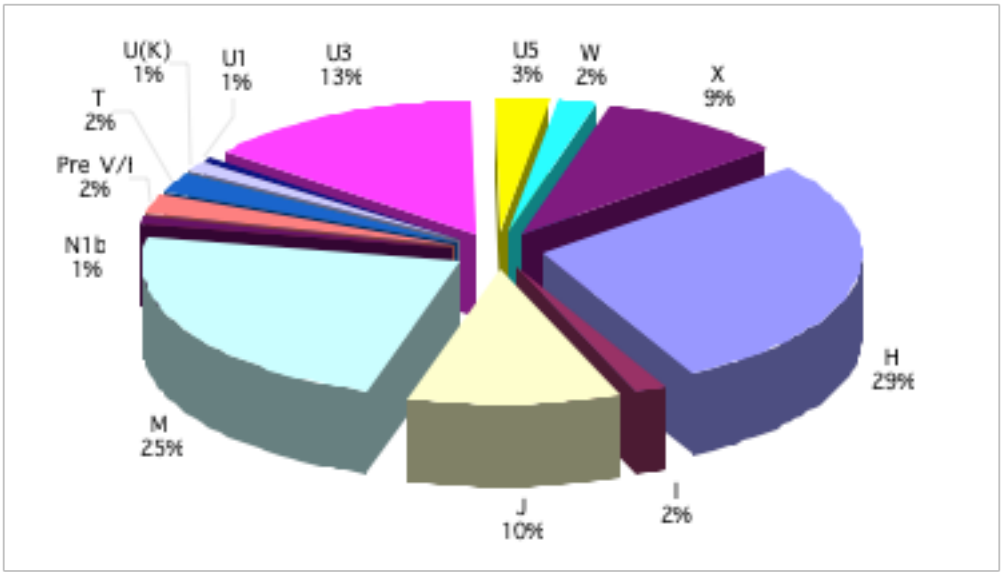

Figure 4-3 Proportional representation of mtDNA haplogroups identified in the Roma

The frequency distribution of mtDNA haplogroups in the Roma differs markedly from that reported for the autochthonous European population (table 4-6). The most striking feature of Romani mtDNA is the high frequency of haplogroup M (defined by +10,394 D*de*I and +10,397 A*lu*I and a HVS1 sequence variant at position 16,223) which is virtually absent in other European populations. Within Europe, this mtDNA haplogroup has only been reported in the Saami (Delghandi, Utsi, & Krauss, 1998) and sporadically in Southeastern European peoples (Richards et al., 2000). Haplogroup M is prevalent in India, occurring at a frequency almost three times that observed in the Roma (Kivisild et al., 1999). Haplogroup H, the most prevalent haplogroup in Europe, is also the most common in the Roma. This haplogroup is geographically ubiquitous, and it is also the most common mtDNA haplogroup in the Near East (Richards et al., 2000). Within India, this haplogroup is rare, but not completely absent. The frequencies of haplogroups U3 and J are much higher in the Roma than observed in the European population. These frequencies are, however, comparable to those estimated for the non-Romani Bulgarians.

Table 4-6
Comparison of mtDNA haplogroup frequencies (%) in Europe, Bulgaria, Spain and India with the Roma

Pop.	H	I	J	M	T	U1	U3	U4	U5	U6	U(K)	Pre HV	W	X	N1b
Europe[1] (	0.50	0.02	0.11	…	0.08	…	0.01	0.01	0.07	0.01	0.07	0.04	0.01	0.02	
Bulgaria[2]	0.23	0.00	0.07	…	0.10	…	0.10	0.07	0.03	…	0.13	0.00	0.00	0.07	
Central Spain[2]	0.52	0.01	0.01	…	0.02	…	0.00	0.01	0.04	…	0.03	0.03	0.02	0.01	
Near East[3]	0.24	0.02	0.09		0.10	0.03	0.05	0.02	0.02		0.05	0.04	0.02	0.03	0.02
India[4]	0.02	0.01	0.01	0.60	0.02		0.13					0.02	0.01		0.01
Roma	0.29	0.02	0.11	0.25	0.02	0.01	0.13	…	0.03	…	0.02	0.02	0.02	0.09	0.01

[1]*Data from Richards et al., (1998) and summarised in Helgason et al., 2000.*
[2]*Data from Simoni et al., 2000.*
[3]*Data from Richards et al., 2000.*
[4]*Data from Kivisild et al., 1999.*

4.1.2.2 Results of HVS1 sequencing

Sequence analysis of the mtDNA HVS1 in the 185 samples identified 61 unique sequences. When combined with the RFLP results this corresponded to 62 unique maternal lineages, as the Cambridge reference sequence (CRS) was found to occur in both haplogroup H and haplogroup Pre V/H (table 4-7).

None of the eleven haplogroup M sequences was found to bear the characteristic East African HVS1 sequence of 16,129, 16,189, 16,223, 16,249, 16,311 (Quintana-Murci et al., 1999), thereby excluding an African origin of this lineage. Thus, haplogroup M sequences in the Roma are most likely of Asian origin.

Three sequences account for almost one-third of all Romani mtDNA. These include the haplogroup U3 sequence defined by a sequence variant at position 16,343, which is the most frequently occurring unique sequence in the Romani sample (12%); the haplogroup M sequence defined by variants at positions 16,129, 16,223, 16,291 & 16,298 (9%); and the haplogroup H sequence defined by variants at positions 16,261 & 16,304 (9%). Conversely, thirty-seven sequences are found just once in the sample.

Table 4-7

Female lineages identified in Romani populations

HVS1 VARIANT(s)	Haplogroup	Total N=185	Turgovzi N=25	Feredjelli N=18	Monteni N=42	Intreni N=16	Lom N=18	Kalderash N=23	Spanish Roma N=25	Lithuanian Roma N=18
CRS*	H	2	.	.	2	.	.	.	.	.
93	H	1	.	.	.	.	.	.	.	1
189	H	1	.	.	.	.	.	1	.	.
223	H	2	.	.	.	.	.	2	.	.
242	H	1	1	.	.	.	.	.	.	.
248	H	1	.	.	.	.	.	.	.	1
261	H	1	.	.	.	.	.	.	1	.
304	H	1	.	.	.	.	.	1	.	.
354	H	3	.	.	.	.	3	.	.	.
362	H	1	.	.	.	.	.	.	1	.
186, 304	H	11	.	.	8	3	.	.	.	.
187, 189	H	1	.	.	1	.	.	.	.	.
218, 278	H	6	.	.	.	.	2	3	1	.
261, 304	H	17	3	.	8	2	1	1	.	2
278, 293, 311	H	1	1	.	.	.	.	.	.	.
51, 145, 304	H	1	1	.	.	.	.	.	.	.
93, 223	H	1	.	.	.	.	1	.	.	.
93, 291	H	2	.	.	1	1	.	.	.	.
129, 172, 223, 311	I	3	1	.	.	.	1	1	.	.
39C, 69, 126	J	1	1	.	.	.	.	.	.	.
69, 126	J	10	.	.	3	2	1	4	.	.
69, 126, 145, 222, 235, 261, 271	J	1	.	.	.	.	.	.	1	.
69, 126, 145, 222, 261, 311	J	3	1	2	.	.	.	.	.	.
69, 126, 193	J	1	1	.	.	.	.	.	.	.
69, 126, 278, 366	J	1	.	.	.	.	.	.	1	.
69, 126, 300	J	1	.	.	.	.	.	.	1	.
69, 126, 311	J	1	1	.	.	.	.	.	.	.
69, 93, 126	J	1	.	.	.	.	.	1	.	.
129, 148, 223, 291, 298	M	2	2	.	.	.	.	.	.	.
129, 223, 230, 233. 304	M	1	.	.	.	1	.	.	.	.
129, 223, 230, 233, 304, 344	M	8	.	.	3	3	1	1	.	.
129, 223, 230, 233, 304, 344, 355	M	4	.	.	3	.	.	1	.	.
129, 223, 256, 291	M	1	.	1	.	.	.	.	.	.
129, 223, 266, 291	M	1	.	.	.	.	.	1	.	.
129, 223, 291	M	8	2	2	.	1	.	3	.	.
129, 223, 291, 298	M	17	.	3	3	2	1	.	4	4
129, 223, 291, 298, 362	M	1	.	.	.	.	.	.	1	.
223, 290, 318T	M	1	.	.	.	.	.	1	.	.
223, 291, 298	M	2	.	.	.	.	2	.	.	.
192A, 320	PRE V/H	3	.	.	3	.	.	.	.	.
CRS*	PRE V/H	1	.	1	.	.	.	.	.	.
126, 294, 296	T	1	.	.	1	.	.	.	.	.
126, 294, 324	T	2	.	.	1	.	.	1	.	.
126, 294, 352	T	1	1	.	.	.	.	.	.	.
222, 224, 261, 311	U(K)	1	.	1	.	.	.	.	.	.
224, 311, 344	U(K)	1	.	.	.	1	.	.	.	.
224, 261, 311	U(K)	1	.	.	.	.	1	.	.	.
183C, 189, 249	U1	1	.	1	.	.	.	.	.	.
343	U3	22	.	.	1	.	.	.	11	10
260, 343	U3	2	.	.	.	.	.	.	2	.
167, 192, 270, 311, 356	U5	1	.	.	.	.	.	.	1	.
189, 270	U5	1	1	.	.	.	.	.	.	.
192, 224, 261, 270	U5	1	1	.	.	.	.	.	.	.
256, 270	U5	1	1	.	.	.	.	.	.	.
28G, 192, 224, 261, 270	U5	1	1	.	.	.	.	.	.	.
172, 223, 231, 292	W	3	1	2	.	.	.	.	.	.
126, 189A, 223, 278	X	9	2	1	3	.	2	1	.	.
92, 126, 189A, 223, 278	X	2	.	2	.	.	.	.	.	.
92, 189A, 223, 278	X	1	.	1	.	.	.	.	.	.
93, 189, 223, 241, 278	X	3	2	1	.	.	.	.	.	.
93, 96T, 189, 223, 241, 278	X	1	.	.	1	.	.	.	.	.
86, 129, 145, 176G, 223	N1b	2	.	.	.	.	2	.	.	.

NB mutated sites are +16,000 in accordance with Anderson et al., (1981) sequence of the mitochondrial genome. All mutations are transitions from the published sequence unless indicated with a letter, which indicates a transversion. *CRS = Cambridge reference sequence and denotes complete identity with the Anderson et al., (1981) sequence.

4.1.2.3 Network analysis of mtDNA haplogroups

4.1.2.3.1 Phylogenetic relationship between Romani mtDNA

In order to examine the evolutionary relationship between Romani mitochondrial lineages, a median joining network (Bandelt, Forster, Sykes, & Richards, 1995) was constructed using all informative sequence variants (figure 4-4). Given the slow mutation rate in the coding region of mtDNA, haplogroup assignments denote classifications of great antiquity that predate the formation of each Romani population. HVS1 variation observed within each haplogroup may represent founding lineages or be due to either female-mediated gene flow or the evolution of new lineages within the population through mutation.

Within the Roma, haplogroup H is represented by 18 unique HVS1 sequences; however, a single sequence (16,261, 16,304) accounts for almost one third of this haplogroup and is widely distributed among populations. This lineage has not been reported in a large survey of Near Eastern and European mtDNA (Richards et al., 2000).

Haplogroup M is represented by eleven unique HVS1 sequences of which nine bear a variant at position 16,129. A transition from the reference sequence at this position defines a subhaplogroup M5[1] (Kivisild et al., 1999). Thus, in the Roma 93.5% of haplogroup M lineages belong to the subhaplogroup M5. The close phylogenetic relationship between these sequences is evident in the network.

Haplogroup U3 occurs at a frequency of 13%, and is almost entirely represented by a single sequence. Previous reports indicate that this lineage is widely dispersed in European and near eastern populations (Richards et al., 2000). It is interesting to note that, within haplogroup X, a transition and transversion are observed at position 16,189. These may be a coincidental finding; however, evidence for sequence-context specific mutability has recently been suggested (Malyarchuk & Derenko, 1999). The possibility of a sequence context effect warrants further investigation. Haplogroup X sequences with a transversion at position 16,129 have not been reported in European, Near Eastern, or Indian populations (Kivisild et al., 1999; Richards et al., 2000)

[1] In a report by Kivisild et al. (1999), the authors denoted this subhaplogroup M4. However, in a later paper by the same authors it is denoted M5 (Bamshad et al., 2001). The more recent nomenclature is used.

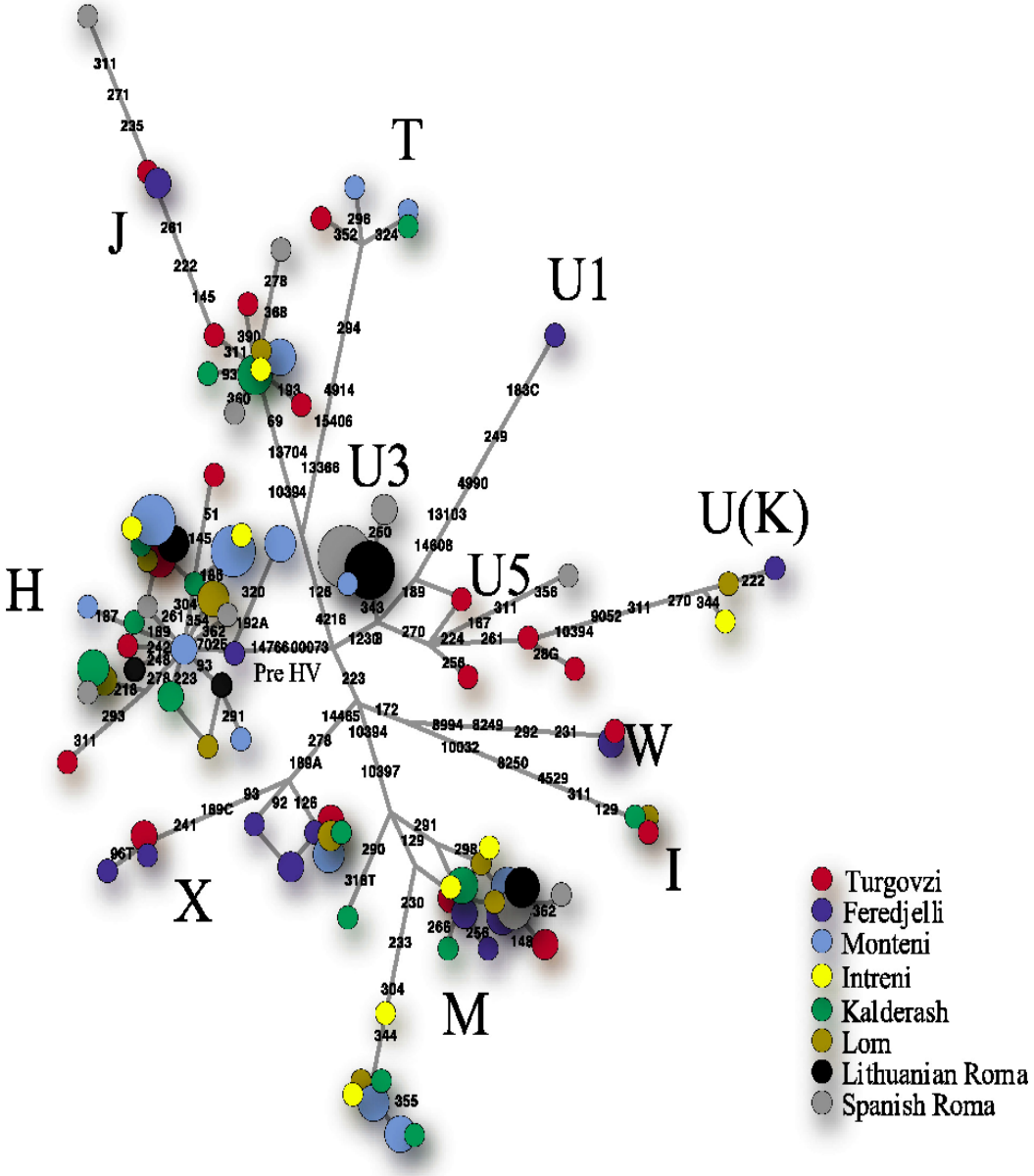

Figure 4-4 Median-joining network of mtDNA sequences identified in the Roma. All 2- and 3-digit numbers are +16,000 according to the Anderson et al., (1981) reference sequence. Nodes are proportional to the frequency of the sequence in the population.

4.1.2.3.2 Network analysis of mtDNA haplogroup M

The variation within mtDNA haplogroup M in the Roma was compared to that described for haplogroup M in Indians (Quintana-Murci et al., 1999; Kivisild et al., 1999), through construction of a median-joining network (figure 4-5). Haplogroup M, which accounts for 60% of all Indian mtDNA, displays a great deal of internal heterogeneity. In the network of haplogroup M, the Romani sequences form a distinct subcluster of sequences of limited diversity. Nine of the eleven Romani haplogroup M sequences are characterised by a variant at position 16,129. Furthermore, six of these sequences are further defined by a sequence variant at position 16,291. Of the sequences that do not bear the M5 diagnostic variant at position 16,129, one is closely related to the other sequences (16,223, 16,291, 16,298). It is interesting to note that position 16129 is known to be a hypermutable site in the mtDNA (Stoneking, 2000; Tully et al., 2000), thus its absence in this sequence might be the result of mutation. The other haplogroup M sequence (16,223, 16,290, 16,318T) is evidently distantly related to all other Romani sequences. A single haplogroup M HVS1 sequence in the Roma, defined by variants at positions 16,129 and 16,291 has been identified in individuals belonging to the Madiga caste in Andhra Pradesh (M. Bamshad, personal communication).

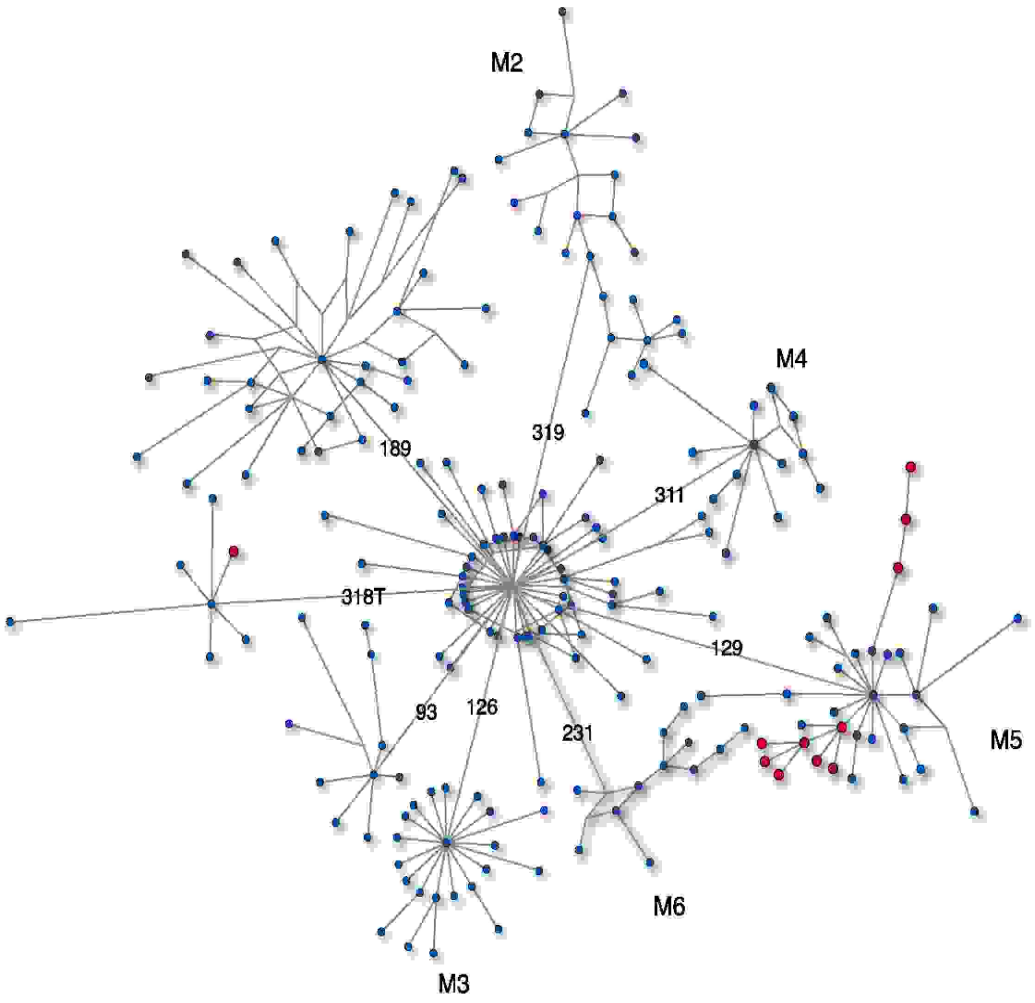

Figure 4-5 Median-joining network of haplogroup M sequences in Indians (Quintana-Murci et al., 1999; Kivisild et al., 1999) and Roma. Haplogroup M sequences identified in the Roma are in red. Subhaplogroup designations and the defining HVS1 variant proposed by Bamshad et al., (2001) are indicated. In addition, three frequently occurring variants that define subclades are shown. All 2- and 3-digit numbers are +16,000 according to the reference sequence (Anderson et al., 1981). Branch lengths are proportional to the number of mutations except those that join subhaplogroups.

4.2 Genetic Relationships Between Romani Populations

4.2.1 Relatedness of Romani Populations as Inferred from Male Lineages

4.2.1.1 Distribution of Y chromosome haplogroups in Romani populations

The frequency of Y chromosome haplogroups within each population was determined (Table 4-8). Haplogroup VI-68 is the only haplogroup present in all populations. The proportional representation of haplogroup VI-68 varies in each population from a minimum value of 11.1% in the Turgovzi to a maximum of 82.4% in the Monteni. Haplogroup VI-68 is the most frequently occurring haplogroup in all Vlach populations. Haplogroup VI-52 was identified in all populations except the Intreni, however, the large proportion of "unknown" lineages in this population precludes an assertion that the haplogroup does not occur in this population. Haplogroup VI-52 is variably represented in the seven populations ranging from 4.8% in the Lithuanian Roma to 52.8% in the Turgovzi.

The most common haplogroup in the Spanish Roma is VI-56 found at a frequency of 33.3%. This haplogroup is also well represented in the Lithuanian Roma (23.8%) but occurs infrequently in other Romani populations (5.6% in the Turgovzi and 10.3% in the Lom). Haplogroup VI-71 is found only in populations speaking Balkan dialects.

Table 4-8
Distribution of Y chromosome haplogroups in Romani populations

Haplogroup	Turgovzi n=36	Feredjelli N=21	Intreni n=17	Monteni n=17	Lom n=19	Kalderash n=11	Spanish Roma N=36	Lithuanian Roma n=21
VI-68	11.1%	19.0%	58.8%	82.4%	68.4%	63.6%	18.5%	47.6%
VI-52	52.8%	23.8%	.	5.9%	15.8%	18.2%	14.8%	4.8%
VI-56	5.6%	.	.	.	10.5%	.	33.3%	23.8%
IX-104	2.8%	14.3%	.	.	5.3%	9.1%	22.2%	9.5%
VI-71	8.3%	23.8%	.	.	.	.	.	.
III-36	5.6%	.	.	5.9%	.	9.1%	.	.
V-52	.	19.0%	.	.	.	.	.	.
IX-108	.	.	.	.	.	.	.	4.8%
VI-57	.	.	.	.	.	.	3.7%	.
VI-58	.	.	.	.	.	.	3.7%	.
Unknown	13.9%		41.2%	5.9%	.	.	3.7%	9.5%

NB The most frequent haplogroup in each population is shaded.

4.2.1.2 Distribution of Y chromosome haplotypes in Romani populations

The frequencies of Y chromosome haplotypes in the eight populations were determined (table 4-9). Haplotype VI-68-a (15-22-10-11-12-16-14) is modal in the sample of 165 chromosomes, representing 26.6% of all Y chromosomes in the Roma. Furthermore, haplotype VI-68-a is found in every population, whereas no other haplotype within the Roma is shared by more than four populations. The frequency of haplotype VI-68-a varies within each population, and it is modal in only the Intreni (0.529), Monteni (0.706) and Kalderash (0.545). However, the closely related haplotype, VI-68-c, is modal in the Lom and Lithuanian Roma. Thus, in all Vlach populations and the Lithuanian Roma, modal haplotypes belongs to haplogroup VI-68. The Turgovzi and Feredjelli have a common modal haplotype, VI-52-a, but otherwise do not share any haplotypes. The modal haplotype in the Spanish Roma is VI-56-b, which is also found at lower frequencies in the Turgovzi, Lom and Lithuanian Roma.

Table 4-9

Y chromosome haplotype frequencies in Romani populations

Haplotype	Total N=165	Turgovzi N=36	Feredjelli N=21	Intreni N=17	Monteni N=17	Lom N=15	Kalderash N=11	Spanish Roma N=27	Lithuanian Roma N=21
VI-68-a		0.111	0.19	0.529	0.706	0.133	0.545	0.185	0.143
VI-68-b		0	0	0	0	0	0	0	0.095
VI-68-c		0	0	0.059	0	0.467	0	0	0.238
VI-68-d		0	0	0	0.059	0	0.091	0	0
VI-68-e		0	0	0	0.059	0	0	0	0
VI-52-a		0.389	0.238	0	0.059	0	0	0	0
VI-52-b		0.083	0	0	0	0	0	0	0
VI-52-c		0.028	0	0	0	0.067	0.091	0	0
VI-52-d		0.028	0	0	0	0	0.091	0	0
VI-52-e		0	0	0	0	0	0	0	0.048
VI-52-f		0	0	0	0	0	0	0.037	0
VI-52-g		0	0	0	0	0	0	0.037	0
VI-52-h		0	0	0	0	0	0	0.037	0
VI-52-I		0	0	0	0	0.067	0	0	0
VI-52-j		0	0	0	0	0	0	0.037	0
VI-52-k		0	0	0	0	0.067	0	0	0
VI-56-a		0.028	0	0	0	0	0	0.037	0
VI-56-b		0.028	0	0	0	0.133	0	0.222	0.143
VI-56-c		0	0	0	0	0	0	0.037	0.048
VI-56-d		0	0	0	0	0	0	0	0.048
VI-56-e		0	0	0	0	0	0	0.037	0
IX-104-a		0	0.095	0	0	0	0	0	0
IX-104-b		0	0.048	0	0	0	0	0	0
IX-104-c		0.028	0	0	0	0	0.091	0	0
IX-104-d		0	0	0	0	0	0	0.037	0.048
IX-104-e		0	0	0	0	0	0	0.074	0.048
IX-104-f		0	0	0	0	0	0	0.037	0
IX-104-g		0	0	0	0	0	0	0.037	0
IX-104-h		0	0	0	0	0	0	0.037	0
IX-104-I		0	0	0	0	0.067	0	0	0
VI-71-a		0	0.048	0	0	0	0	0	0
VI-71-b		0	0.143	0	0	0	0	0	0
VI-71-c		0	0.048	0	0	0	0	0	0
VI-71-d		0.083	0	0	0	0	0	0	0
III-36-a		0.028	0	0	0	0	0	0	0
III-36-b		0.028	0	0	0.059	0	0.091	0	0
V-52-a		0	0.048	0	0	0	0	0	0
V-52-b		0	0.095	0	0	0	0	0	0
V-52-c		0	0.048	0	0	0	0	0	0
IX-108-a		0	0	0	0	0	0	0	0.048
VI-57-a		0	0	0	0	0	0	0.037	0
VI-58-a		0	0	0	0	0	0	0.037	0
Ht-a		0.056	0	0	0	0	0	0	0
Ht-b		0.028	0	0	0	0	0	0	0
Ht –c		0.028	0	0	0	0	0	0	0
Ht –d		0.028	0	0	0	0	0	0	0
Ht –e		0	0	0	0	0	0	0	0.048
Ht –f		0	0	0	0	0	0	0	0.048
Ht –g		0	0	0	0	0	0	0.037	0
Ht –h		0	0	0.059	0	0	0	0	0
Ht –j		0	0	0.235	0	0	0	0	0
Ht –k		0	0	0.118	0	0	0	0	0
Ht –l		0	0	0	0.059	0	0	0	0

NB Modal haplotypes within each population are highlighted.

4.2.1.3 Male-specific genetic distances between Romani populations

Population pairwise R_{ST} values were computed for the eight populations using Y STR data (table 4-10). In general, the genetic distances are smaller between populations within the same migrational grouping. It is apparent that the greatest genetic distances are observed between Rudari and Balkan populations. This is striking, given that genetic distances between each of these populations and the geographically distant Spanish and Lithuanian Roma are smaller. This relationship is observed in a neighbour-joining tree depicting the distance matrix (figure 4-6). Populations cluster primarily on the basis of migrational/linguistic groupings. The Lithuanian and Spanish Roma are placed between the Vlach and Balkan Roma.

Table 4-10
Matrix of population pairwise R_{ST} values

	Feredjelli	Turgovzi	Lithuanian Roma	Spanish Roma	Intreni	Monteni	Kalderash
Turgovzi	0.025						
Lithuanian Roma	0.120	0.128					
Spanish Roma	0.059	0.099	0.007				
Intreni	0.383	0.276	0.144	0.223			
Monteni	0.251	0.206	0.103	0.167	0.043		
Kalderash	0.130	0.143	0.105	0.072	0.180	0.038	
Lom	0.106	0.100	-0.039	0.011	0.084	0.024	0.031

NB Significant values ($P<0.05$ from 1000 permutations) are highlighted.

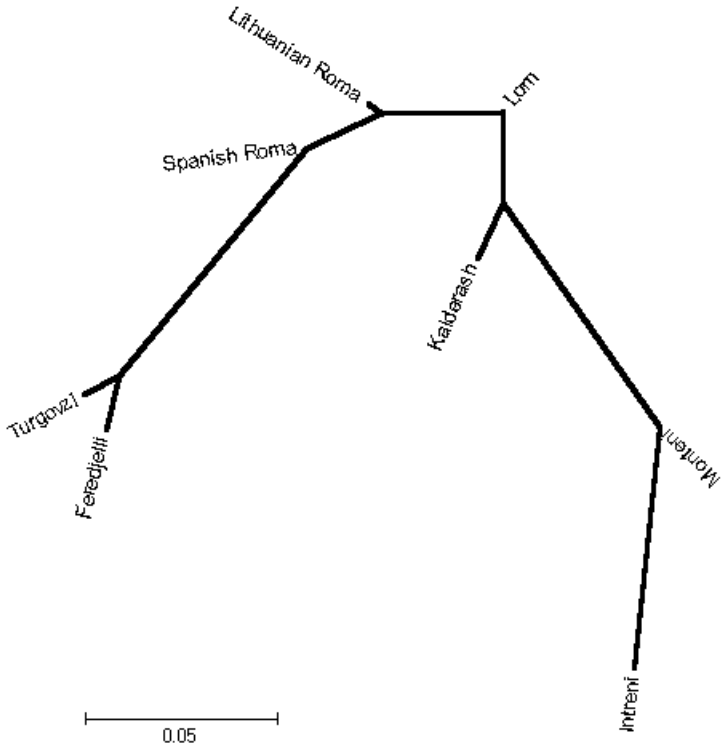

Figure 4-5 Unrooted neighbour-joining tree based on population pairwise R_{ST} distances determined using Y STR data

4.2.1.4 Genetic structure of Y chromosome diversity in the Roma

To examine population structure in the Roma, an analysis of molecular variance (AMOVA) was performed. Molecular variance was assessed under a number of population groupings (table 4-11). The greatest variation amongst groups was apportioned when populations were grouped according to migrational groupings (13.28%, P<0.05). Conversely, this grouping yielded the smallest apportionment of variation to populations within groups (1.45%, P<0.05). Grouping populations by nationality yielded the lowest apportionment of genetic variation amongst groups, being effectively zero. Grouping populations by religion yielded a high apportionment of variation amongst groups (11.97% P<0.05). Grouping by metagroups did not produce statistically significant apportionment of variation amongst groups.

Table 4-11
Apportionment of molecular variance for Y STR data under different population groupings

Group definition	Variation amongst groups	Variation among populations within groups	Variation within populations
Whole population N=1		12.07% P<0.00001	87.93% P<0.00001
Nationality N=3	-6.83% P=0.80645	16.07% P<0.00001	90.76% P<0.00001
Religion N=4	11.97% P=0.01857	1.89% P=0.03519	86.14% P<0.00001
Metagroup N=5	7.16% P=0.17693	5.93% P<0.00001	86.91% P<0.00001
Major migrations N=3	13.28% P=0.01173	1.45% P=0.01369	85.27% P<0.00001

NB N refers to number of defined groups in analysis

4.2.2 Relationship Between Romani Populations as Inferred from Female Lineages

4.2.2.1 Distribution of mtDNA Haplogroups in Romani populations

The proportional representation of mtDNA haplogroups in each Romani populations was analysed (table 4-12). Haplogroup M is the only haplogroup that is found in all Romani populations, ranging from 16.0% in the Turgovzi to 43.8% in the Intreni. Haplogroup H is found at highest frequencies in Vlach populations. Within the Balkan populations, haplogroup H accounts for almost one-quarter of female lineages in the Turgovzi, but is entirely absent in the Feredjelli. Haplogroup U3 is the most frequent haplogroup in both the Spanish and Lithuanian Roma, but is otherwise absent in Romani populations aside from the Monteni in which it is rare. Conversely, haplogroup X is found in all Vlach and Balkan Romani populations except the Intreni, but is not observed in the Lithuanian or Spanish Roma.

Table 4-12
Distribution of mtDNA haplogroups in Romani populations (values in %)

Population	H	I	J	M	Pre V/H	T	U(K)	U1	U3	U5	W	X	N1b
Turgovzi N=25	24.0	4.0	16.0	16.0	0	4.0	0	0	0	16.0	4	16.0	0
Feredjelli N=18	0	0	11.1	33.3	5.6	0	5.6	5.6	0	0	11.1	27.8	0
Monteni N=42	47.6	0	7.1	21.4	7.1	4.8	0	0	2.4	0	0	9.5	0
Intreni N=16	37.5	0	12.5	43.8	0	0	6.25	0	0	0	0	0	0
Lom N=18	38.9	5.6	5.6	22.2	0	0	5.6	0	0	0	0	11.1	11.1
Kalderash N=23	34.8	4.3	21.7	30.4	0	4.3	0	0	0	0	0	4.3	0
Spanish Roma N=25	12.0	0	12.0	20.0	0	0	0	0	52.0	4.0	0	0	0
Lithuanian Roma N=18	22.2	0	0	22.2	0	0	0	0	55.6	0	0	0	0

NB Modal haplotypes within each population are highlighted.

4.2.2.2 Distribution of HVS1 sequences in Romani populations

The distribution of mtDNA HVS1 sequences in the seven Romani populations is provided in table 4-7. No single mtDNA HVS1 sequence was found to be common to all populations. However, at least one of two haplogroup M sequences that differ only by a sequence variant at position 16,298 is found in every population. The haplogroup U3 sequence with a single variant at position 16,343 is the most frequently encountered sequence in the entire sample. However, the two U3 sequences bearing the 16,343, and the 16,343 and 16,260 variants respectively are very common in the Spanish and Lithuanian Roma and otherwise are found only once, in the Monteni. The haplogroup X sequence with variants at positions 16,126, 16,189A, 16,223, and 16,278 occurs in every Balkan and Vlach population. This sequence and other haplogroup X sequences are absent in the Spanish and Lithuanian Roma. The haplogroup J sequence with mutations at positions 16,069 and 16,126 is found in the Vlach groups but in neither of the Balkan populations nor in the Spanish and Lithuanian Roma. In contrast to the sequence lineages that adhere to population groupings, the haplogroup H sequence containing the 16,261 and 16,304 variants appears randomly distributed - present in all populations except the Feredjelli and Spanish Roma.

4.2.2.3 Female-specific genetic distances between Romani populations

Population pairwise genetic distances between the seven populations were computed using intermatch-mismatch distances (table 4-13).

Table 4-13
Intermatch-mismatch distances between populations

	Feredjelli	Turgovzi	Lithuanian Roma	Spanish Roma	Intreni	Monteni	Kalderash
Turgovzi	0.050						
Lithuanian Roma	0.769	0.480					
Spanish Roma	0.644	0.360	-0.060				
Intreni	0.642	0.380	0.657	0.750			
Monteni	0.647	0.221	0.559	0.607	-0.017		
Kalderash	0.148	0.003	0.506	0.390	0.178	0.167	
Lom	0.001	-0.020	0.419	0.356	0.264	0.257	-0.052

NB Significant values (P<0.05 from 1000 permutations) are highlighted.

Generally, the largest genetic distances are observed between the Lithuanian or Spanish Roma, and Balkan or Vlach Roma. Significant distances are seen between Balkan and Rudari (Intreni and Monteni) populations. However, the distances between the two Balkan populations and the Kalderash and Lom are small and statistically insignificant. Negative distances are observed between the Lithuanian and Spanish Roma, the Intreni and Monteni, the Lom and Kalderash, and the Lom and Turgovzi.

An unrooted neighbour-joining tree was created from this distance matrix (figure 4-7). As can be seen in the tree, the Lithuanian and Spanish Roma cluster closely together as do the Intreni and Monteni. The Kalderash, Turgovzi, Lom and Feredjelli are separated from these populations but, do not appear to adhere to any particular clustering.

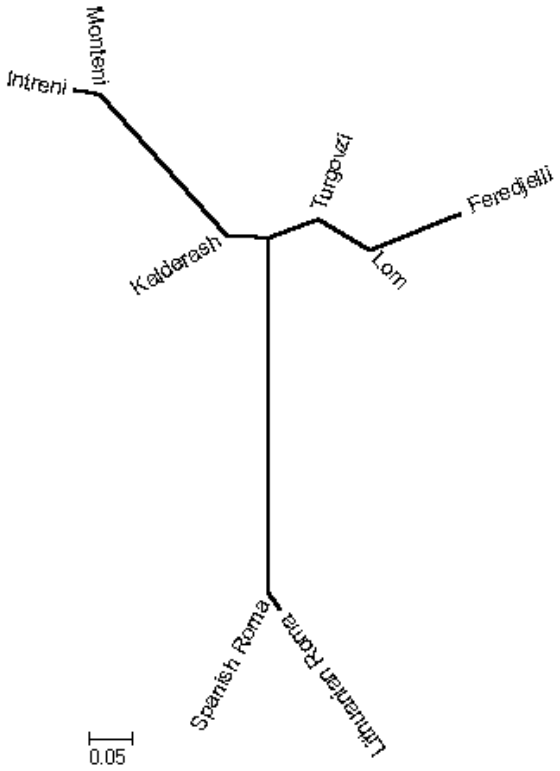

Figure 4-6 Unrooted neighbour-joining tree based on mtDNA HVS1 intermatch-mismatch distances between Romani populations.

4.2.2.4 Genetic Structure of mtDNA diversity in the Roma

In order to examine female-specific genetic structure in the Roma, an analysis of molecular variation (AMOVA) was performed using mtDNA HVS1 data. A variety of population groupings were created to explore different demographic and social parameters and their relevance to genetic structuring (table 4-14). These different groupings did not yield dramatically different apportionments of variation amongst groups. The greatest apportionment of genetic variation among groups was observed when populations were grouped according to historical migrations (6.85%). However, the reciprocal apportionment of variation among populations within groups was not statistically significant (1.59%, P=0.05865). Statistical significance was reached for all three variance components when populations were grouped by metagroups. This grouping apportioned 6.73% (P<0.05) of genetic variation among groups, and only 0.62% (P<0.00001) of variation among populations within groups.

Table 4-14
Apportionment of molecular variance for mtDNA data under different population groupings

Group definition	Variation amongst groups	Variation among populations within groups	Variation within populations
Whole population N=1		6.54% P<0.00001	93.46% P<0.00001
Nationality N=3	4.87% P=0.04301	4.20% P<0.00001	89.73% P<0.00001
Religion N=4	6.75% P=0.00978	1.03% P=0.14467	92.22% P<0.00001
Metagroup N=5	6.73% P=0.01369	0.62% P<0.00001	92.65% P<0.00001
Major migrations N=3	6.85% P=0.01857	1.59% P=0.05865	91.56% P<0.00001

NB N refers to number of groups in analysis

4.2.2.5 Relatedness of Roma to worldwide populations as determined using female lineages

The genetic relatedness of the Roma to other European and worldwide populations was calculated from mtDNA data by determining intermatch-mismatch population pairwise distances. A neighbour-joining tree was constructed that displays the relationship between these populations (Figure 4-7). In this tree it is apparent that Romani populations form distal branches that are separated from all other populations. The branch distances between Romani populations are great, which reflects genetic sub-structure and genetic divergence of these populations from each other. This contrasts with the autochthonous European populations that cluster closely together, reflecting a low level of genetic substructure. The Roma are situated at greatest distance to Middle Eastern populations and lie approximately midway between European and Asian populations.

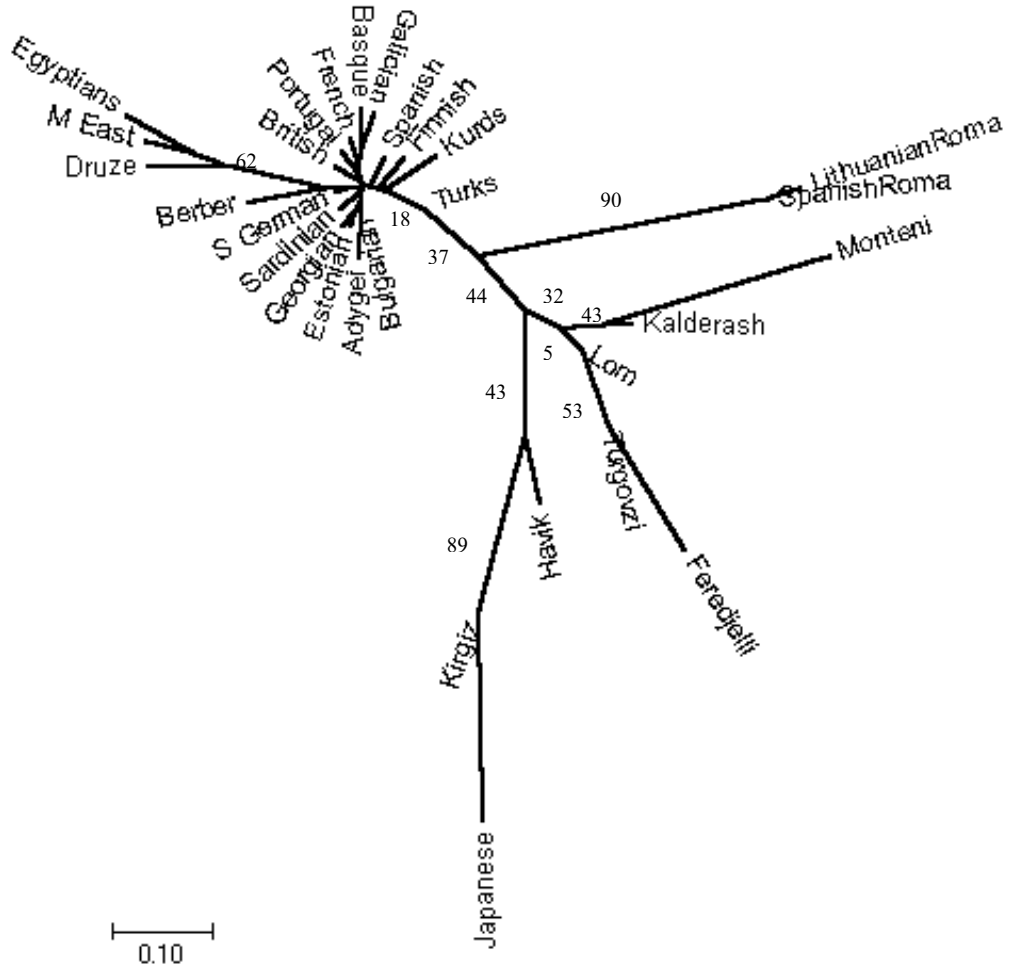

Figure 4-7 Unrooted neighbour-joining tree depicting intermatch-mismatch population pairwise genetic distances between Romani and worldwide populations as determined from mtDNA data. Bootstrap values for major branches are in percentage values from 1000 iterations.

4.3 Intrapopulation Genetic Diversity of Romani Populations

4.3.1 Intrapopulation Analysis of Paternal Lineages

Y chromosomal variation differed greatly in the eight populations, as calculated by the mean number of pairwise differences and the haplotype diversity index. The data are presented in descending order according to the average pairwise differences (Table 4-15).

Table 4-15
Diversity indices for Y chromosome haplotypes in Romani populations

Population	No. Y Chrs	Average Pairwise Differences	Haplotype Diversity
Spanish Roma	27	3.72 +/- 1.94	0.926 +/- 0.001
Feredjelli	21	3.35 +/- 1.79	0.900 +/- 0.002
Turgovzi	36	3.10 +/- 1.65	0.835 +/- 0.003
Lithuanian Roma	21	3.02 +/- 1.64	0.900 +/- 0.001
Kalderash	11	2.82 +/- 1.61	0.728 +/- 0.022
Lom	15	2.74 +/- 1.54	0.781 +/- 0.011
Monteni	17	1.50 +/- 0.95	0.514 +/- 0.023
Intreni	17	1.16 +/- 0.79	0.684 +/- 0.010

Y chromosome haplotype diversity is greatest in the Spanish Roma with 17 haplotypes observed in the sample of 27 males. Similarly high diversity indices were determined for the Feredjelli, Turgovzi and Lithuanian Roma, all of whom have an average of more than 3 pairwise differences between haplotypes, and haplotype diversities greater than 0.8. In contrast, the two Rudari population, the Monteni and the Intreni, show a strikingly low male genetic diversity, with averages of less than 1.6 differences between haplotypes within the population and haplotypes diversities below 0.7. Y chromosome diversity within the Lom and Kalderash is of an intermediate value relative to the other populations.

4.3.2 Intrapopulation Analysis of Maternal Lineages

Three different diversity indices were calculated from mtDNA data. The data are presented in descending order according to the average pairwise differences (table 4-16).

Table 4-16
Diversity indices for mtDNA data in Romani populations

Population	N	K	A	D'	Π	P
Feredjelli	18	12	22	0.9542	0.016449	5.92
Turgovzi	25	19	31	0.9767	0.016407	5.91
Lom	18	12	22	0.9542	0.013725	4.94
Intreni	16	9	15	0.9250	0.013356	4.80
Monteni	42	15	25	0.9129	0.013111	4.72
Kalderash	23	15	22	0.9486	0.012330	4.44
Spanish Roma	25	11	23	0.7933	0.010741	3.87
Lithuanian Roma	18	5	9	0.6601	0.007298	2.63
All	**185**	**61**	**61**	**0.9562**	**0.013702**	**4.93**

NB N= sample size, K =no. of unique sequences, A= no. of variable positions, D'= sequence diversity, π = nucleotide diversity, P= average number of pairwise differences.

When each Romani population is analysed independently a large variation in mtDNA diversity is apparent. The two Balkan Romani populations, the Turgovzi and Feredjelli, are extremely diverse with average pairwise differences values of 5.91 and 5.92 respectively and high sequence and nucleotide diversities. These values place them amongst the most heterogeneous European populations based on mtDNA (see table 4-17 for examples of European populations). In contrast, extremely low levels of mtDNA diversity are indicated by the same statistics calculated for the Spanish and Lithuanian Roma, and the diversity values determined for these populations are very low when compared to other European populations. Female specific diversity in the Vlach Romani populations is intermediate in comparison to the other Romani populations.

Comparisons to results from studies of other populations (table 4-17) show that the entire European Romani population is well within the range of genetic diversities observed in other populations. An average number of 4.96 pairwise differences between mtDNA sequences in the Roma is indicative of greater diversity than European genetic isolates, such as the Icelanders, Sardinians and Saami.

Table 4-17

Average number of pairwise differences of mtDNA sequences in the Roma and other populations

Population	N	Average pairwise differences
Turkish[1]	96	5.45
Spanish[1]	89	5.02
European Roma	**169**	**4.93**
Iceland[2]	73	4.40
Sardinians[2]	69	4.22
Saami[2]	115	3.99

[1]Data summarised in Salas, Comas, Lareu, Bertranpetit, & Carracedo (1998).

[2]Data summarised in Arnason, Sigurgislason, & Benedikz (2000)

CHAPTER 5

DISCUSSION

5.1 Genetic Evidence for the Origins of the Roma

5.1.1 The Composition and Origin of Romani Male Lineages

The determination of deep-rooted paternal lineages in the Roma revealed a single predominant Y chromosome haplogroup. Y chromosomes belonging to haplogroup VI-68 account for almost two-fifths of Y chromosomes in the sample of 169 Romani males. Previously, this haplogroup has only been found in India and Pakistan, where it occurs relatively infrequently, and in Central Asia where it appears to be rare (Underhill et al., 2000). This suggests that Y chromosome haplogroup VI-68 is restricted to populations of the Indian subcontinent and proximate geographical locations. Hence, its occurrence in the European Roma points to the Indian origin of Romani males. The prevalence of this haplogroup in the Roma (almost twice as frequent as the next most frequent haplogroup, VI-52), justifies the assertion that it is representative of the founding population.

Y STR haplotype analysis revealed the restricted diversity within this haplogroup with only five unique haplotypes identified. The most frequent of these haplotypes, VI-68-a, represents 71% of the 63 Y chromosomes belonging to this haplogroup. Furthermore, the four other haplotypes within this haplogroup are closely related, separated by single mutations at YSTR loci. This suggests that diversity within this haplogroup has been generated through mutation rather than male-mediated gene flow. The coalescent age of the VI-68 haplogroup in the Roma was dated at 2,020 years before present (95% CI 1,786-2,344 years). This date can be roughly equated with a profound bottleneck event in the proto-Roma around 2,000 years ago. It is conceivable that this event represented the splitting of a small population from a larger parental population.

Haplogroup VI-56 represents 10.7% of the Romani Y chromosomes. This haplogroup has been mainly identified in Middle Eastern populations (Underhill et al., 2000). Y STR haplotypes within this haplogroup are closely related; suggesting that haplogroup diversity in the Roma has arisen largely through mutation. Thus, it is likely that this lineage was introduced into the Romani population by a limited number of related male founders. The apparent restriction of this haplogroup to Middle Eastern populations suggests a possible contribution by non-Indian and non-European peoples to the Romani gene pool. Based on linguistic evidence, Roma are believed to have had extended stays in Persia and Armenia prior to their arrival in Europe (Fraser, 1992; Hancock, 1999a). The presence of haplogroup VI-56 suggests that the Roma underwent some degree of male-mediated admixture during these sojourns. The age of the VI-56 haplogroup was determined to be 2,435 (95%CI 2,135-2,826) years before present, which is older than the Indian specific haplogroup in the Roma. This is possibly due to the fact that the history of the Y chromosome may be confounded by population bottlenecks (de Knijff, 2000). However, it is possible that some diversity was already present within a Middle Eastern male population when it fused with Indian migrants, which would account for the calculated age of the haplogroup.

In striking contrast to haplogroups VI-68 and VI-56, haplogroup VI-52 displays considerable internal heterogeneity. This haplogroup accounts for 21% of the Romani Y chromosomes, but comprises 12 relatively distantly related haplotypes. Network analysis of this haplogroup reveals a complex topology with different haplotypes separated by numerous mutations. In a survey of world-wide populations, this haplogroup was only found in European populations (Underhill et al., 2000). Another study showed it to be a common European lineage with a strong East to West clinal distribution (Semino et al., 2000). Thus, the most likely explanation for this haplogroup in the Roma is through multiple independent admixture events. The clinal distribution of this haplogroup in Europe implies that the majority of males would have come from Eastern European populations. The long-term Eastern European residency of most of the Romani populations included in this study is consistent with this claim.

Haplogroup IX-104 accounts for 8% of patrilineages. This haplogroup is found in populations throughout the world including Africans, Europeans, Middle Easterners,

Indians, Americans and Australians (Underhill et al., 2000). Within Europe, it accounts for over half of all the male lineages in Western Europe and almost a fifth of lineages in Eastern Europe (Semino et al., 2000). However, the worldwide distribution of haplogroup IX-104 makes attempts to discern its origins in the Roma problematic. Network analysis shows that this haplogroup is composed of five haplotypes that are closely related and an additional four more distantly related haplotypes. It is possible that Y chromosomes belonging to this haplogroup may have different histories of admixture in the Roma. The close evolutionary relatedness of some of the haplotypes suggests that some of these admixture events may have occurred sufficiently long ago to allow diversity through mutation to arise within the population.

As well as the aforementioned four haplogroups, an additional six haplogroups were identified in the Roma. Each of these haplogroups represents 5% or less of the known Y chromosomes, but are nonetheless important as they may provide evidence of the initial composition of the Roma, or of subsequent admixture events. Four of these haplogroups, which are represented by just one chromosome in the sample, are found in Indian populations. These include haplogroups VI-71, IX-108, VI-57 and VI-58. Of these haplogroups, VI-57 is the only one that has not been found in European or Middle Eastern populations and thus can justifiably be considered as descended from an Indian male progenitor. Interestingly, a single haplogroup identified in the Roma, III-36, has only previously been found in African populations from Ethiopia, southern Africa and in the Khoisan (Underhill et al., 2000). Although this haplogroup only represents 2.4% of the Romani Y chromosomes, its presence in the Roma is intriguing. Hancock (1999b; 2000) argues that the proto-Roma comprised a military force, which may have included some East Africans. Further studies are required to illuminate the history of this haplogroup in the Roma.

5.1.2 The Composition and Origin of Romani Female Lineages

Nearly one-quarter of Romani matrilineages belong to the mtDNA haplogroup M. This haplogroup is generally considered to be very rare in Europe, where it has only been found in population outliers such as the Saami (Delghandi, Utsi, & Krauss, 1998). Haplogroup M is estimated to represent 60% of maternal lineages in India (Kivisild et

al., 1999). It is also occurs in East African populations, however HVS1 sequence variants allow the discrimination between the Asian and East African subtypes of haplogroup M (Quintana-Murci et al., 1999). Haplogroup M sequences in the Roma can confidently be assigned to the Asian haplogroup M lineage. Studies of Indian mtDNA demonstrated the enormous amount of variation within haplogroup M (Bamshad et al., 2001; Kivisild et al., 1999). Network analysis makes it strikingly apparent that the haplogroup M sequences in the Roma comprise a small subset of the diversity observed within India. Moreover, Romani haplogroup M sequences are closely related and overwhelmingly belong to the mtDNA subhaplogroup M5. This suggests that the variation in haplogroup M observed in the Roma has resulted from mutation rather than heterogeneous origins. Furthermore, the limited diversity in extant sequences can be explained by a small number of related female founders.

Determination of the Indian population that is most closely related to the Roma requires close analysis of this subhaplogroup. Within India, it is unclear whether haplogroup M5 is more prevalent in particular populations (Bamshad et al., 2001; Kivisild et al., 1999). A single haplogroup M sequence (16,129, 16,291) is shared by the Roma and Indians described by Kivisild et al., (1999). These individuals belong to the Madiga caste in the upper east coast of Andhra Pradesh (M. Bamshad, pers comm). However, conclusions cannot be drawn on the basis of a single sequence. Further studies of Indian populations might serve to illuminate related populations.

The most frequently occurring maternal lineage in the Roma sample is haplogroup H, which accounts for 29% of all mtDNA. Haplogroup H is the most common haplogroup in Europe (Richards, Macaulay, Bandelt, & Sykes, 1998) and the Near East (Richards et al., 2000). It also occurs in India, however it is infrequent and represents just 2% of mtDNA in a sample of over 500 individuals (Kivisild et al., 1999). Within the Roma, haplogroup H is heterogeneous. Network analysis shows that the eighteen haplogroup H HVS1 sequences form a cluster of nodes of variable evolutionary relationships. Thirty-one percent of the haplogroup is represented by a single HVS1 sequence defined by mutations at positions 16,261 and 16,304. This points to the antiquity of this mtDNA sequence in the Roma. The geographically widespread

distribution of haplogroup H makes it difficult to assign any of the sequences to hypothesised parental populations.

Haplogroup U3 is the third most frequent mtDNA haplogroup in the Roma. As is apparent in the network analysis, this haplogroup is represented primarily by a single HVS1 sequence. This points to the close biological relatedness of individuals bearing haplogroup U3. Haplogroup U3 is relatively uncommon in Europe (Helgason, Sigurethardottir, Gulcher, Ward, & Stefansson, 2000), although the frequency varies in different European populations (Simoni, Calafell, Pettener, Bertranpetit, & Barbujani, 2000). Within the Near East it occurs at a frequency of 5% (Richards et al., 2000), and it has not been reported in Indian populations (Kivisild et al., 1999). Thus, this haplogroup was most likely introduced into the Roma at some stage subsequent to their exit from India. The almost complete absence of variation within this haplogroup suggests its introduction by a limited number of related individuals.

The other mtDNA haplogroups in the Roma are all found in Indian, Near Eastern and European populations. Thus, determining the population origins of these female lineages is problematic. This might be overcome through identification of subgroups within these haplogroups and determination of their distribution.

5.2 Genetic Relationships between Romani Populations

Highly resolved paternal and maternal lineages shared amongst Romani populations provide evidence of genetic relatedness. This relatedness can be due to either common origins or gene flow. Relationships between populations can be quantified through genetic distance analysis. Whilst interpretations must be made cautiously owing to the confounding effects of genetic drift, careful consideration of the data permits a number of conclusions.

5.2.1 Male Specific Genetic Structure in the Roma

Haplogroup VI-68 is the only Y chromosome haplogroup that is found in every Romani population. Therefore, these separated populations are related to each other through a common Y chromosome haplogroup of Indian origin. The frequency of this haplogroup in different Romani populations varies widely. The Balkan Romani

populations display low frequencies of this haplogroup as do Spanish Roma. In these populations haplogroup VI-68 represents less than 1 in 5 Y chromosomes. The population with the highest frequency of the VI-68 haplogroup is the Monteni, in which it accounts for 82.4% of all Y chromosomes. The Monteni belong to the Rudari metagroup, as do the Intreni where VI-68 represents at least 58.8% of Y chromosomes. These two populations migrated to Bulgaria from Rumania after the end of slavery in the nineteenth century (Marushiakova & Popov, 1997). Similarly, the Kalderash and Lom emigrated from Wallachia and Moldavia. In these two populations, Y chromosome VI-68 represents over 60% of the males. Thus, it is apparent that Vlach Romani groups are characterised by higher frequencies of the Indian-specific Y chromosome haplogroup VI-68 than other Romani groups.

Haplotype analysis reveals the close biological affinity of males with Y chromosomes belonging to VI-68 in these eight populations. Of the 52 YSTR haplotypes that were identified in the Roma, only one was common to all populations. This lineage, VI-68-a, accounts for 27% of Y chromosomes and thus can be referred to as the Romani modal male lineage. The occurrence of a common highly resolved male lineage in separated population has been observed previously only in the Ashkenazi and Sephardic Cohen priests (Thomas et al., 1998). For Jewish priests, a common Y chromosome is not unexpected owing to the paternal inheritance of the vocation. In the Roma, the presence of an identical male lineage in every population points to the common origin of these populations and the long-term preservation of group identity. The genesis of each Romani population can be understood as a process of population fission from a parental population. For the VI-68-a lineage to be found in every extant Romani population, it must have been highly represented in the parental population. This provides strong evidence for long-term endogamous practices in the Roma and proto-Roma.

Although it is found in all populations, the Romani modal male lineage occurs at differing frequencies. In the Turgovzi, Feredjelli and Spanish Roma VI-68-a represents the only lineage belonging to haplogroup VI-68. This suggests that these populations were formed from a small number of founding males from the parental population of Indian immigrants. In contrast, in the other five populations at least two haplotypes are

found within the VI-68 haplogroup. The diversity within haplogroup VI-68 in these populations may be due to a combination of a greater number of founding males bearing this lineage and lesser levels of subsequent admixture, which would serve to maintain high frequencies of haplogroup VI-68.

Genetic distances and haplotype sharing indicate a general trend in the genetic structure of Romani males. A neighbour-joining tree constructed from R_{ST} population pairwise distances shows distinct clustering of populations according to historical migrations. This is reflected in AMOVA results that indicate that the largest variation amongst groups (and thereby, the least variation between populations within groups) is observed when populations are grouped according to major historical migrations. Thus, genetic structure of the Romani male population corresponds to historical divisions arising from major migrations within Europe, rather than nationality. This points to the maintenance of group identity following population fissions, and limited gene flow between historically and socially separated populations that are now geographically proximate.

5.2.2 Female Specific Genetic Structure in the Roma

The Asian-specific haplogroup M is the only mtDNA haplogroup found in all Romani populations. Fifty-six percent of Romani mtDNA belonging to this haplogroup have one of two HVS1 sequences that differ by a single mutation. One or other of these sequences is found in every population, with the Feredjelli the only population in which both occur. Furthermore, every Romani population, with the exception of the Lithuanian Roma, has additional related haplogroup M sequences. The presence of these closely related maternal lineages attests to the common biological ancestry of females in Romani populations.

The only other female lineage with a widespread distribution in the Roma is the haplogroup H sequence, with mutations at position 16,261 and 16,304. This lineage is found in five populations, but is absent in the Feredjelli and Spanish Roma. Nevertheless, its distribution in geographically and historically distant groups points to the relatedness of Romani populations. The population origins of this lineage cannot be discriminated; however, its widespread distribution in the Roma points to its early

existence, possibly in India but more likely in the Middle East or soon after arrival in Europe.

Other mtDNA lineages testify to the independent histories of the populations. Haplogroup X occurs in the Balkan and Vlach populations, but is absent in the Lithuanian and Spanish Roma. This haplogroup is subdivided in the Roma by a transition or transversion at position 16,189. A haplogroup X HVS1 lineage with a transversion at position 16,189 is found in every Balkan and Vlach Romani population. Its prevalence in Balkan and Vlach Romani populations suggests that it is a female founding lineage. However, it is completely absent in the Spanish and Lithuanian Roma. Early historical records report Romani groups of 30-400 people (Fraser, 1992), and it is possible that very few females travelled west with these small migrating groups. Therefore, it is plausible that some lineages were not represented in these populations or else were subsequently lost through genetic drift.

The haplogroup J sequence, defined by mutations at positions 16,069 and 16,126, is found only in Vlach speaking Roma. The Monteni, Lom and Kalderash have different histories, but are all descended from Roma who were enslaved in Wallachia and Moldavia. This shared mtDNA lineage could possibly be a signature of this legacy. Similarly, the Spanish and Lithuanian Roma share the haplogroup U3 sequence with a mutation at position 16343. Haplogroup U3 represents over 50% of mtDNA in these populations, and is almost completely absent in all other populations. The sharing of this sequence at high frequency by the Spanish and Lithuanian Roma suggests that they have a recent common origin. Furthermore, the Lithuanian and Spanish Roma display the highest frequencies of the Y chromosome haplogroup VI-56, which is absent or rare in most other Romani populations. It is interesting to observe that VI-56 and U3 are most frequent in Middle Eastern populations (Richards et al., 2000; Underhill et al., 2000). Thus, these lineages provide evidence of possible admixture prior to the Roma entering Europe. Their over-representation in Romani populations that migrated westward suggests that groups splintering from the early migrant population could have included a larger proportion of admixed individuals of Middle Eastern origins.

Genetic distance analysis using mtDNA HVS1 data shows a sharp distinction between Spanish and Lithuanian Roma on one hand, and other populations. Within the

Balkan and Vlach groups, genetic distances between the Intreni and Monteni and the two Balkan Romani populations are large. However, genetic distances are considerably smaller between the two other Vlach populations, the Kalderash and Lom, and the two Balkan Romani populations. This is displayed in the neighbour-joining tree in which the Intreni and Monteni form a separate branch to the other Vlach and Balkan Romani populations. Thus, it is apparent that female genetic structuring does not conform to population groupings by historical migrations. Whilst there is a clear delineation between the Spanish and Lithuanian Roma and all other groups, substructuring within populations resident in Bulgaria appears to be complex. The most statistically robust AMOVA results are obtained when populations are grouped according to metagroups. This suggests different female histories for Romani populations resident in Bulgaria, and possibly reflects varying levels of female admixture in the different populations.

When mtDNA data are used to construct a population tree comparing the Romani populations to world-wide populations, the contrast is illuminating. Whereas regional autochthonous populations cluster closely together, large branch distances separate the Romani populations. Endogamous practices and small effective population sizes enhance the effects of genetic drift resulting in rapid population differentiation. In addition to this feature of the tree, the Roma are situated midway between European and Asian populations which reflects their genetic heritage.

5.3 Genetic Variation within Romani Populations

5.3.1 Intrapopulation Diversity of Paternal Lineages

The analysis of internal male-specific genetic diversity reveals widely varying degrees of genetic homogeneity within Romani populations. Populations can be grouped into those displaying considerable genetic heterogeneity, the Balkan and Western European Roma, and those showing extremely limited diversity, the Vlach Roma. Analysis of the male lineages in the Balkan and Western European Roma indicates that diversity has arisen through greater admixture with autochthonous Europeans. The homogeneity of Vlach male lineages points to strict adherence to male endogamy with very low levels of male-mediated gene flow. Until recently many Vlach

populations have been nomadic (Marushiakova & Popov, 1997) which could preserve traditional practices including endogamy. Furthermore, the enslavement of these populations in Moldavia and Wallachia may have served to restrict external genetic contributions.

5.3.2 Intrapopulation Diversity of Female Lineages.

Intrapopulation analysis using mtDNA data yielded contrasting results to those observed for male lineages. Mitochondrial DNA data indicate that the Lithuanian and Spanish Roma are by far the most restricted groups with diversity indices that are much lower than have been reported for other European populations. Whilst the two Balkan Romani populations show the greatest diversity, values for Vlach populations are of similar magnitude. Therefore, these data indicate a stricter adherence to female-specific endogamy in the Spanish and Lithuanian Roma than is observed in Balkan and Vlach groups.

Comparison of mtDNA diversity within the Roma to that of other populations points to their genetic heterogeneity. This is consistent with the Roma being composed of genetically differentiated population isolates. Diverse mtDNA due to admixture results in high pairwise differences between sequences (Arnason, Sigurgislason, & Benedikz, 2000), and this is the probable explanation for the high value determined for the Roma.

5.4 Summary of Findings

The investigation of maternal and paternal lineages in the Roma has identified predominant founding lineages of Indian origins. This supports claims of an Indian origin of different Romani populations. The homogeneity of these lineages suggests that the Roma are predominantly descended from a single ethnic population in India. Additional possible founding female lineages suggest that there may have been greater female diversity amongst the founder population. These findings disprove claims that the Roma comprise an indigenous European population (Okely, 1983). Furthermore, they contradict claims that the Roma were comprised of a conglomerate of different ethnic groups (Hancock, 1999b; Marushiakova & Popov, 1997). The data support a

scenario in which a limited number of related emigrants left India and made their way to Europe as a cohesive group. This implies that the population would have had a common reason and purpose for exiting India. However, these data cannot confirm nor disprove Hancock's (1999b, 2000) claim that the proto-Roma comprised a military force. Linguistic evidence points to extended sojourns in Persia and the Middle East (Fraser, 1992) and the genetic evidence collated in the present study provides evidence of possible Middle Eastern contributions to the Roma. Within Europe, the Roma have fractured into numerous diverse groups. The resultant social and cultural diversity is reflected in genetic diversity. Therefore, whilst Romani groups are related through a common ancestral population, they have become genetically differentiated through the stochastic process of genetic drift and differing degrees and sources of admixture. The establishment of each new population has represented a restrictive population bottleneck. The number of founders in each newly formed population would also impact on the current genetic profile of the population. Population bottlenecks combined with continued adherence to endogamous practices have resulted in populations with restricted genetic diversities. Thus, the Roma are best described as a mosaic of genetically related population isolates.

Genetic structuring in the male component of the population is related to the major migrations of the Roma into the Balkans, to Western Europe and out of Wallachia and Moldavia. However, this structuring does not appear to be the case for females. Social practices such as endogamy have likely shaped this structuring, and it is possible that some Romani populations may be more relaxed in letting unrelated females into their communities than males resulting in a less apparent female-specific genetic structure.

Section II

POSITIONAL CLONING OF THE HMSNL GENE

CHAPTER 6

SUBJECTS AND METHODS

6.1 Study Design and Subjects

6.1.1 Summary of Previous Findings

In the study by Kalaydjieva et al., (1996), a genome scan for segment sharing in an extended pedigree was used to localise the HMSNL gene to chromosome 8q24. All available polymorphic markers in the region were analysed for linkage in families from three Romani populations and recombination mapping reduced the critical interval to 3cM on 8q24.3. The disease locus was defined by a conserved haplotype constructed from four polymorphic loci: D8S558-D8S378-D8S529-D8S256. The two internal markers were homozygous in all patients, and the two markers bracketing the haplotype showed evidence of recombinations.

Refined genetic mapping was embarked upon by researchers at the Centre for Human Genetics, Edith Cowan University and the Medical School of the University of Sofia, Bulgaria. This was achieved by the identification of the publicly available marker AFM116yh8 and four novel markers; SLAP$(CA)_n$, pJ19, pJ10 and 474(CA1). Subsequent to the initial description of HMSNL, a number of additional Romani patients were identified in Bulgaria and throughout Europe. Genotyping of chromosome 8q24 markers confirmed the identical disease haplotype in these affected individuals thus expanding the sample size of the study. Two positional candidate genes, sialyl transferase 4A (*SIAT4A*) and src-like adaptor (*SLA*), were identified. *SIAT4A* was excluded by a recombination. *SLA* was sequenced in affected individuals and found not to contain any disease-causing mutations.

6.1.2 Research Questions

Kalaydjieva et al., (1996) noted the homogeneity of disease haplotypes in different Romani populations and postulated that HMSNL was caused by a founder mutation on 8q24.3. Genotyping of additional markers in the HMSNL region confirmed allelic homogeneity. Therefore, the search for the HMSNL gene proceeded under the hypothesis of a single founder mutation. This study aimed at positionally cloning the HMSNL gene and identifying the genetic defect in HMSNL affected individuals. To this end a number of research aims were developed as follows:

1. To assemble a map of contiguous genomic clones in the HMSNL region.
2. To determine:
    a. which known ESTs, STSs and polymorphic loci are found in the HMSNL region.
    b. which genes are contained within the HMSNL critical region and what their genomic structure is as determined by complete sequencing of the genomic region.
3. To perform fine-scale genetic mapping, through the use of newly identified polymorphic STRs and the identification of recent and historical recombinations in divergent Romani groups to reduce the candidate region for the HMSNL gene.
4. To identify the disease gene and primary genetic mutation that results in the HMSNL phenotype
5. To determine, for the purposes of future research, how the common origin and subsequent divergence of Romani groups impacts on the approach to refined genetic mapping.

6.1.3 HMSNL Affected Individuals and Families Involved in the Study

Refined genetic mapping was performed in 60 HMSNL affected individuals and 114 unaffected family members from 23 families. All patients were diagnosed as having HMSNL, based on the clinical presentation described in Kalaydjieva et al., (1996, 1998). Bulgarian Roma came from one of three groups: the Lom, Kalderash and Monteni. In addition, individuals and families were recruited through international collaboration. These included families from Spain (Colomer et al., 2000), Slovenia (Butinar et al.,

1999), Italy (Merlini et al., 1998), Germany (Baethmann, Gohlich-Ratmann, Schroder, Kalaydjieva, & Voit, 1998), France and Rumania (Kalaydjieva et al., 2000). All families were of declared Romani ethnicity except for the family with an affected child identified in Germany. The non-consanguineous parents of this individual were of Bulgarian nationality, but not of declared Romani ancestry (Baethmann et al., 1998). Haplotype analysis confirmed that this individual was homozygous for the same disease allele found in the Roma.

6.2 Methods

6.2.1 DNA Sample Preparation

DNA samples were extracted from blood samples collected on 3MM Whatman filter paper (Whatman) as described in section 3.2.1. Eluted DNA was quantified using spectrophotometry as described in section 3.2.2. Stock DNA was diluted to a working concentration of 10ng/μL with dH_2O.

6.2.2 Physical Characterisation of the HMSNL Region

Physical characterisation of the HMSNL region entailed the construction of a map of contiguous genomic clones and STS content mapping.

6.2.2.1 BAC library screening and BAC DNA isolation

For the purposes of constructing a map of contiguous genomic clones spanning the HMSNL region, the CITB Human Bacterial Artificial Chromosome DNA Pools Release IV library (ResGen) was screened. Library screening involved three rounds using a PCR assay to probe for genomic clones. In the first round the superpools were screened. The results of this initial phase dictated which plate pool and row/column pools were screened in subsequent rounds. Probing of the genomic library in this manner led to a unique address, corresponding to a unique BAC clone. BAC clones were ordered from the commercial supplier (ResGen). The bacteria were grown for 12-14 hours at 37°C in 100ml of LB broth containing 25μg/mL of the antibiotic kanamycin.

BAC DNA was isolated using the QIAfilter Plasmid Midi kit (Qiagen). For the purposes of subsequent experiments, the stock BAC DNA was diluted 1000-fold.

6.2.2.2 Chromosome walking

Coverage of the HMSNL region with genomic clones was achieved using chromosome walking. For this purpose, the ends of BAC genomic inserts were sequenced. These newly generated STSs were used as probes for subsequent rounds of library screening. The genomic insert ends of the BAC clone were sequenced as follows:

Four mg of purified BAC DNA was sequenced in two separate reactions using T7 and SP6 universal primers. These primers are complementary to the vector sequence at either ends of the genomic insert. Sequencing was performed using Dye Terminator Cycle Sequencing Ready Reaction Kits (Applied Biosystems). The thermocycling program for sequencing reactions was 30 cycles of 1 min at 94°C, 1 min at 60°C and 4 mins at 72°C. Sequencing products were precipitated and purified as described in section 3.4.3.3. Sequenced samples were prepared and run on the 373A DNA Analyser (Applied Biosystems) as described in section 3.4.3.4 and edited using Sequence Navigator software 1.0.1 (Applied Biosystems).

PCR primers were designed from the two BAC insert end sequences. A PCR assay was used to map the end of the new BAC in the parental clone. The unmapped BAC insert end was used for the subsequent round of BAC library screening. Table 7-1 provides a list of BAC insert end primers and PCR protocols generated during the study.

6.2.2.3 STS content mapping

All STSs including anonymous STSs, ESTs, microsatellite loci, and genes were screened in the contiguous map of genomic clones using standard PCR assays (table 6-1). PCR reactions were performed in 2400 and 9600 GeneAmp PCR Systems (Applied Biosystems) using protocols optimised for $MgCl_2$ concentration and annealing temperatures. Details of the PCR primers and reaction conditions are provided in chapter 7: tables 7-1 (STSs generated from BAC ends) and 7-2 (microsatellite loci).

Table 6-1
Standard PCR mixture for STS mapping

Reagent	Volume
10x PCR Buffer	1µL
2.5mM dNTPs	1µL
$MgCl_2$	0.4/0.6/0.8/1.0µL
Primer A	1µL
Primer B	1µL
*Taq*1 DNA polymerase	0.05µL
dH_2O	4.55/4.35/4.15/3.95µL
BAC DNA	1µL

6.2.3 Refined Genetic Mapping of the HMSNL Locus

6.2.3.1 Identification of novel microsatellite DNA in the HMSNL critical region

Identification of novel microsatellite DNA was achieved by screening genomic clones for simple repetitive DNA. BAC DNA was subcloned to create a sublibrary. The sublibrary was probed with labelled repetitive oligonucleotides and positive clones were sequenced.

6.2.3.1.1 Subcloning of BAC DNA

BAC DNA was randomly digested using the restriction endonuclease, *Sau*3A1 (New England Biolabs Inc.), at 37°C for 1 hour in a 1.5 mL Eppendorf tube (table 6-2). *Sau*3A1 digests genomic DNA to fragment sizes averaging 300bp.

Table 6-2
Sau3A1 restriction digest reaction mixture

Reagent	Volume
BAC DNA	2.0µL
10x reaction buffer	1.5µL
DTT (5mM)	1.5µL
BSA (1mg/mL)	1.5µL
*Sau*3A1	1.0µL
dH_2O	7.5µL

Digest products were purified using a silica/PBS/3M NaI "glass milk" matrix as follows:

To the digest mixture was added 8μL of glass milk and 60μL of NaI. The mixture was vortexed and left at room temperature for 10 mins with intermittent agitation. The mixture was centrifuged at 13,000 rpm for 1 min and the supernatant discarded. The pellet was resuspended in 300μL of New Wash (Invitrogen) by vortexing, and centrifuged for 1 minute and the supernatant discarded. This washing process was repeated an additional two times. All traces of New Wash were removed following the final wash. To elute the DNA, the pellet was resuspended in 8μL of dH_2O and incubated at 55°C for 5 mins. The mixture was centrifuged at 13,000 rpm for 2 mins and the eluate transferred to a clean 1.5mL Eppendorf tube.

The phagemid, pBluescript II (KS) [Stratagene] was digested with the restriction endonuclease, *Bam*H1 (New England Biolabs Inc.) at 37°C for 1 hour to create a complementary cloning site (table 6-3). The vector, pBluescript II (KS) contains the amp^r gene, which confers resistance to ampicillin, allowing the use of this antibiotic as a selection agent. It also contains the *lac*Z component of *β-galactosidase*, which is interrupted by successful cloning, resulting in the absence of *β-galactosidase* activity.

Table 6-3
BamH1 restriction digest reaction mixture

Reagent	Volume
pBluescript DNA	2.0μL
10x reaction buffer	1.5μL
*Bam*H1	1.0μL
dH_2O	10.5μL

Ligation of BAC DNA fragments into pBluescript was performed using the Rapid DNA Ligation Kit (Roche). T4 DNA ligase catalyses the formation of phosphodiester bonds between neighbouring 3'-hydroxyl and 5'-phosphate ends of double stranded DNA. The ligation reaction mixture was prepared (table 6-4) and incubated at room temperature for four hours.

Table 6-4
Ligation reaction mixture

Reagent	Volume
*Sau*3A1 digested BAC DNA	4μL
BamH1 digested pBluescript	2μL
5x DNA dilution buffer	2μL
dH2O	2μL
2x T4 DNA ligation buffer	10μL
T4 DNA Ligase	1μL

Library Efficiency DH5α Competent Cells (Gibco BRL) were transformed with the ligated products. For this procedure, 5μL of ligated products were added to 500μL of DH5α cells and the mixture was gently agitated. The bacteria were heat shocked at 42°C for 45 s and then rescued with the addition of 100μL of LB broth followed by incubation at 37°C for 1 hour. The entire volume of bacteria was then plated on to MacConkey agar (Gibco BRL) plates containing 1mg/mL of ampicillin as a selecting agent. Transformed bacteria were cultured overnight at 37°C.

Positively transformed bacteria were identifiable as white in colour, as these bacteria were unable to ferment lactose due to the inactivation of *β-galactosidase* activity. Bacteria that did were not successfully transformed were red. Approximately 150 individual positive clones were picked and replica-plated on to a gridded Hybond nitrocellulose membrane (Amersham) and an agar plate. Bacteria on the two replica plates were grown overnight at 37°C.

After incubation, the agar plate containing gridded subclones was stored at 4°C. The gridded nitrocellulose membrane was denatured with a 1.5M NaCl/0.5M NaOH solution and neutralised with a 1.5M NaCl/0.5M Tris solution. The membrane was washed with 2x SSC and the DNA fixed to the membrane by heating at 80°C for 2 hours.

6.2.3.1.2 Probing gridded membranes for repetitive DNA

Nitrocellulose membranes containing the gridded library were screened with di-, tri-, and tetra-nucleotide repeat probes labelled with [^{32}P]γ-ATP. To label repeat oligonucleotides, a reaction mixture containing T4 polynucleotide kinase (PNK)

[Promega] was prepared (table 6-5) and allowed to react at 37°C for 20 minutes. The reaction was stopped by the addition of 2μL of 0.5M EDTA. To remove unincorporated [^{32}P]γ-ATP, the reagents were run through a S-300 MicroSpin Column (Pharmacia Biotech).

Table 6-5
Reaction mixture for labelling repeat oligonucleotides with [^{32}P]γ-ATP

Reagent	Volume
Buffer	2μL
100mM DTT	2μL
Oligonucleotides (10pmol/μL)	4μL
^{32}Pγ-ATP	5μL
dH_2O	5μL
T4 PNK	2μL

Labelled repeat oligonucleotide probes were added to 20mL of 12x SSC and the mixture was added to the library membranes for overnight hybridisation at 45°C. The following day the membranes were washed with 12x SSC at 50°C and then with progressively decreasing concentrations of SSC at increasing temperatures. This process was completed with a final wash of 1x SSC at 60°C. The membrane was then exposed to Cronex 4 autoradiographic film (Kodak Eastman) for four hours at –80°C. The film was developed in a Curix 60 X-ray film developer (AGFA). Positive clones were identified by an increased radioactive signal, compared to that caused by background hybridisation.

6.2.3.1.3 Sequencing of positive subclones

The corresponding positive clones on the agar replica plate were picked and grown overnight at 37°C in 2 mL of LB containing 1mg/mL of ampicillin. Plasmid DNA was extracted using the QIAQuick Miniprep kit (Qiagen) as per the manufacturer's instructions. Plasmid inserts were sequenced in both directions using the universal PCR primers, T3 and T7, with Dye Terminator Ready Reaction Cycle Sequencing kits (Applied Biosystems). Thermocycling sequencing reactions entailed 30 cycles of 30 s at 94°C, 30 s at 50°C and 1 min at 72°C. Sequencing products were precipitated and purified as described in section 3.4.3.3. Samples were prepared and electrophoresed on

the 373A DNA Analyser (Applied Biosystems) as described in section 3.4.3.4, and edited using Sequence Navigator software (Applied Biosystems). Repetitive sequences were identified and PCR primers designed for the amplification of microsatellites.

6.2.3.2 Genotype analysis of microsatellites in the HMSNL region

6.2.3.2.1 PCR amplification of microsatellites with inclusion of [^{32}P]α-CTPs

A standard PCR mixture was used to incorporate [^{32}P]α-CTP into the amplified DNA fragments (table 6-6). Table 7-2 provides a list of the PCR primers and protocols for the newly identified microsatellites. Microsatellite loci were amplified in affected individuals and family members.

Table 6-6
Standard PCR mixture for incorporation of [^{32}P] α-CTP into amplified microsatellite fragments

Reagent	Volume
Buffer	1μL
dNTPs (2μM)	1μL
$MgCl_2$	1μL
Primer A	1μL
Primer B	1μL
*Taq*1 polymerase	0.05μL
dH_20	3.75μL
DNA	1μL
^{32}Pα-dCTP	0.2μL

6.2.3.2.2 Analysis of microsatellite alleles

[^{32}P]α-CTP labelled PCR products were electrophoresed on 6% polyacrylamide gels using a Pokerface II apparatus (Hoeffer). Gels were prepared using 100mL of 6% acrylamide made from a stock solution of 40% 19:1 bisacrylamide/acrylamide. Polymerisation occurred through the addition of 50μL of TEMED and 500μL of 10% ammonium persulphate. Wells were formed using a plastic shark toothcomb. To prepare samples, 2μl of formamide loading buffer (98% formamide, 10mM NaOH, 0.1% bromophenol blue and 0.1% xylene cyanol) was added to labelled PCR products, the DNA fragment denatured by heating at 95°C for 5 minutes, and then placed on ice. A 2μL aliquot of the sample was loaded on to the gel and electrophoresed for 2.5 hours

at 1400V. The gels were fixed using a 10% methanol/10% acetic acid solution and dried in a Savant gel dryer. Dried gels were exposed to Cronex 4 film (Eastman Kodak) for 12 hours at –80°C and developed using a Curix 60 X-ray film developer (AGFA).

Allele calling was performed manually, assigning the number 1 to the largest observed allele. Control samples were used for each microsatellite on each gel to ensure compatibility of allele calling between gels.

6.2.3.3 Haplotype analysis and fine-structure mapping of the HMSNL locus

The physical order of microsatellite markers was determined through STS content mapping. Haplotypes were constructed manually using genotypic information from affected offspring and both parents (where available) to resolve the phase of alleles. Haplotypes were examined for possible parental and historical recombinations.

6.2.4 Genomic Sequence Analysis of the HMSNL Region

A minimal tiling path of BAC clones spanning the HMSNL region was selected. These clones were forwarded to the Institute of Molecular Biotechnology in Jena, Germany where large-scale genomic sequencing was performed as part of the Human Genome Project. Genomic sequence information was submitted by Dr Karen Blechschmidt to the public database and is accessible through the accession numbers indicated in table 7-5.

6.2.5 Candidate Gene Analysis

The genomic structures of two positional candidate genes were determined through the comparison of genome sequence with the published cDNA sequence of Wnt1-inducible signalling protein 1 [*WISP1*] (Pennica et al., 1998) and N-myc down-regulated gene 1 [*NDRG1*] (Kokame, Kato, & Miyata, 1996). This was performed using the homology search functions of BLAST v1.4 (Altschul et al., 1997) housed at the National Centre for Biotechnology Information (http://www.ncbi.nlm.nih.gov/BLAST/). Genomic structures of *WISP1* and *NDRG1* can be accessed through the Locus Link (http://www.ncbi.nlm.nih.gov/LocusLink/index.html) IDs provided in table 7-6.

6.2.5.1 Sequence analysis of *WISP1*

The five exons of *WISP1* were amplified in separate reactions (table 6-7) using standard 10μL PCR (table 6-1). All coding regions and at least 50bp of intronic sequence on each side of the exon were analysed for sequence variants.

Table 6-7
PCR primers and protocols for WISP1

Primer sequence	Protocol	Size of PCR fragment
Exon 1-F CAT ATC TGG TGC TCC TGA TGG -R GTA GCA GGA CCC AGT AGAGAA G	63-55°C (Δ-0.5°C/cycle) 20 cycles @ 55°C 2.0mM $MgCl_2$	288bp
Exon 2-F GAC AGG AAT GCA ATG GCA G -R GGT GTA TCT CCT GCT GAA C	63-55°C (Δ-0.5°C/cycle) 20 cycles @ 55°C 1.0mM $MgCl_2$	488bp
Exon 3-F GCA TGG TCC ACA TGG AGC C -R GGT GGT CAG AGT TCC AGG	35 cycles @ 55°C 1.0mM $MgCl_2$	424bp
Exon 4-F GTG TGG TGA AAG TGA GGG TTG -R GCT TGT GAA GTC TAG ACA TCC	35 cycles @ 55°C 1.5mM $MgCl_2$	304bp
Exon 5-F GTA AGG TGG AAT GCT CCC AC -R CAG ATC AGG GTA ACT AAG GC	63-55°C (Δ-0.5°C/cycle) 20 cycles @ 55°C 2.5mM $MgCl_2$	516bp

PCR fragments were cleaned with the QIAQuick PCR Purification Kit (Qiagen). Sequencing of PCR products was performed with the same PCR primers using the Dye Terminator Ready Reaction Cycle Sequencing kit (Applied Biosystems). Thermocycling sequencing reactions entailed 30 cycles of 30 s at 94°C, 30 s at 50°C and 1 min at 72°C. Sequencing products were precipitated and purified as described in section 3.4.3.3. Samples were prepared and electrophoresed on the 373A DNA Analyser (Applied Biosystems) as described in section 3.4.3.4 and edited using Sequence Navigator 1.0.1 software (Applied Biosystems).

6.2.5.2 Sequence analysis of *NDRG1*

The 16 exons of *NDRG1* were amplified in separate reactions (table 6-8) using standard 10μL PCR (table 6-1). All coding regions and at least 50bp of intronic sequence on each side of the exon were included in the search for sequence variants.

Table 6-8
PCR primers and protocols for NDRG1

PCR Primer	Protocol	Size of PCR fragment
Exon 1-F GAC TGC GAG GGT GTG GGA G -R CTT ACT CCT GGA GTA CGC	63-58°C (Δ-0.5°C/cycle) 20 cycles @ 58°C 1mM $MgCl_2$	313bp
Exon 2-F CTT CTT GCC ATT GGT CTT G -R GCA TGC CCA TAA GTA CAA G	35 cycles @ 55°C 1.5mM $MgCl_2$	282bp
Exon 3-F GAT TCA GGT CAT AGA AAG G -R AGA GAA GAC GGG ATG AGG	35 cycles @ 55°C 1mM $MgCl_2$	172bp
Exon 4-F CAC GCG GAT GCC ATG AAC -R GCA TTT CTG GCT TTT CCA G	63-58°C (Δ-0.5°C/cycle) 20 cycles @ 58°C 3mM $MgCl_2$	331bp
Exon 5-F CTT TGA CAC CGA GAC ACC -R GAG CAA AGC ACC TGA ACC	63-58°C (Δ-0.5°C/cycle) 20 cycles @ 58°C 1mM $MgCl_2$	268bp
Exon 6-F CTA ATG GCT TCT CTG TGT C -R GTC AGT CCA GAT CAA AGC	63-58°C (Δ-0.5°C/cycle) 20 cycles @ 58°C 1mM $MgCl_2$	178bp
Exon 7-F AGG CTC CCG TCA CTC TG -R GTC TTC CTT CAT CTT AAA ATG	35 cycles @ 55°C 2mM $MgCl_2$	176bp
Exon 8-F CCT AGT GTT TCA GAT TGC TG -R GAG AGC TCG TAG TCT CAG	63-58°C (Δ-0.5°C/cycle) 20 cycles @ 58°C 1mM $MgCl_2$	238bp
Exon 9-F GGA GTC CAG CAA TGC CAC -R CTG AGC ACC ACA CAA TGC	63-58°C (Δ-0.5°C/cycle) 20 cycles @ 58°C 2mM $MgCl_2$	224bp
Exon10-F GAG TAG TGA CCA GCT CAG -R CAA ACT CAG AGC CTG CCT C	35 cycles @ 55°C 1mM $MgCl_2$	287bp
Exon11-F ACA GGG CCT CTC TCA AGT TG -R CTG GGT AAT GCT CAG TCT C	35 cycles @ 55°C 1mM $MgCl_2$	346bp
Exon12-F CAG GCC TGG GAG TGG GAC AAT C -R GCA GGC AGG GCC ACT TCA AC	35 cycles @ 55°C 2mM $MgCl_2$	201bp
Exon13-F CAA GCC ACA TCT GCT GAA TCC -R CTT TGC AGC CTC AGA TCA CC	35 cycles @ 55°C 1mM $MgCl_2$	390bp
Exon14-F GAC ACC AGC AGC CTT GCC TG -R CCT AGG GAA TCA GAG TCC TC	35 cycles @ 55°C 1mM $MgCl_2$	389bp
Exon15-F GGA AAC TGG CTC AGA CAG G -R CAT GCC CTC CAC ACA CCT AAC	63-58°C (Δ-0.5°C/cycle) 20 cycles @ 58°C 2mM $MgCl_2$	432bp
Exon16-F GTG GAC ATG GAG AGG ACG -R GTC TCC ACC AGA GCT CAC TC	63-58°C (Δ-0.5°C/cycle) 20 cycles @ 58°C 1mM $MgCl_2$	576bp

PCR fragments were cleaned with the QIAQuick PCR Purification Kit (Qiagen). Sequencing of PCR products was performed with the same PCR primers using the Dye Terminator Ready Reaction Cycle Sequencing kits (Applied Biosystems). Thermocycling reactions entailed 30 cycles of 30 s at 94°C, 30 s at 50°C and 1 min at 72°C. Sequencing products were precipitated and purified as described in section 3.4.3.3. Samples were prepared and electrophoresed on the 373A DNA Analyser (Applied Biosystems) as described in section 3.4.3.4 and edited using Sequence Navigator 1.0.1 software (Applied Biosystems).

6.2.6 Analysis of the R148X Mutation Using *Taq*1 Restriction Endonuclease

For the purposes of the restriction digest assay, exon 7 of *NDRG1* was amplified using the primers and protocol described in table 6-8. A restriction digest reaction was performed on PCR products using the restriction endonuclease, *Taq*1 (New England Biolabs) for 4 hrs at 65°C (table 6-9).

Table 6-9
Taq1 digest of exon 7 of NDRG1 for R148X mutation assay

Reagent	Volume
PCR product	7.0μL
10x Reaction Buffer	1.0μL
*Taq*1	0.1μL
BSA	0.1μL
dH_2O	6.8μL

Restriction products were electrophoresed for 30 mins at 80V on a 4% agarose gel (3:1 standard agarose: metaphor agarose [BioWhittaker Molecular Applications]) stained with ethidium bromide. Separated DNA fragments were visualised with UV light on a transilluminator. The undigested PCR product was 176bp long and the digest products were 104bp and 72bp long.

The PCR assay was redesigned to in order avoid a primer annealing site mutation. The new primers designed for amplifying exon 7 were F-AACTGTGGAGAATACGGG and R-CTGTGCAGGCAGTTACGGCAGC and yielded a PCR product of 316bp. A 10μL PCR (table 6-1) was most efficient with the use of HotStar*Taq* (Qiagen) requiring an initial denaturation and enzyme activation of 15 mins

at 96°C followed by 35 cycles of 30 s at 96°C followed by 30 s 55°C and extension at 72°C for 45 s. A $MgCl_2$ concentration of 1mM (i.e. 0.8µL of 25mM $MgCl_2$ in a 10µL reaction) was used. The *Taq*1 restriction endonuclease assay was performed as described above (table 6-9) which produced digest products of 190bp and 126bp.

CHAPTER 7

RESULTS

7.1 Physical Mapping of the HMSNL Region

7.1.1 A Map of Contiguous Genomic Clones Spanning the HMSNL Region

The initial round of BAC library screening was performed using microsatellite loci that defined the core disease haplotype (Kalaydjieva et al., 1996). This included D8S378, D8S529, D8S256 and AFM116yh8. Chromosome walking proceeded by screening the BAC library in a redundant manner ensuring dense coverage of the region. A total of thirty-two genomic clones were identified as mapping to the region. This included thirty BAC clones and two PAC clones. Screening of the BAC library failed to identify any clones covering the region from 326J4 to 458A3. The PAC clone, 709A24988Q2 (PAC 709), was found to extend beyond BAC 458A3. However, a gap remained between BAC 326J4 and PAC 709. A genomic clone that spans this gap was later identified through searches of genome sequence databases (218N23). Thus, a highly redundant map of contiguous clones, providing complete coverage of the HMSNL genomic region, was constructed (figure 7-1). The physical distance spanned by this contig was estimated to be 1Mb.

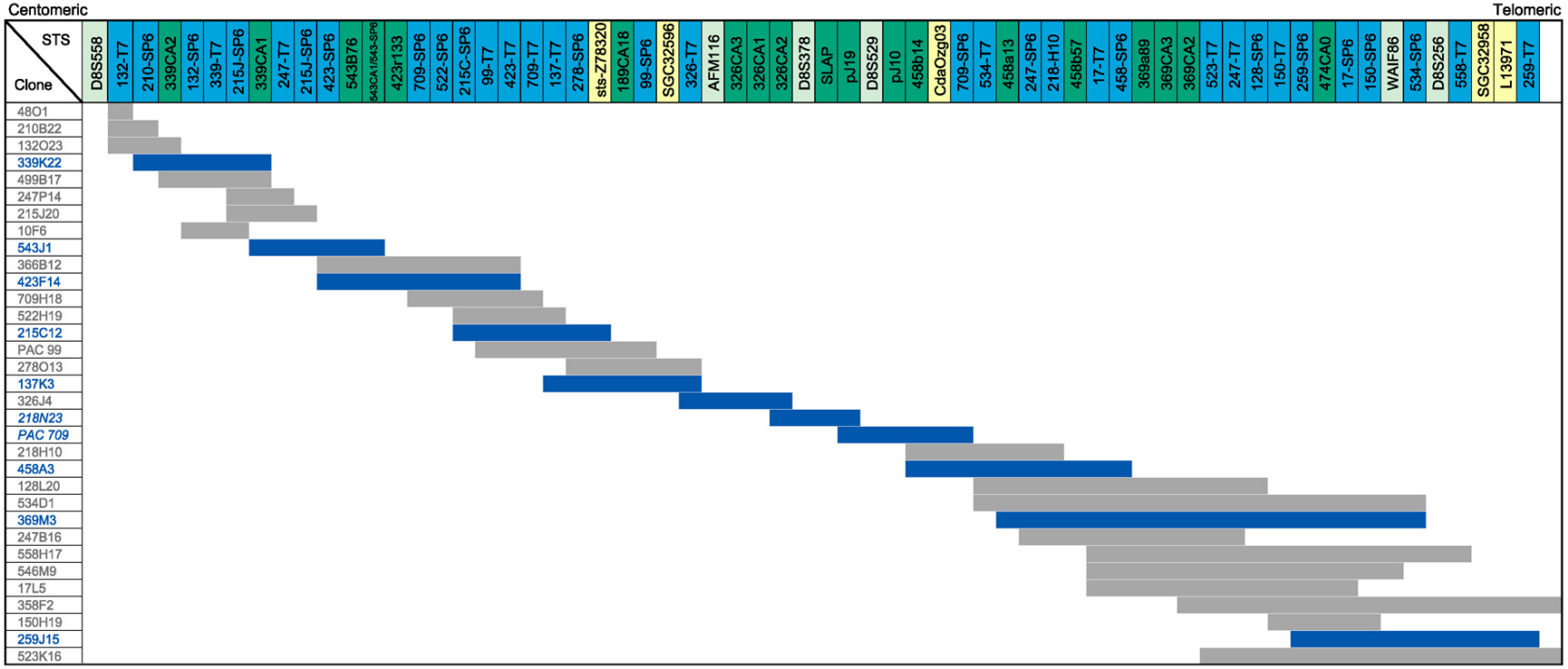

Figure 7-1 Map of contiguous genomic clones providing coverage of the HMSNL critical region. Blue clones represent a minimal tiling path, which was sequenced by the Centre for Molecular Biotechnology, Jena. Corresponding clone names are provided in the left column. STSs which were physically mapped to the contig are indicated. Publicly available microsatellite DNA are indicated in light green. Novel microsatellite DNA, identified in this study, are in dark green. Novel anonymous STSs derived from BAC insert ends are coloured aqua. ESTs mapped in the region are in yellow.

7.1.2 STS Content Mapping in the HMSNL Region

Anonymous STSs were derived from sequencing the ends of BAC inserts. Polymorphic markers mapped in the region included those available in public databases and novel microsatellites identified during the course of the study. Expression sequence tags (ESTs) and known genes that had been tentatively localised to chromosome 8q24 were also screened in the contig. All STSs were systematically screened against the contig to aid ordering of genomic clones relative to each other (figure 7-1). In addition, the mapping of known and newly identified polymorphic markers facilitated refined genetic mapping by determining marker order for haplotype analysis. The identification of genes and/or ESTs mapping to the HMSNL region identified positional candidate genes.

7.1.2.1 STSs localised in the map of contiguous genomic clones

Thirty-five novel anonymous STSs generated from the sequencing of BAC insert ends were used to position and order genomic clones. One of these, 543J1-SP6, contained a $(CA)_n$ repeat that was determined to be polymorphic and used in refined genetic mapping. PCR protocols were developed for each of these STSs (table 7-1).

Table 7-1

Primer sequences and protocols for novel STSs in the HMSNL contig map

STS	PCR primer sequence	PCR protocol	Product size
132O23-T7	F- ctt cca aat tcc atc ttg R- tat tgg aaa gtg gtc agg	35 cycles @ 55˚C 1.5mm $MgCl_2$	131bp
210B22-SP6	F- atg ctt tga gca ctg tgg R- cgt agt tcc cca tac aag	35 cycles @ 55˚C 1.5mm $MgCl_2$	145bp
132O23-SP6	F- aag cca cca agg ctc ag R- aag ctg gtg gac ttg gtg	35 cycles @ 55˚C 1.5mm $MgCl_2$	119bp
339K22-T7	F- atg agc ttc tgg gtg tac R- tag aga gtc gaa cac cag	35 cycles @ 55˚C 1.5mm $MgCl_2$	95bp
215J20-SP6	F- tct cat gta acg tcc ttg R- cca gac ttg aaa ttg acc	35 @ 50˚C 1.5mM $MgCl_2$	183bp
247P14-T7	F- tta tac tca tat tct gta tg R- aga ggc atg agc cac ag	35 @ 50 1.5mM $MgCl_2$	204bp
423F14-SP6	F- gct caa gct cca cat ggc ac R- gtg atc atc ccc tgt tcc tcc	35 @ 50 1.5mM $MgCl_2$	128bp
543J1-SP6	F- gtc tta ctg ctg tat ctc c R- cca caa tac gaa tgt atg	35 @ 55˚C 1.5mM $MgCl_2$	121bp
709H18-SP6	F- gtc cca gcc tct atc tcc tg R- gag gtg gaa ttt ccc ata gc	35 @ 55˚C 1.5mM $MgCl_2$	130bp
522H19-SP6	F- cct tga tga tgc cag gtg ac R- gag gat taa aca gga gga tgc	35 cycles @ 55˚C 1mM $MgCl_2$	168bp
215C12-SP6	F- gga tca cag tct agt ccc agg R- gct gtg ggg gag aca gct g	35 cycles @ 55˚C 1.5mM $MgCl_2$	235bp
PAC99-T7	F- gca gca caa gca gat cat ttt gc R- cac ccc ttc ccc aac acc tc	35 cycles @ 55˚C 1mM $MgCl_2$	90bp
423F14-T7	F- cac aac atc cac tga ttc tc R- gga tgt gca gga tat taa gg	35 @ 55 1mM $MgCl_2$	155bp
709H18-T7	F- cta tgc aag aca atg ggt R- gaa aac tga ata taa ttt gg	35 @ 55 1mM $MgCl_2$	126bp
137K3-T7	F- gct gat aca aaa tat acg ttg tg R- gga tgt act atg aga atc caa cg	35 cycles @ 55˚C 1.5mM $MgCl_2$	138bp
278O13-SP6	F- gat cga tgc tgg tcg acg aat cc R- tat act gca tgg att att gtt gcg	35 cycles @ 55˚C 2.5mM $MgCl_2$	158bp
PAC99-SP6	F- gag ctg act aac act att agg R- ctc aac agc att tgg tat gaa cag	35 @ 55˚C 1mM $MgCl_2$	155bp
326J4-T7	F- gca tgc att tta ggg caa tgg R- ctc att aca tag gcc aag ttc ac	35 @ 58˚C 1.5mM $MgCl_2$	280bp
PAC709-SP6	F- caa gtc aca ccg tcg ac R- gtt tgc cca gag ctg ag	35 cycles @ 55˚C 1mM $MgCl_2$	171bp
534D1-T7	F- ctt cta atc ttg caa ttt cc R- cac ata gaa gta gaa tgc	35 cycles @ 55˚C 1mM $MgCl_2$	138bp
247B16-SP6	F- cat gcc cat aca gat cac R- gca cag ctg atg act gg	35 @ 58˚C 1.5mM $MgCl_2$	133bp
218H10-SP6	F- ggc ctt gga tga agt cc R- gat cca gac aag agg atg	35 @ 55˚C 1mM $MgCl_2$	256bp
17L5-T7	F- ggt ctc ccc ggc tcc acc R- ctg aaa ttt aat cca ggt cca g	35 @ 55˚C 1mM $MgCl_2$	145bp
458A3-SP6	F- cca caa atg tct ttc cac R- agg atg cag aca cag agg	35 @ 50˚C 1.5mM $MgCl_2$	121bp
523K16-T7	F- ctc aca gtt tta ttt cct gga c R- gtc gat gac agc cat aac tca c	35 @ 58˚C 1.5mM $MgCl_2$	183bp
247B16-T7	F- cta tga tct cct tac ata tg R- caa gca ctt gca tat tta g	35 @ 55˚C 1.5mM $MgCl_2$	114bp
128L20-SP6	F- ctt ctg tat atc ctt agg c R- gat aat tgg cca cca tta c	35 @ 55˚C 1.5mM $MgCl_2$	155bp
150H19-T7	F- gca ttg gag gtg gca ggc ttt gc R- cag gcc agt gag ggt tgg tgt c	35 @ 60˚C 1.5mM $MgCl_2$	153bp
259J15-SP6	F- ggt tga ctg agg ttc aaa tgt R- acc aga ccc tag gca ggt gca	35 cycles @ 50˚C 2mM $MgCl_2$	198bp
17L5-SP6	F- cct gac ttg tat aaa cag R- cat gaa ctg aaa ccc cag c	35 @ 55˚C 1mM $MgCl_2$	148bp
150H19-SP6	F- gca tct cca ttg ctg ggc ctc R- ctg gga atg tgt aac ctg tac c	35 @ 60˚C 1.5mM $MgCl_2$	150bp
534D1-SP6	F- cat aca tgc agg tgc ttt g R- ccc ata ggc aac agc aac	35 @ 58˚C 1.5mM $MgCl_2$	121bp
558H17-T7	F- gaa tgg gtc taa acc ttc R- gcc aag tgc cca gca tg	35 @ 58˚C 1.5mM $MgCl_2$	224bp
259J15-T7	F- gaa tgc tgg gac ctg cct g R- gct gct agg gcc tgg act c	35 @ 55˚C 1.5mM $MgCl_2$	178bp

7.1.2.2 Polymorphic markers localised in the contig

Physical mapping placed four published microsatellite loci in the contig (AFM116yh8, D8S378, D8S529 and D8S256). D8S558 was determined to be centromeric of the contig and HMSNL critical region. In addition to the four microsatellite loci, a published SNP, WIAF86, mapped to the region. Nineteen novel polymorphic microsatellites were identified in the course of the HMSNL study. PCR protocols were developed for each genetic marker (table 7-2). All microsatellites were $CA_{(n)}$ repeats except for pJ19, which is a CTG repeat. Markers were physically mapped in the contig, thereby determining their relative order (figure 7-1).

Table 7-2
PCR primers, protocols, and approximate allele sizes of novel microsatellite DNA

Locus name	PCR primers	PCR protocol	Approximate allele size
339CA2	F- cgg acc caa atc aat ttt c R- cca ttt aca gtg cag atg	56˚C-50˚C (Δ-0.5˚C/cycle) 20 cycles @ 50˚C 1mM $MgCl_2$	122bp
339CA1	F- ttg atc tgg gag aat gat g R- aca tat aca ctg cca cg	35 cycles @ 55 ˚C 1mM $MgCl_2$	100bp
543b76	F- gtg gca gag tga gac act R- tat act atg acc att ctc tg	35 cycles @ 50˚C 1.5mM $MgCl_2$	120bp
543CA1	F- gtc tta ctg ctg tat ctc c R- cca caa tac gaa tct atg	35 cycles @ 50˚C 1.5mM $MgCl_2$	150bp
423r133	F- cat tac agg cat ctg cca tg R- gtc aac atg gcg aac gct g	35 cycles @ 55 ˚C 1mM $MgCl_2$	120bp
189CA17	F- gaa aag gtc aat atg cca gg R- gat tga gtt gtc tat ttg tc	35 cycles @ 55 ˚C 1mM $MgCl_2$	140bp
326CA3	F- tca tgg gat aaa aca tta gtg aa R- gat ttg caa ttt att caa gaa cac	35 cycles @ 55 ˚C 1mM $MgCl_2$	160bp
326CA1	F- gaa atg ctg gca gaa gtc ttg aaa g R- ttg act ccc tgc att tat acc aat ctt	35 cycles @ 50˚C 1.5mM $MgCl_2$	190bp
326CA2	F- gtg cac caa aat ctc aca aat cac R- cca att cac cgc aag tca gac act	35 cycles @ 50˚C 1.5mM $MgCl_2$	300bp
SLAP (CA)	F- tgc gtc aga aga ctg tgg ac R- tgg cca tgg ttt tca tgt gc	35 cycles @ 55 ˚C 1mM $MgCl_2$	170bp
pJ19	F- acc aca gcc cag tgc ctg att cc R-ttt act tgg cac cca ggc ttc tca	35 cycles @ 55 ˚C 1mM $MgCl_2$	140bp
pJ10	F- agg gtc tta gtc cca aca R- aga aag aac tga cca gcc	35 cycles @ 50˚C 1.5mM $MgCl_2$	170bp
458b14	F- ctc tcc ctc caa agt ctc c R-aaa gca gag gaa gcg ctg g	35 cycles @ 55 ˚C 1mM $MgCl_2$	170bp
458a13	F- aag tat ccc tgt tat tca gc R- ctt act tcc agg ata aac ac	56˚C-50˚C (Δ-0.5˚C/cycle) 1mM $MgCl_2$	110bp
458b57	F- aga cag tct tct tga ctg g R- tgt acc caa gtc cca tcc	35 cycles @ 55 ˚C 1mM $MgCl_2$	120bp
369a89	F- ctc atc tac aca ctc gcg cg R- ggc cga tga gac ggt cg aaa	35 cycles @ 50˚C 1.5mM $MgCl_2$	200bp
369CA3	F- gat ata att atg cag ata gg R- gtt att tgt ctt atc agt c	35 cycles @ 50˚C 1.5mM $MgCl_2$	188bp
369CA2	F- ctc cta cct gct gtc tgc R- gct gag aag tcc atg atc	35 cycles @ 55 ˚C 1mM $MgCl_2$	199bp
474CA1	F- tca ggc agg ctg gat tca g R- agc aga gcc atg gca cat g	35 cycles @ 55 ˚C 1mM $MgCl_2$	170bp

7.1.2.3 ESTs and known genes identified in the contig

ESTs and genes that had been assigned to the 8q24 region and were publicly available in the databases were considered as potential positional candidate genes. As only a gross chromosomal localisation had been determined for many of these genes, a rational approach was taken in which plausible functional candidates were preferentially selected for screening. This was based on knowledge of function for genes and the expression patterns for ESTs. A panel of 11 genes and 15 ESTs putatively mapped to the region were screened against all genomic clones. Five ESTs were found in the contig (table 7-3). No known genes were mapped to the contig. STS content mapping allowed the relative positioning of the ESTs in the contig (figure 7-1).

Table 7-3
ESTs identified in the HMSNL critical region

EST	Accession number[φ]	Unigene cluster[ξ] (# of ESTs)	Corresponding Gene
HSZ78320	Z78320	-	-
SGC32596	H87187	Hs.300598 (4 ESTs)	-
SGC32958	T32458	Hs.293696 (99 ESTs)	-
L13972	L13972	Hs.60617 (25 ESTs)	*SIAT4A*
Cdaozg03	Z39096	Hs.75789 (509 ESTs)	*NDRG1*

[φ]http://www.ncbi.nlm.nih.gov/Genbank/
[ξ] http://www.ncbi.nlm.nih.gov/LocusLink/

7.2 Genetic Mapping of the HMSNL Region

The aim of refined genetic mapping of the HMSNL locus was to utilise historical and parental recombinations to narrow the critical region to a genomic segment amenable to mutational analysis of positional candidate genes. Accordingly, microsatellites were genotyped in affected individuals and parents. The genotyping of parents allowed the phase of alleles to be resolved. Dense marker haplotypes were constructed using genotypic data from 24 microsatellite loci. This enabled the discernment of variant alleles due to microsatellite mutations from those resulting from recombinations. Recombination breakpoints were physically mapped using the results of STS content mapping.

7.2.2 Haplotype Analysis and Fine-structure Mapping of the HMSNL region

Disease haplotypes from 60 individuals (i.e. 120 disease haplotypes) were constructed (table 7-4). Twenty unique disease haplotypes were observed in affected individuals from nine populations. These haplotypes were non-randomly distributed amongst populations; with only three found in more than one population.

Within the disease haplotypes, microsatellite mutations were distinguished by the conservation of the common disease haplotype on both sides flanking the locus. A conservative approach was taken to designating the breakpoints of recombinant haplotypes. Microsatellite mutations were observed at four of the loci: 339CA2, 189CA17, D8S378 and 458b14. Within these, locus D8S378 displays the greatest mutability with four different alleles arising due to mutations observed on HMSNL haplotypes.

Table 7-4

HMSNL disease haplotypes constructed using 24 polymorphic microsatellite loc over a 3cM region. Affected individuals were from nine Romani groups. Microsatellite mutations are indicated in yellow. Recombinant haplotypes are red. The conserved minimum region of complete homozygosity is shaded in blue.

Thirteen unique recombinant haplotypes were observed. These are the result of ten recombination events, with subsequent mutations differentiating recombinant haplotypes. A single maternal recombination was observed in a Kalderash individual, the breakpoint of which mapped between markers pJ10 and 458b14 (haplotype L). The other 12 recombinant haplotypes were the result of historical recombinations. Seven recombination events were observed on the centromeric side of the haplotypes, with the breakpoints of five of these recombinations mapping to the region between marker pJ10 and 458b14 (haplotypes B, L, Q, S & T). Thus, five different recombinations define the centromeric boundary of the critical region. Each of these five recombinant haplotypes occurs in a different population. On the telomeric side, a minimum of three recombinational events is observed (haplotypes F and S show evidence of identical recombinations as do haplotypes G, H and O). Two historical recombinations extend to the region between markers D8S256 and 474CA1 (haplotype O) and markers 474CA1 and 369CA2 respectively (haplotype P). A single historical recombination maps to the region between markers 369a89 and 369CA3. Thus, the telomeric boundary of the HMSNL critical region is defined by a single historical recombination observed in individuals from the Kalderash (haplotype F) and Spanish Roma (haplotype S).

Fine-scale haplotype mapping of the HMSNL region reduced the region of homozygosity to one bracketed by pJ10 and 369CA3 and encompassing the loci 458b14, 458a13, 458b57 and 369a89. Within this haplotype, marker 458b14 displays significant heterogeneity with four alleles. However, none of these can confidently be considered as resulting from recombination. The conserved haplotype, 1/2/4/5-3-6-2, constructed from the marker order 458b14-458a13-458b57-369a89, was not found in any unaffected individuals.

7.3 Sequence Analysis of the HMSNL Region

Genomic sequencing of the HMSNL region was performed by the Centre for Molecular Biotechnology in Jena, Germany under the auspices of the Human Genome Project.

7.3.1 Genomic Sequencing of the Entire HMSNL Region

A tiling path composed of ten genomic clones provides a minimally redundant map of contiguous clones spanning the entire HMSNL region (genomic clones in blue in figure 7-1). This included eight clones identified through BAC library screening and a single PAC clone. These entire clones were sequenced (clones and Genbank accession numbers in table 7-5). With the addition of clone 218N23, identified in searches of public databases, complete coverage of the region was obtained. Genomic sequencing of the HMSNL region resulted in a total of 1,002,463 nucleotides (≈1Mb) of DNA sequence.

Table 7-5
Genomic clones sequenced in the HMSNL critical region

Clone	Accession number[φ]
543J1	AF216667
423F14	AF257497
215C12	AF228727
137K3	AF230666
326J4	AF230667
218N23	AF305872
709	AF235100
458A3	AF192304
369M3	AF186190
259J15	AF186191

[φ]http://www.ncbi.nlm.nih.gov/Genbank/

7.3.2 Genomic Structure of Genes in the HMSNL Region

Sequence analysis of the 1Mb region revealed the genomic structure of five genes (table 7-6). The genomic structure of three of these genes had previously been reported: *TG* (Baas, van Ommen, Bikker, Arnberg, & de Vijlder, 1986), *SIAT4A* (Chang, Eddy, Shows, & Lau, 1995) and *SLA* (Meijerink et al., 1998). In addition to the five genes, a pseudogene designated progesterone receptor-associated p48 protein, was identified in the region.

Table 7-6
Genes contained within the HMSNL region

Gene	Locus ID[Ψ]
Thyroglobulin (*TG*)	7038
Src-like adaptor (*SLA*)	6503
N-myc down-regulated gene 1 (*NDRG1*)	10397
Wnt1-inducible signalling protein 1 (*WISP1*)	8840
Sialyl transferase 4A (*SIAT4A*)	6482

[Ψ] Available at http://www.ncbi.nlm.nih.gov/LocusLink

7.3.3 Integration of Genomic Sequence with Physical and Genetic Maps

The critical region was reduced to a 202kb genomic segment flanked by the markers pJ10 and369CA3. Sequence analysis showed that *SLA* and most of *TG* lie centromeric of pJ10. *SIAT4A* is located telomeric of marker 369CA3. Four exons of *TG* were found in this critical region; however *TG,* a precursor of thyroid hormones, was considered to be an implausible functional candidate. Thus, two complete genes were contained within the critical region: *NDRG1* and *WISP1*. *WISP1* spans 38kb of genomic sequence and comprises 5 exons oriented on the sense strand. *NDRG1* spans 60kb of genomic sequence and comprises 16 exons oriented on the antisense strand.

A fine structure integrated map of the genomic structure of the two genes determined by sequencing, and the physical and genetic map of the HMSNL critical region was constructed (Figure 7-2). The three markers 369a89, 458b57 and 458a13 that are homozygous in all affected individuals are contained within the intronic sequence of *NDRG1* and span a region of 40kb.

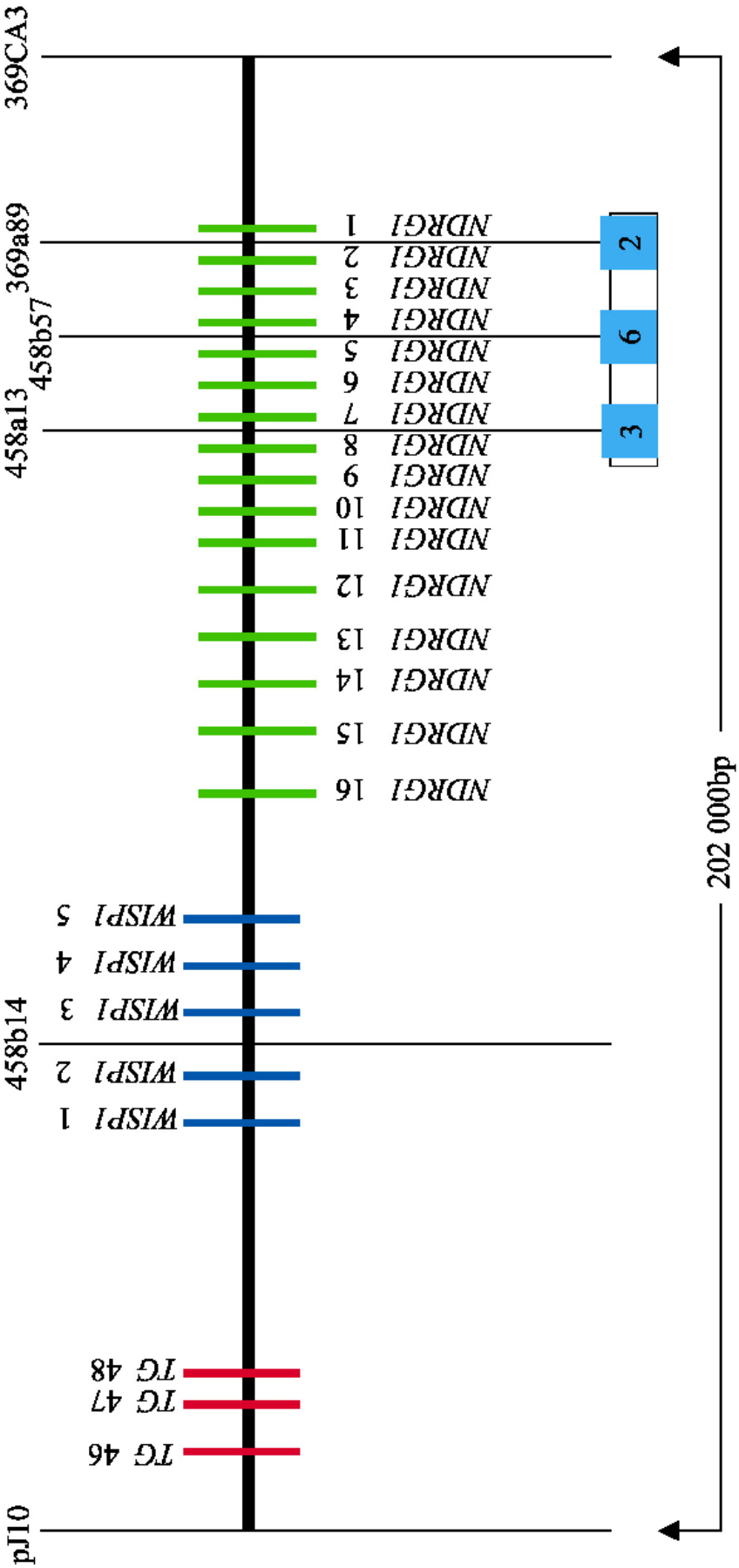

Figure 7-2 Integrated physical and genetic map of the HMSNL critical region. The genomic structure of *WISP1* and *NDRG1* as determined by large-scale sequencing is shown. A three marker haplotype is homozygous in all 120 disease chromosomes.

7.4 Mutation Analysis of HMSNL Candidate Genes

A panel of DNA samples from three homozygous affected individuals and three unaffected non-carrier family members, as determined by haplotypes, was screened for mutations in all exons of *WISP1* and *NDRG1* using direct sequencing.

7.4.1 Sequence Analysis of *WISP1* in Affected Individuals

The five exons of *WISP1* were sequenced in the panel of six DNA samples. No sequence variants were found in the patients. An unaffected individual was heterozygous for a single C→T transition in exon five of *WISP1*. This SNP would result in a predicted silent mutation at codon 307, retaining an asparagine codon at this position.

7.4.2 Sequence Analysis of *NDRG1* in Affected Individuals

The sixteen exons of *NDRG1* were sequenced in the panel of six DNA samples. A cytosine to thymine transition was observed at a CpG site in codon 148 contained in exon 7 of the three affected individuals. This C→T mutation results in the replacement of an arginine codon with a stop codon in the transcribed messenger RNA and a predicted truncated protein product. The R148X mutation was observed to be in the homozygous state in the DNA of all three affected individuals (figure 7-3). Sequence analysis of obligate carriers revealed heterozygosity at this site. The remaining 15 exons of *NDRG1* did not display any sequence variation in affected individuals.

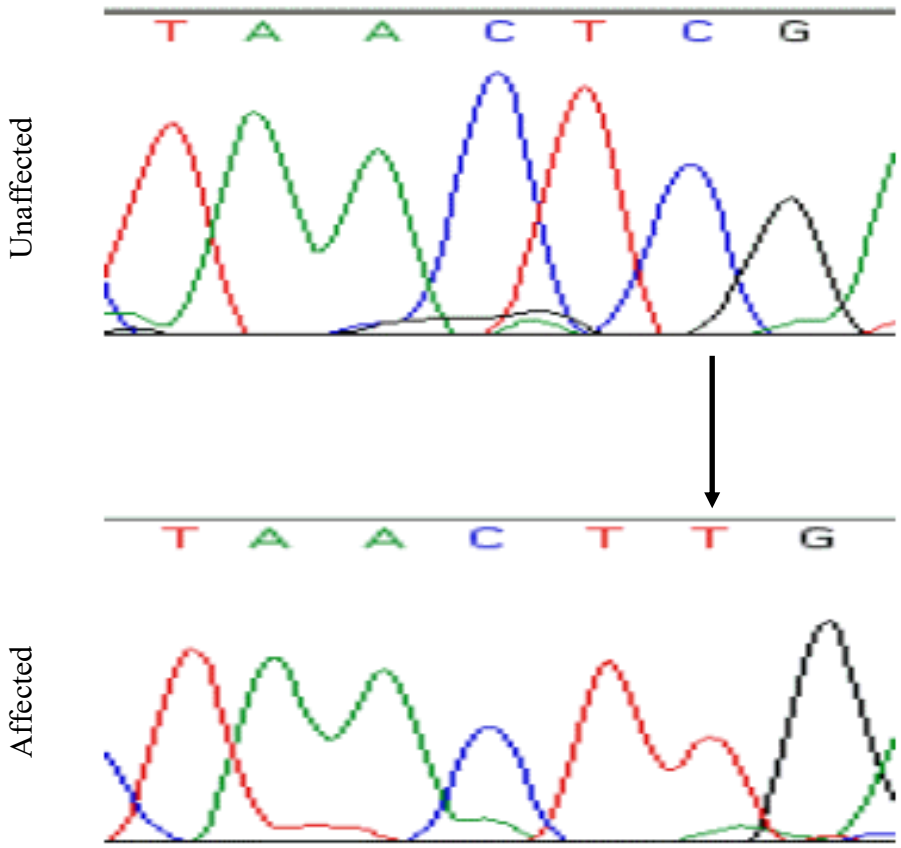

Figure 7-3 Chromatogram showing the C→T transition in DNA sequence in exon 7 of *NDRG1* in affected individuals.

7.4.3 Segregation Analysis of the R148X Mutation in HMSNL Families

Members of all HMSNL families were tested for the putative disease-causing mutation.

7.4.3.1 Analysis of R148X mutation using T*aq*1 restriction endonuclease

The C→T transition results in the abolition of a T*aq*1 restriction site. Segregation analysis in the HMSNL families was performed using a T*aq*1 digest assay (Figure 7-4). The R148X mutation was found to be in the homozygous state in all 60 affected individuals included in the study. Carriers were found to be heterozygous for the mutation. Unaffected family members who were predicted to be non-carriers on the basis of haplotype analysis did not have the R148X mutation.

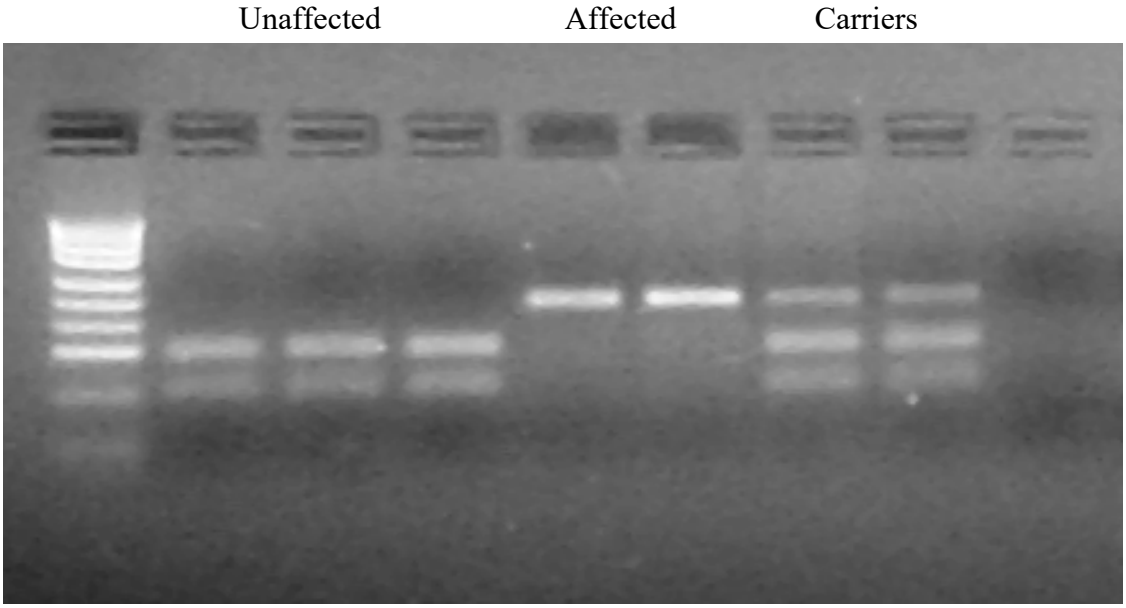

Figure 7-4 Agarose gel containing products of T*aq*1 digests of exon 7 of *NDRG1* in samples of HMSNL affected, carrier, and noncarrier individuals.

7.4.3.2 Identification of a null allele mutation in exon 7 of *NDRG1*

Segregation analysis of the R148X disease-causing mutation revealed an anomalous pattern in some unaffected family members. Haplotype analysis had identified these individuals as carriers of the disease allele. However, in these individuals the PCR failed to amplify the wildtype allele. Thus, genotype analysis of samples from these individuals suggested that they were homozygous for the disease-causing mutation. Sequence analysis identified a T→C SNP within the annealing site of the original forward PCR primer (Figure 7-5). Therefore, it is likely that primer annealing was compromised by this SNP resulting in the non-disease allele failing to amplify during PCR.

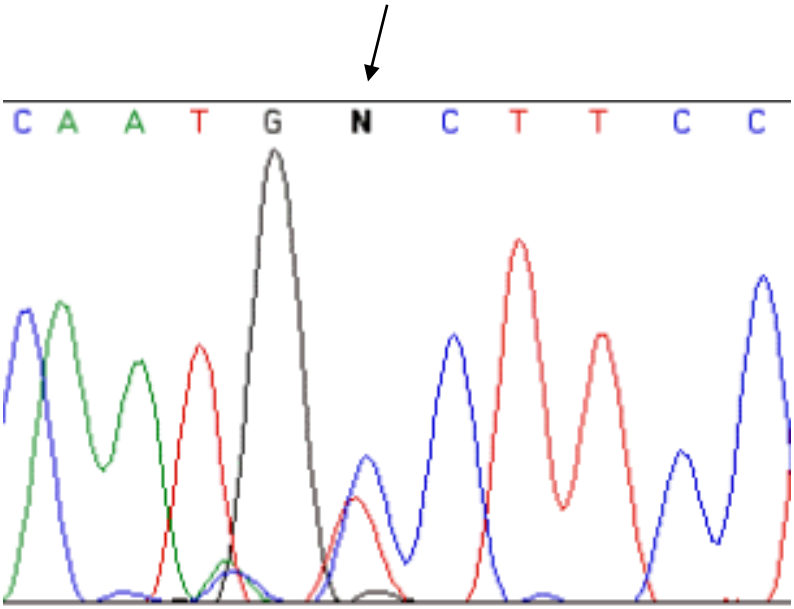

Figure 7-5 Sequence confirmation of a primer T→C SNP in the original PCR primer used for the R148X assay. Both alleles are observed in this sample.

Haplotypes from individuals in which the null allele mutation was observed were constructed using the genotyping results from six linked polymorphic loci (table 7-7). The haplotypes are evidently related, although the most recent common ancestor (MRCA) would be ancient as suggested by the variation observed at different loci. However, the null allele mutation is in complete linkage disequilibrium with allele 4 of the 458a13 locus. This locus is located 1.5kb proximal to exon 7 of *NDRG1*. Allele 4 is the most frequent allele at this locus in normal chromosomes, occurring at a frequency of 0.295 in non-disease haplotypes. However, complete linkage disequilibrium with the nearby locus and the dispersal of this null allele mutation in seven extended pedigrees from four different Romani populations suggests that it is a founder mutation.

Table 7-7

Haplotypes associated with the NDRG1 exon 7 null allele mutation

Family	pJ10	458b14	458a13	458b57	369a89	369CA3
Lom 2	4	1	4	4	4	1
Monteni 1	?	1	4	4	8	2
Lom 3	4	1	4	4	4	1
Lom 4	4	5	4	6	4	?
Kalderash 1	4	1	4	2	?	4
Italian Roma 1	1	3	4	4	7	?
Monteni 2	3	1	4	6	7	3

CHAPTER 8

DISCUSSION

8.1 A Physical Map of the HMSNL Region

An essential step in positional cloning is characterisation of the chromosomal segment to which the disease gene has been mapped. This is achieved through the identification of genomic clones that represent the region of interest. The construction of a map of contiguous clones covering the region provides a resource can that can then be used for the development of a physical map. In addition, these genomic clones can be screened to identify novel polymorphic markers and genes.

A human BAC library was screened to obtain clones mapping to the region of 8q24 in which the HMSNL disease gene had been localised using genetic linkage analysis. BACs were chosen as the cloning system as they contain genomic fragments of a reasonable size (an average of 120kb). Although the larger insert size of YACs makes them a more efficient mean of covering a genomic region, the high frequency of chimerism, rearrangements and deletions preclude their usefulness. On the other hand, cosmid clones have small genomic inserts, which makes them a less effective means of covering large regions of the genome. Although the genomic inserts of BACs are of a significant size, the construction of a contig is a laborious process. The identification of each unique BAC clone from the genomic library requires at least 152 individual PCRs.

Probing of the BAC library was initiated using polymorphic markers that had been used to map the gene. Chromosome walking was initiated from clones identified using these markers. In the early rounds of assembling genomic clones there are a number of uncertainties, including which direction along the chromosome one is "walking". In this case, it was not until large-scale sequencing of the region was in progress that it became apparent that the centromeric subcontig was incorrectly oriented. In retrospect, the judicious use of methods to orient the BAC clones relative to each other and to determine physical distances between STSs may have reduced the amount of time and labour expended in constructing the BAC contig. Methods that could be

used include radiation hybrid mapping and fluorescent *in situ* hybridisation (FISH), however both are of limited use when examining such small regions.

Library screening identified thirty overlapping BAC genomic clones. These clones provided a highly redundant representation of the genomic region. This is an inevitable outcome as there is no way of knowing the degree of additional coverage that a clone will provide prior to it identification. Fingerprint analyses of clones will allow an estimation of the overlap, but does not greatly ameliorate the amount of labour involved. Genomic representation in the Human Bacterial Artificial Chromosome DNA Pools Release IV library (Research Genetics) was insufficient to provide complete coverage of the region. Screening of a PAC library provided one redundant clone and one clone that extended into the unrepresented genomic region. This provided almost complete coverage of the region; however, the gap in the contig was only closed after genomic sequencing by a clone available in the public database (http://www.ncbi.nlm.nih.gov/). The difficulty in obtaining a clone representing this region suggests that some regions of the human genome may not be as readily cloneable as others.

Redundant clone coverage aids in STS content mapping by generating novel STSs from genomic insert ends and allowing the comparative physical mapping in clones. This provides a means of determining the physical order of STSs. In this study, 63 STSs were physically mapped. Of the mapped STSs, 35 were BAC insert ends. These sequence fragments were routinely checked for homology with genes and ESTs using BLAST (Altschul et al., 1997) searches, but this approach was not fruitful. Therefore, the STSs served only to establish the order of clones, thereby aiding in physically mapping ESTs, genes and polymorphic loci. In this study, the sequencing of the SP6 end of BAC 543J1 yielded a $CA_{(n)}$ repeat which was polymorphic and subsequently used in genetic mapping.

Screening genomic clones identified nineteen novel microsatellites. They were physically mapped in the contig along with the three known markers; D8S526, D8S256 and AFM116yh8. The correct physical order of polymorphic markers is essential for mapping recombination breakpoints identified through refined genetic mapping. Five ESTs were placed on the contig. Two of these were from fully characterised cDNAs:

Cda0zg03, which is derived from *NDRG1*, and L13972, which is derived from *SIAT4A.* The other three ESTs were derived from unknown genes. Physical mapping placed L13972 and another of the unknown ESTs telomeric of marker D8S256. This marker was shown to be involved in recombination by Kalaydjieva et al., (1996), therefore these two ESTs were excluded as positional candidates early in the study.

Construction of a BAC contig and physical mapping provided almost complete coverage of a 1Mb region containing 63 ordered STSs. This contig served as the framework in which refined genetic mapping and positional cloning could be pursued.

8.2 Fine-Scale Recombinant Mapping of the HMSNL Locus

Fine structure recombination mapping of the HMSNL locus was performed using 24 microsatellite loci. Initial localisation of the gene reported complete homozygosity in affected individuals at loci D8S378 and D8S529 (Kalaydjieva et al., 1996). Flanking markers D8S558 and D8S256 showed evidence of historical recombinations and therefore defined the boundaries of the critical region, estimated to be 3cM (Kalaydjieva et al., 1996). Of the 19 new microsatellites mapped in this interval, 18 were identified by probing sub-libraries of genomic clones, whilst one was fortuitously found by sequencing a BAC insert end. The construction of disease haplotypes using a dense marker map (i.e. 24 markers in a 3cM interval) allowed the discrimination of allelic variation at loci as a result of mutation from allelic variation due to recombination. Therefore, although 21 of the 24 loci (i.e. 87.5%) displayed more than one allele across the disease haplotypes, only 5 loci showed evidence of mutation. To reduce the workload, not every microsatellite locus was genotyped in every individual. New microsatellites were first tested for polymorphisms and those that had exhibited some variation were analysed in recombinant haplotypes. In some cases, it was apparent that a locus was bounded by the conserved haplotype and therefore inappropriate for refining the region. In these cases, the alleles were inferred based on the most common disease-associated allele. Unique disease haplotypes were defined by recombination and mutation. Thus, it is possible that additional unique disease haplotypes would be identified if all possible mutations were observed; however, these are not useful in refining the critical region.

Fine structure genetic mapping aimed at identifying historical and parental recombination in order to reduce the critical region. Therefore, haplotypes were constructed for every affected individual regardless of their relationship to other persons in the study. From the sample of 120 disease haplotypes, thirteen unique recombinant haplotypes were identified. Three of these were only differentiated from each other by mutations at marker D8S378 and an additional pair of haplotypes had a common recombination on the telomeric side, but were differentiated by an additional centromeric recombination in one of the haplotypes. Therefore, the refined genetic mapping revealed 10 recombination events that served to reduce the size of the critical region. Only one recombinant haplotype was due to a parental recombination. This recombinant haplotype alone would have served to exclude the centromeric region up to marker pJ10. However, the exclusion of this region was buttressed by four different historical recombinations, all of which extended at least as far as marker pJ10. On the telomeric side, a single historical recombination excluded the region up to marker 369CA3. Therefore, the 3cM locus defined by Kalaydjieva et al., (1996) was reduced to a critical region of approximately 202kb, bracketed by markers pJ10 and 369CA3. A four-marker conserved haplotype was found in all affected individuals in this region. This was defined by three homozygous loci, 458a13, 458b57 and 369a89, which are in complete linkage disequilibrium with the gene defect. Marker 458b14 displayed allelic heterogeneity, which was conservatively considered as resulting from microsatellite mutations.

The refined genetic mapping of the HMSNL disease locus benefited greatly from the inclusion of affected individuals and families from different Romani populations. The five recombinations that were used to define the centromeric boundary were found in individuals from five different Romani populations and the telomeric boundary was defined by a historical recombination observed in individuals from two different populations. Thus, it is apparent that genetic mapping in Romani populations bears a number of parallels to findings in other population isolates. Initial localisation of the gene benefited from the preservation of relatively large IBD chromosomal segments in the Lom kindred, enabling a search for shared segments in a 10cM genome scan (Kalaydjieva et al., 1996). The possibilities for refined mapping, however, were limited

in this population and this task was facilitated by the examination of individuals from a variety of Romani populations. This is analogous to the localisation of disease genes in the "New Finnish" population, which is genetically young and homogeneous, and the refinement of disease loci in the "Old Finnish" population, which displays greater genetic heterogeneity (Jorde, Watkins, Kere, Nyman, & Eriksson, 2000). Within-population variation of HMSNL disease haplotypes is low in comparison to the variation observed in the entire sample of individuals from all Romani populations. It is conceivable that the HMSNL locus could have been mapped by searching for large IBD segments within individuals from any one of the populations with a sufficiently large sample size. However, this would not have been feasible if shared segments were sought in the entire sample given the overall heterogeneity of disease haplotypes. Conversely, refined genetic mapping would not have been as successful if it was performed in affected individuals from only one of these populations. As these populations are descended from a common ancestral population, an IBS chromosomal fragment can be inferred as being IBD from a distant ancestor. Reduction of the critical region was dependent on the study of individuals from socially and geographically separated Romani groups that nevertheless share a common ancestor. Knowledge of endogamous practices and limited genetic diversity in Romani populations, and observations from this study, suggest that identical disease haplotypes may be represented many times over within a population; whereas, sampling from a variety of Romani populations increases the likelihood of observing different disease haplotypes, which will further refine the critical region.

8.3 Identification of the HMSNL Disease Gene and Mutation

Exhaustive fine-scale recombination mapping resulted in a marked reduction in number of positional candidates for HMSNL. The 1Mb HMSNL region, which was cloned in this study, contained five genes. *NDRG1*, *SIAT4A*, *TG* and *SLA* had been located in the contig through physical mapping of ESTs. *SIAT4A*, *SLA* and most of *TG* were excluded on the basis of historical recombinations observed in the conserved disease haplotype. The genomic structures of *NDRG1* and *WISP1* were determined through large-scale genomic sequencing. *NDRG1* is transcribed as a 3,020bp mRNA,

which encodes a protein of 394 amino acids. *WISP1* is transcribed as a 2,830bp mRNA encoding a protein of 367 amino acids. On the basis of refined genetic mapping, these two genes and exons 46-48 of *TG* were the only positional candidates. Thus, prioritising refined genetic mapping over sequence analysis of positional candidate gene greatly reduced this final stage of investigation.

The entire genomic coding regions of *NDRG1* and *WISP1* were examined. The common mutation identified in all HMSNL individuals is a C→T transition at codon 148 in *NDRG1*. The genome level mutation results in a stop codon replacing an arginine codon in the transcribed messenger RNA. Thus, the predicted mutant mRNA would produce a polypeptide of 147 amino acids. This can be hypothesised to result in either a truncated protein of aberrant function or a degraded protein, either of which would be expected to result in biological malfunction at the cellular level. No other nucleotide variants were identified in *NDRG1* or *WISP1* in affected individuals. The disease-causing mutation segregates in all families in complete agreement with autosomal recessive inheritance. Therefore, the evidence that the C→T transition in *NDRG1* is the disease-causing mutation fulfils the majority of criteria outlined by Cotton and Scriver (1998). These authors claimed that disease-causing mutations should be rare and therefore, population controls should be screened to demonstrate this to be the case. However, this is not deemed to be an appropriate criterion in the Roma where disease alleles can occur at frequencies well over 1% (Kalaydjieva et al., 1999; Plasilova et al., 1999). In this study, the problems described by Bonné-Tamir et al. (1997) in proving that a single genome-level mutation is the disease-causing mutation in population isolates were overcome. This is because refined genetic mapping reduced the conserved disease allele to one containing just two genes, which could be completely characterised.

NDRG1 encodes a gene of unknown function. The cDNA for *NDRG1* was first reported by Kokame, Kato, & Miyata (1996) and termed reducing agents-tunicamycin-responsive protein (RTP). It was subsequently reported by van Belzen et al., (1997) who called it differentiation-related gene (DRG1) and Zhou, Salnikow, & Costa, (1998) who called it CAP43. *NDRG1* expression has been shown to be up-regulated in growth-arrested cells (Piquemal et al., 1999; van Belzen et al., 1997) and repressed in transformed cells (Kurdistani et al., 1998; van Belzen et al., 1997). The protein has been

observed in the cytosol and found to accumulate in the nucleus during DNA damage (Kurdistani et al., 1998). Database searches for a serial array of gene expression (SAGE) tag analysis, related ESTs and cDNA expression libraries, reveal that it is ubiquitously expressed in a variety of tissues and cells.

The NDRG1 protein is highly conserved across species. The protein shares 93% identity with the mouse homologue, 64% identity with the rat homologue, 29% identity with the *Drosophila melanogastor* homologue and 30% with the *C. elegans* homologue. However, these proteins show no homology to known protein motifs, apart from a possible phosphopantetheine-binding site (Kokame et al., 1996) and some similarity to the ligand-binding domain of the inositol 1,4,5-triphosphate receptor (Krauter-Canham et al., 1997).

Taken together, previous studies of NDRG1 suggest that it may act as a signalling molecule or a chaperone. However, elucidation of the molecular pathology of HMSNL requires functional studies of *NDRG1*. Neuropathology studies of HMSNL peripheral nerve suggest that the primary defect may be in Schwann cell malfunction, possibly compromising axon-Schwann cell signalling (Kalaydjieva et al., 1998). A number of other genes involved with myelination have been identified, including protein myelin protein 22kd (PMP22), protein zero (P0), and connexin32 (reviewed in Scherer, 1999). NDRG1 may interact with these structural proteins and its aberrant expression may be the underlying cause of the HMSNL pathology. In addition, other genes involved in demyelinating neuropathies such as myotubularin–related protein-2 (Bolino et al., 2000), and others yet to be identified, represent possible protein partners in signalling pathways or structural complexes.

8.4 A Putative Founder Null Allele Mutation

An interesting incidental finding of this study is the observation of a null allele mutation that compromised the genomic assay for R148X mutation. This rogue allele is a normal chromosome with a SNP at the 3' end of the annealing site of the forward PCR primer used to amplify exon 7 of *NDRG1*. The presence of this SNP prevented the non-R148X allele from being amplified by the standard PCR protocol. Thus, individuals predicted by haplotype analysis to be heterozygous for the disease-causing mutation

appeared as homozygotes. SNPs are considered to be unique event mutations, which suggests that the chromosomes bearing this SNP in the population are related. Indeed, the haplotypic background on which this SNP occurs shows some evidence of evolutionary relatedness. However, only the most proximal marker is in complete linkage disequilibrium with the SNP, which suggests that the mutation is ancient.

As PCR is commonly used for diagnostic purposes, this observation is of great relevance. SNPs occur at an average of once every 1kb in the human genome (Kruglyak, 1999). Thus, the likelihood of a primer site mutation is not insignificant. Previous reports of haemochromatosis screening described a similar situation of a primer site mutation resulting in a null allele using a common diagnostic PCR (Jeffrey, Chakrabarti, Hegele, & Adams, 1999). It was argued that the consequences of this polymorphic allele in diagnostic analyses were minimal, particularly as it was used in conjunction with biochemical assays (Gomez et al., 1999; Merryweather-Clarke et al., 1999; Noll, Belloni, Stenzel, & Grody, 1999). In a population isolate, if such an allele is subjected to founder effect, its frequency could be greatly increased and pose a significant problem for predictive and diagnostic DNA testing. This is of even graver concern if alternative supporting diagnostic measures are not possible. This suggests that DNA testing should be duplicated using different primer pairs and/or in conjunction with haplotype analysis to ensure that this problem is minimised.

SECTION III

STUDY OF THE GENETIC EPIDEMIOLOGY OF DISEASE ALLELES IN THE ROMA

CHAPTER 9

SUBJECTS AND METHODS

9.1 Introduction

9.1.1 Summary of Previous Findings

Studies of genetic disorders in the Roma thus far suggest a non-random distribution of disease prevalence in different Romani populations. Clustering of disorders in geographically and socially separated Romani groups is evident (Abicht et al., 1999; Kalaydjieva et al., 1996; Piccolo et al., 1996). The disease gene frequency for autosomal recessive disorders in the Roma has been determined in a number of studies (Kalaydjieva et al., 1999; Plasilova et al., 1999; Todorova, Ashikov, Beltcheva, Tournev, & Kremensky, 1999). The carrier rates for these disorders in the general Roma population has ranged from 2-6%. However, preliminary findings suggest an increased carrier rate in high-risk groups. This is exemplified by the E378K founder mutation in *CYP1B1* causing primary congenital glaucoma, for which a carrier frequency in a high risk group is 11% (Plasilova et al., 1999) as compared to 5% estimated for the general Romani population (Ferak, Gençik, & Gençikova, 1982). The population structure of the Roma and group-specific endogamy implies that common deleterious genes may be found at vastly different frequencies in Romani populations. This distribution is likely to be affected by the time of population founding, the number of founders and differing population histories.

Disease gene haplotype analysis provides a means of investigating the history of a gene in a population, the history of the population, and relationships between populations. This analysis has been applied both to continental populations (Mateu et al., 2001; Morral et al., 1994) and to structured populations (Varilo, Nikali, Suomalainen, Lonnqvist, & Peltonen, 1996). Disease haplotype analyses of private founder mutations in Romani populations have demonstrated conserved disease

haplotypes exhibiting some degree of variation (Kalaydjieva et al., 1996; Kalaydjieva et al., 1999; Piccolo et al., 1996). However, a systematic study of variation in disease gene haplotypes in different Romani populations has not been performed. Such investigations could provide additional insights into the histories and relationships of Romani population.

The age of a disease gene can be estimated by the degree of haplotypic decay of the ancestral chromosome. Haplotype decay is directly correlated with the extent of linkage disequilibrium around rare variants. Linkage disequilibrium mapping in isolated populations is an alternative approach to disease gene identification, and has been proposed as a means of identifying genes involved in complex traits (Jorde, Watkins, Kere, Nyman, & Eriksson, 2000). However, there is uncertainty about which populations are most appropriate for linkage disequilibrium studies (Kruglyak, 1999a, 1999b; Terwilliger, Zollner, Laan, & Paabo, 1998). Future efforts to map the genes underlying both monogenic and complex traits in the Roma can be expected to benefit from an examination of linkage disequilibrium around known disease loci.

9.1.2 Research Questions

This study aimed at investigating the distribution, haplotypic diversity and history of two private founder mutations in Romani populations: R148X in *NDRG1* on chromosome 8q24 and C283Y in *SGCG* on chromosome 13q12.

The specific research questions were:

1. How do carrier and gene frequencies of the two founder mutations vary amongst Romani populations?
2. What are the implications of autosomal recessive gene frequencies for gene mapping in the Roma?
3. What is the history of the disease genes in these populations?
4. What does haplotype diversity indicate about population history in the different populations and the historical relationship between different Romani population?
5. How does linkage disequilibrium behave around rare variants in different Romani populations?

6. What are the appropriate approaches to LD mapping in the Roma?

9.1.3 Subjects and Study Design

Screening for the two founder mutations was performed in individuals from eight Romani populations (table 9-1). These Romani populations provide representation of the Balkan, Vlach and Western European migrational groupings. For this study, not all populations investigated in the population genetic study (section I) were investigated. This is because the Monteni and Kalderash sample used in section I included families selected for HMSNL, thus disallowing an unbiased sample for examining the frequency of the HMSNL mutation. Population samples of the Lom, Kalaidjii and Lingurari were obtained by Drs Kalaydjieva and Angelicheva in Bulgaria during a field trip in June 2000. The total sample size for C283Y testing was 573, because Turgovzi and Feredjelli data were obtained from the community genetic study (section IV). The total sample size for R148X testing was 348, as smaller samples of the Turgovzi and Feredjelli populations were screened.

Table 9-1
Populations included in the study of the R148X and C283Y mutations

Population	Metagroup	Migrational group	Geographic location	Sample size
Turgovzi	Xoroxane	Balkan	Omurtag, NE Bulgaria	50-224
Feredjelli	Xoroxane	Balkan	Omurtag, NE Bulgaria	19-101
Spanish Roma	Caló	West European	Madrid, Spain	68
Lithuanian Roma	Russian	West European	Vilnius, Lithuania	20
Intreni	Rudari	Vlach	Letnitza, N Bulgaria	28
Lom	Jerlii	Vlach	Lom, NW Bulgaria	62
Kalaidjii	Jerlii	Vlach	NW & SW Bulgaria	43
Lingurari	Rudari	Vlach	N & S Bulgaria	45
Total				**348-573**

The C283Y disease haplotypes were characterised in twenty-four unrelated C283Y homozygous affected individuals from the Turgovzi population. These individuals were from the families identified in the genetic screening program (section IV). Where available, both parents of affected individuals were genotyped in order to resolve the phase of alleles. Genotyping data from the non-C283Y chromosomes in these parents were used to estimate allele frequencies in the population.

R148X haplotypes were determined in the course of the HMSNL gene identification study (Section II). Allele frequencies at each locus in the population were estimated from genotyping data from non-disease chromosomes in unrelated individuals.

9.2 Methods

9.2.1 Mutation Assays

The genome level C→T mutation in exon 7 of *NDRG1* that results in the HMSNL-causing R148X mutation was tested using the Taq1 restriction enzyme digest assay described in section 6.2.6.

Testing for the C283Y mutation was performed using an assay to determine the G→A transition in codon 283 located in exon 8 of *SGCG*. The transition of a guanine to an adenine creates a R*sa*1 restriction endonuclease site, enabling the use of this enzyme in discriminating between mutant and normal chromosomes. A 168 base pair fragment that includes the mutation site was amplified using the PCR primers 5'-CCT GTC TGT GGC CGG TGT GA-3' and 5'-GCG TTT ACT TCC CAT CCA CGC TGC-3' (Piccolo et al., 1996) in a 20 μL reaction mixture (table 9-2).

Table 9-2
PCR protocol for amplifying the fragment containing the SGCG C283Y mutation

Reagent	Volume
10x PCR Buffer	2.0μL
25mM $MgCl_2$	1.2μL
Forward Primer (20ng)	2.0μL
Reverse Primer (20ng)	2.0μL
dNTPs (5mM)	2.0μL
Taq Polymerase	0.05μL`
dH_2O*	8.75μL/10.75μL
DNA	2.0μL/FTA Genecard

*The volume of dH_2O was adjusted if the DNA sample was bound to a FTA Genecard

DNA amplification was performed in a GeneAmp 2400 thermocycler (Applied Biosystems) using the conditions described by Piccolo et al., (1996): an initial denaturation at 94°C for 5 mins was followed by 35 cycles of 1 min at 94°C, 1 min at 59°C and 1 min at 72°C. A final extension period of 7 mins at 72°C was allowed followed by cooling of the sample to 4°C.

A R*sa*1 restriction enzyme digest was performed subsequent to amplification of the 168 base pair DNA fragment (table 9-3). The reaction was incubated at 37°C for one hour.

Table 9-3
Rsa1 restiction digest reaction for assessment of C283Y genotypic status

Reactant	Volume
10x Buffer	1.5μL
PCR product	7.0μL
dH_2O	5.5μL
R*sa*1	1.0μL

To visualise the results, the digest products were electrophoresed on a 3.5% agarose gel at 90 volts for 30 minutes. In addition, each gel contained control samples of a known homozygote normal, heterozygote and homozygote mutant which were PCR-amplified and digested with R*sa*1 for direct comparison with unknown samples. The gel was stained with ethidium bromide and products were visualised using an ultraviolet transilluminator. An image of the gel was recorded using the Kodak gel documentation system (Eastman Kodak).

9.2.2 Characterisation of Disease Haplotypes

NDRG1 R148X disease haplotypes were constructed for the purposes of refined genetic mapping of the HMSNL disease locus (section II). Additional genotypic data were used from Kalaydjieva et al., (1996) to extend disease haplotypes.

SGCG C283Y disease haplotypes were constructed using genotyping results from five polymorphic microsatellites spanning a physical distance of 2.75Mb around *SGCG* (http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/map).

9.2.2.1 Amplification of chromosome 13q12 microsatellites

The five microsatellite loci were amplified using the fluorescently labelled primer pairs (table 9-4). The primers for microsatellites D13S232, D13S283 and D13S787 were labelled with HEX. Primers for D13S115 and D13S292 were labelled with FAM.

Table 9-4

Primer sequences for microsatellite loci used to define the C283Y haplotype

D13S787	Forward	ATC AGG ATT CCA GGA GGA AA
	Reverse	ACC TGG GAG TCG GAG CTC
D13S232	Forward	TGC TCA CTG CTC TTG TGA TT
	Reverse	GGC ACA GAA ATA AAT GTT GAT G
D13S115	Forward	TGT AAG GAG AGA GAT TTC GAC A
	Reverse	TCT TAG CTG CTG GTG GTG G
D13S283	Forward	TCT CAT ATT CAA TAT TCT TAC TGC A
	Reverse	GCC ATT CCA AGC GTG T
D13S292	Forward	TAA TGG CGG ACC ATG C
	Reverse	TTT GAC ACT TTC CAA GTT GC

.

Each microsatellite was amplified in a separate reaction (table 9-5). Amplification of these loci required the use of the heat activated DNA polymerase, TaqGold (Applied Biosystems).

Table 9-5
PCR mixtures and reactions for the five microsatellite loci used to define the C283Y haplotype

	D13S787	D13S232	D13S115	D13S283	D13S292
10x PCR buffer	2.0µL	2µL	2µL	2µL	2µL
25mM $MgCl_2$	1.2µL	1.2µL	2µL	2µL	1.8µL
Forward Primer (20ng/µL)	2.0µL	1µL	2µL	1µL	1µL
Reverse Primer (20ng/µL)	2.0µL	1µL	2µL	1µL	1µL
dNTPs (5mM)	2.0µL	2µL	2µL	2µL	2µL
TaqGold Polymerase	0.05µL	0.05µL	0.05µL	0.05µL	0.05µL
dH_2O	8.75µL	8.75µL	7.95µL	7.95µL	8.15µL
DNA	2.0µL	2µL	2µL	2µL	2µL
Thermocycling program	94°C-15 mins 35 cycles: 50s @9 4°C 50s @ 54°C 50s @ 72°C 72°C – 5 mins 4°C	94°C-15 mins 35 cycles: 40s @ 94°C 50s @ 55°C 40s @ 72°C 72°C – 5 mins 4°C	94°C-15 mins 35 cycles: 40s @ 94°C 50s @ 55°C 40s @ 72°C 72°C – 5 mins 4°C	94°C-15 mins 35 cycles: 40s @ 94°C 50s @ 57°C 40s @ 72°C 72°C – 5 mins 4°C	94°C-15 mins 35 cycles: 40s @ 94°C 50s @ 57°C 40s @ 72°C 72°C – 5 mins 4°C

9.2.2.2 Determination of 13q12 microsatellite DNA sizes

Microsatellite allele sizes were determined using the 373A DNA Analyser (Applied Biosystems). Gel and sample preparation was as described in section 3.3.4 and section 3.3.5 respectively. Two control samples of previously determined size were included on each gel to ensure compatibility in sizing between the unknown samples and previous results.

Polymorphic DNA fragments were assigned an allele number based on the designation of the smallest fragment as allele 1 and numbered sequentially (table 9-6). This nomenclature is not consistent with previously published allele calling, however the published allele numberings do not follow any logical order.

Table 9-6
Allele designations for chromosome the five 13q12 STRs used to define the C283Y haplotype

D13S115			D13S232			D13S787			D13S292			D13S283		
Size	**A**	**B**	**Size**	**A**	**B**	**Size**	**A**	**B**	**Size**	**A**	**B**	**Size**	**A**	**B**
163	1	7	108	1	7	249	1		203	1	4	128	1	6
165	2	6	110	2	6	253	2		205	2	3	129	2	3
167	3	5	112	3	5	257	3		207	3	1	131	3	
169	4	4	114	4	.	261	4		209	4		133	4	
171	5	3	116	5	.	265	5					135	5	
173	6	2	118	6	.							137	6	1
175	7	1	120	7	4							139	7	
177	8											141	8	
												143	9	
												145	10	2
												147	11	
												149	12	4
												151	13	5
												153	14	7

A= allele numbering used in this study, B= The Genome Database allele calling (http://www.gdb.org)

9.2.2.3 Construction of C283Y haplotype

Inheritance of alleles was determined by genotyping affected individuals and both parents. Haplotypes were constructed manually using the marker order D13S115-D13S232-D13S787-D13S292-D13S283 as reported on the Soton genetic map (http://cedar.genetics.soton.ac.uk/public_html/ldb.html).

9.2.6 Statistical Analyses

Statistical analyses were performed using Arlequin 2.000 (Schneider, Kueffer, Roessli & Excoffier, 1996). Genetic diversity statistics were calculated using equations provided in section 3.5. Networks were drawn using Network 2.0 (Bandelt, Forster, Sykes & Richards, 1995).

The age of disease alleles were determined using the method of Stephens et al., (1998) described in section 3.5.4. For this the rate of change, *r*, was estimated based on mutation and recombination rates. The age of disease alleles was also determined using the method of Risch et al. (1995). This method assesses linkage disequilibrium at a

linked polymorphic locus to determine a moment estimator, *t*, of the age of the mutation. If θ is small, then log(1-θ) ≈ θ. Therefore,

$$\hat{t} \approx \frac{\log\left(\frac{p_{affected} - p_{normal}}{1 - p_{normal}}\right)}{\log(1-\theta)}$$

(Guo & Xiong, 1997; Risch et al., 1995).

A generation age of 20 years was used for all calculations.

Linkage disequilibrium was determined at disease linked loci using the equation

$$P_{excess} = \frac{(p_{affected} - P_{normal})}{(1 - P_{normal})}$$

where $P_{affected}$ is the frequency of the most common disease-associated allele and P_{normal} is the frequency of that allele in the general population (Hastbacka et al., 1992; Hastbacka et al., 1994).

CHAPTER 10

RESULTS

10.1 Population Distribution and Frequencies of Founder Mutations

The R148X and C283Y founder mutations were analysed in individuals from 8 Romani populations. Carrier rates and gene frequencies were estimated for each population.

10.1.1 Population Distribution of the R148X Mutation

Population screening identified the R148X mutation in six Romani populations (table 10-1). These populations include Balkan and Vlach Roma, and the geographically distant Romani populations from Spain and Lithuania. The generalised carrier frequency of the R148X mutation in the Roma surveyed is 5.4%. However, a large variation in carrier frequencies was observed between different populations. The carrier frequency of the R148X mutation is highest in the Lom population, in which the carrier frequency is 19.4%. In most other populations in which the disease allele is found the carrier frequency is below 3%. The exception is the Lithuanian Roma in which the carrier frequency is 5%; however, the small sample size for this population makes this estimate unreliable. The mutation was not identified in the Feredjelli, a Balkan Romani population nor in the Intreni, a Vlach population. Disease allele frequencies were over 0.01 in the six populations in which the R148X mutation is found. The highest frequency of the R148X mutation was observed in the Lom (0.097).

Table 10-1
Summary of results of screening for the R148X mutation in Romani populations

Population	Number of carriers	Carrier frequency	Gene frequency
Turgovzi N=50	1	2%	0.01
Feredjelli N=19	0	0	0
Spanish Roma N=68	2	2.9%	0.015
Lithuanian Roma N=20	1	5%	0.025
Intreni N=28	0	0	0
Lom N=62	12	19.4%	0.097
Kalaidjii N=43	1	2.3%	0.012
Lingurari N=45	1	2.2%	0.011
Total N=335	**18**	**5.4%**	**0.027**

NB sample size (N) refers to number of individuals tested

10.1.2 Population Distribution of the C283Y Mutation

The C283Y mutation was found in just two Romani populations (table 10-2). Screening of the Turgovzi and Feredjelli populations was performed as part of the pilot community genetics program (Section IV). The Turgovzi had been deemed to be at high-risk based on the prevalence of LGMD2C-affected individuals. This high-risk status is confirmed by a carrier frequency of 6.25%. This corresponds to a disease allele frequency of 0.03 in this population. The only other population in which the C283Y mutation was found was the Spanish Roma in which the carrier frequency is 1.6%. Therefore, the C283Y mutation is either absent or extremely rare in Vlach groups. A generalised estimate for the carrier rate of the C283Y mutation in the Roma is 2.6%.

Table 10-2
Summary of results of screening for the C283Y mutation in Romani populations

Population	Number of carriers	Carrier frequency	Gene frequency
Turgovzi N=224	14	6.25%	0.03
Feredjelli N=101	0	-	-
Spanish Roma N=63	1	1.6%	0.008
Lithuanian Roma N=20	0	-	-
Intreni N=27	0	-	-
Lom N=61	0	-	-
Kalaidjii N=41	0	-	-
Lingurari N=36	0	-	-
Total N=573	**15**	**2.6%**	**0.013**

NB sample size (N) refers to number of individuals tested

10.2 Analysis of the R148X and C283Y Founder Mutations

10.2.1 The R148X Mutation in *NDRG1*

The R148X disease chromosome was characterised using a 24 marker haplotype, which was produced for the purposes of refined genetic mapping of the HMSNL locus. Three uninformative loci were discarded and data from additional seven loci, which had been used in the disease gene localisation study (Kalaydjieva et al., 1996), were used for the analysis of linkage disequilibrium.

10.2.1.1 R148X haplotype analysis

Twenty unique R148X haplotypes were identified in the sample (chapter, table 7-4). As is evident in this table, there is limited sharing of haplotypes between populations. Haplotype M is the ancestral haplotype (36 of 120 chromosomes) and is

present in four of the nine populations; namely the Lom, Monteni, Rumanian and Spanish Roma. All other shared haplotypes are found in the Kalderash and one or more populations. Haplotype E, which differs from haplotype M by a mutation at locus 458b14, is prevalent in the Kalderash (14 of 37 chromosomes) and found once in the Monteni. Similarly, haplotype D is common in the Kalderash (13 of 37 chromosomes) and is also found in the French Roma. Haplotype A is found in the Kalderash, Italian Roma and the German individual.

Diversity within the disease haplotypes has been generated through microsatellite mutation and recombination. Network analysis of the phylogenetic history of the R148X haplotype visually depicts the evolution of diversity within this locus (figure 10-1). In this figure, it is apparent that the ancestral haplotype M is the predominant and most widely dispersed haplotype. Diversity within the HMSNL haplotypes is generally unique to each population. Thus, diversity appears largely to have arisen within each subpopulation subsequent to the introduction of the disease chromosome. This is most apparent in the Lom, Kalderash and Monteni in whom significant sample sizes have been investigated. These three populations show different haplotype profiles and the most frequent haplotype is different in each of these populations. The putative Romani ancestral haplotype M is completely absent in the Kalderash.

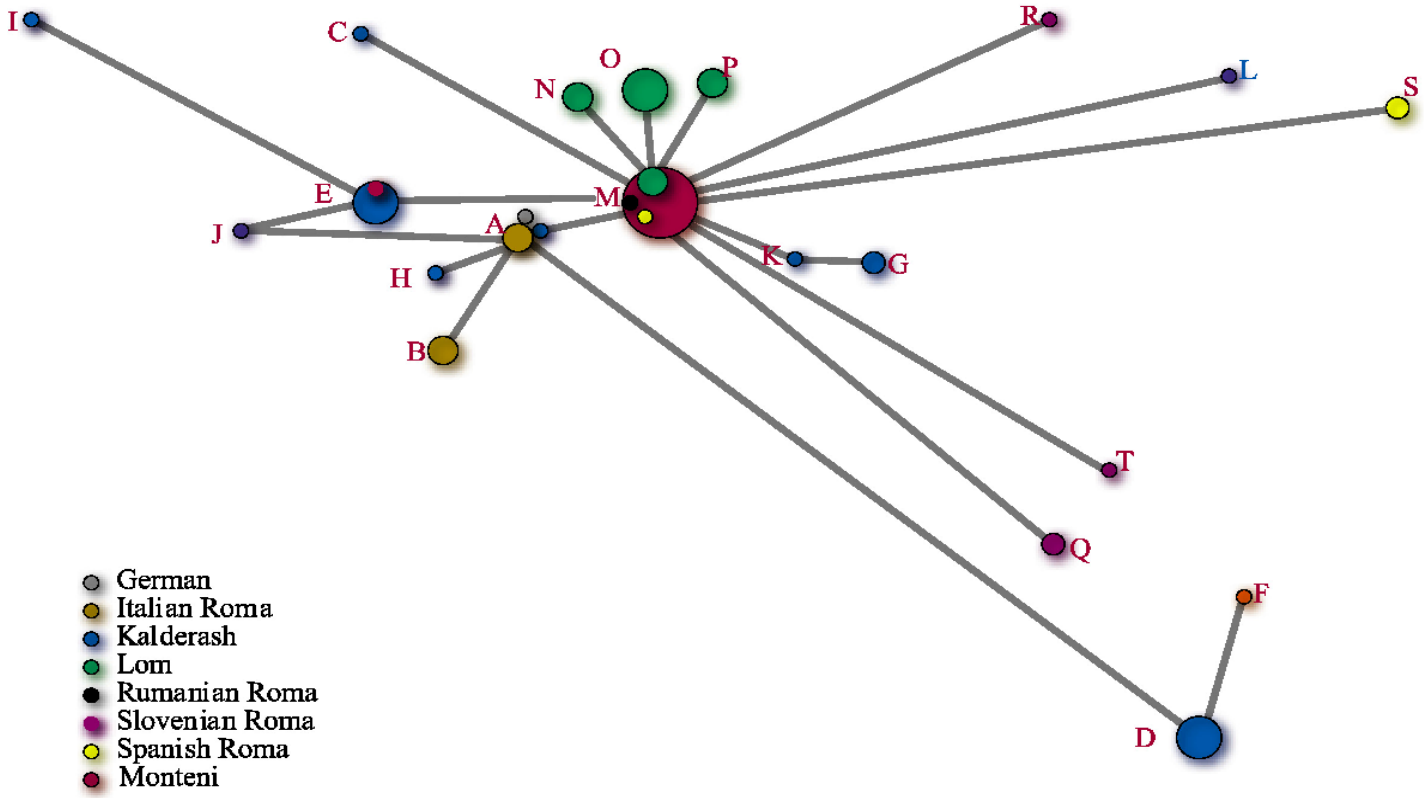

Figure 10-1 Network of R148X haplotypes in Romani populations. Nodes are proportional to the number of times the haplotype is observed in the sample. Branch lengths approximate the number of mutations at each locus and the number of loci in the recombinant segment.

Average haplotype diversity was determined for the three populations with sample sizes over 25 (table 10-3). The greatest haplotype diversity is seen in the Kalderash sample (0.75) in which 10 unique haplotypes are observed. However, even though only four haplotypes are observed in the Lom, haplotype diversity is virtually the same as in the Kalderash (i.e. 0.74). In contrast to these two populations, haplotype diversity in the Monteni population is very low (0.15).

Table 10-3
Average R148X haplotype diversity within Romani populations

Population	No. of unique haplotypes	Haplotype diversity
Kalderash (n=38)	10	0.75 ± 0.0024
Lom (n=28)	4	0.74 ± 0.0017
Monteni (n=26)	3	0.15 ± 0.0089

10.2.1.2 Age of the R148X mutation

The estimated proportion of ancestral haplotypes in the population and rate of change within the haplotype was used to calculate the coalescent age of the disease haplotype (Stephens et al., 1998). $P_{ancestral}$ was estimated as the proportion of the most frequent haplotype for each population. The rate of change, *r*, calculated for the 24 microsatellites was estimated as $24(\mu)+\theta$, where the microsatellite mutation rate, μ, is $1.2x10^{-3}$ mutations/generation (Weber & Wong, 1993) and the recombination rate over the 1Mb region is estimated to be 0.01. Thus, $r = 24\ (1.2x10^{-3)} + 0.01 = 0.0388$.

Table 10-4
Coalescent age estimates of the R148X mutation. The ancestral haplotype in each population is indicated

Population	$P_{ancestral}$	Generations	Age (20 years/gen)
Total	0.30 (haplotype M)	31	620
Lom	0.39 (haplotype O)	24	480
Kalderash	0.37 (haplotype E)	26	520
Monteni	0.92 (haplotype M)	2	40

Using the method of Stephens et al., (1998), a date of 620 years was calculated for the HMSNL mutation in the entire Roma population. Within the Lom and

Kalderash, the mutation is dated around 500 years BP, whilst the homogeneity of the Monteni R148X haplotypes suggests the mutation entered the population very recently.

The ago of the R148X mutation was estimated based on linkage disequilibrium with proximate loci (Guo & Xiong, 1997; Risch et al., 1995). Linkage disequilibrium at the two loci that define the centromeric and telomeric borders of the HMSNL critical region was used to calculate the age of the disease allele (table 10-4). As the genomic region has been completely characterised the exact physical distance between the marker and the disease-causing mutation is known. To estimate the recombination fraction, θ, the relationship 1cM = 1Mb was used. Thus, for marker pJ10, θ = 0.00187 and for marker 474CA1, θ = 0.00114.

Table 10-5
Age estimates of the R148X mutation based on linkage disequilibrium with nearby polymorphic loci

	T_{pJ10}	T_{474CA1}	$T_{average}$
Generations	57.5 gen	101 gen	79 gen
T (20y/generation)	1,150 years	2,020	1,580 years

Using markers that flank either side of the critical region for the HMSNL locus, the number of generations since the R148X mutation occurred range from 57.5 to 101. An average of values from these two polymorphic loci places the age of the mutation 1,580 years before present.

10.2.1.3 Linkage disequilibrium in the HMSNL region

The extent of linkage disequilibrium (LD) in the HMSNL locus was analysed in the Monteni, Lom, Kalderash and the whole Romani population using disease chromosomes and population controls, namely the non-transmitted chromosomes observed in HMSNL families. LD was estimated using the P_{excess} statistic for 27 microsatellite loci spanning a total genetic distance of 3.3cM (figure 10-2). $P_{affected}$ values were determined without discrimination between disease associated alleles that were a result of recombination from those resulting from mutation. This simulates a population-based case-control scenario in which parents of affected individuals are not

necessarily available, and therefore the phase of alleles in affected individuals is unknown.

As can be seen, the P_{excess} values increase with closer proximity to the disease-causing mutation. P_{excess} maxima correspond with the most likely location of the disease-causing variant. The P_{excess} value for the sample of all disease chromosomes peaks at the three loci that are invariably homozygous in all affected individuals. In contrast, P_{excess} in the Kalderash and Monteni show three independent peaks. The peak P_{excess} value for the Lom extends from marker 189CA17 to 369a89. The P_{excess} maximum in the whole sample corresponds to the only location at which P_{excess} maxima in all three populations overlap.

An aberrant P_{excess} value is observed at the D8S378 locus. This is a result of hypermutability at this locus within the Kalderash sample, which causes undetectable LD in the Kalderash population and a dramatic inconsistency in the LD trend over the HMSNL locus in the entire sample. This isolated drop in LD indicates a locus- and population-specific phenomenon, reminiscent of a microsatellite instability phenotype (Thibodeu, Bren & Schaid, 1993)

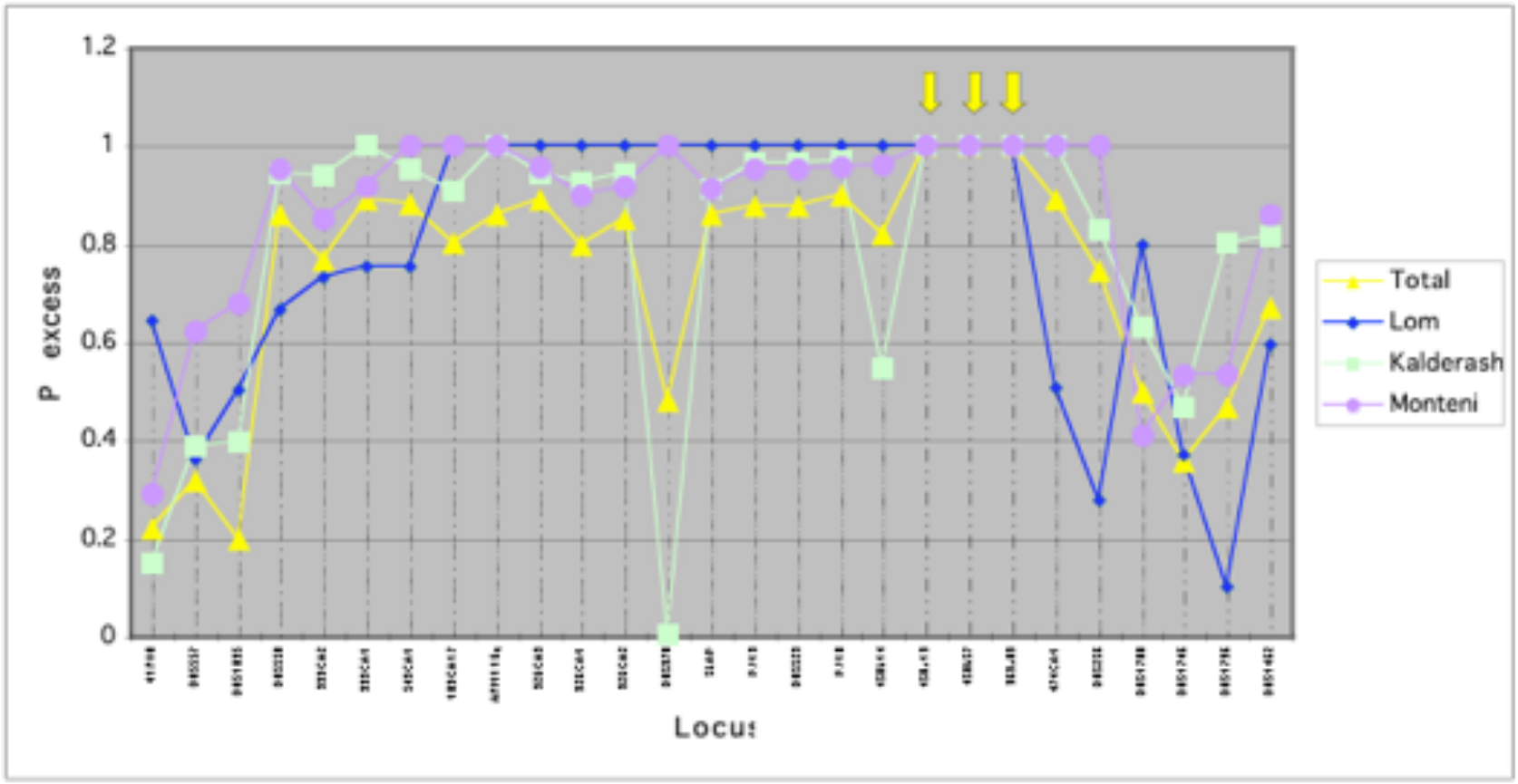

Figure 10-2 Linkage disequilibrium over the HMSNL region assuming a simulated phase-unknown scenario. LD was measured over a 3.3cM region using a disease haplotype constructed from 27 polymorphic loci. The overlapping region of maximum linkage disequilibrium in all three populations is indicated.

10.2.2 The C283Y Mutation in *SGCG*

10.2.2.1 C283Y haplotype analysis

Eleven unique haplotypes were identified in the sample of forty-eight Turgovzi disease chromosomes (table 10-5). Haplotypes are closely related to each other and clearly derived from an ancestral chromosome, designated haplotype 1. All C283Y disease haplotypes were found to be associated with allele 3, an 112bp fragment, of marker D13S232. This was consistent with the study of Roma from Western Europe by Piccolo et al., (1996) that found the C283Y mutation invariably associated with this allele[1]. Haplotype 1 is the most frequent in the Turgovzi representing 60.4% of all disease haplotypes. Haplotype 2 accounts for 14.6% (7 of 48 chromosomes) and haplotype 7 represents 8.3% (4 of 48 chromosomes). Haplotypes 2 and 7 differ from the ancestral haplotype at only one locus each, a mutation at D13S115 and a recombination between D13S292 and D13S283 respectively. The other eight C283Y haplotypes were found only once in the Turgovzi sample.

C283Y haplotypes in the Turgovzi were compared to the C283Y haplotypes from West European Roma (M. Jeanpierre, pers comm). This sample included eighteen disease haplotypes from unrelated Iberian Roma, four from Roma resident in Germany and 2 disease haplotypes each from French and Italian Roma. A comparison of the frequency of these haplotypes with the Turgovzi sample shows a striking difference (table 10-6). In the Iberian Romani sample, haplotype 7 is the most frequent and accounts for 66.7% (12 of 18) of chromosomes and the Turgovzi ancestral haplotype 1 is found at a frequency of 5.6%. Haplotype diversity in the Turgovzi is 0.616 ± 0.0058, and in the Iberian Roma it is 0.569 ± 0.0201.

[1] In the initial study by Piccolo et al., (1996) the authors used the GDB allele calling which designates the 112bp fragment as allele 5

Table 10-6
C283Y haplotypes in Romani populations

Haplotype	D13S115	D13S232	D13S787	D13S292	D13S283	P_{Total}	Turgovzi[1]	Iberian[2]	Other
1	2	3	3	3	10	0.459	29	1	4
2	3	3	3	3	10	0.094	7	-	-
3	7	3	2	3	2	0.014	1	-	-
4	2	3	2	3	10	0.014	1	-	-
5	5	3	3	3	10	0.014	1	-	-
6	2	3	3	3	14	0.040	1	-	2
7	2	3	3	3	2	0.216	4	12	-
8	2	3	3	3	6	0.013	1	-	-
9	3	3	3	3	1	0.013	1	-	-
10	5	3	2	3	2	0.027	1	1	-
11	2	3	3	1	2	0.013	1	-	-
12	7	3	3	3	2	0.013	-	1	-
13	1	3	3	3	2	0.013	-	1	-
14	2	3	2	3	2	0.013	-	1	-
15	5	3	3	3	2	0.013	-	1	-
16	2	3	2	1	2	0.027	-	-	2
Total							48	18	8

N.B Recombinations are shown in red and microsatellite mutations are shown in yellow.

A network relating C283Y haplotypes was constructed (figure 10-3). C283Y haplotypes are nonrandomly distributed in the Roma. Three haplotypes (1, 7 and 10) are shared between the Turgovzi and Iberian Roma, however they occur at different frequencies within the populations. Haplotypes in the Turgovzi are distributed throughout the network. Four unique haplotypes are observed in the Iberian Roma. However, haplotypes found in the Iberian Roma can be considered as a subset of those observed in the Turgovzi. Haplotypes in other European Roma are identical or closely related to Turgovzi haplotypes.

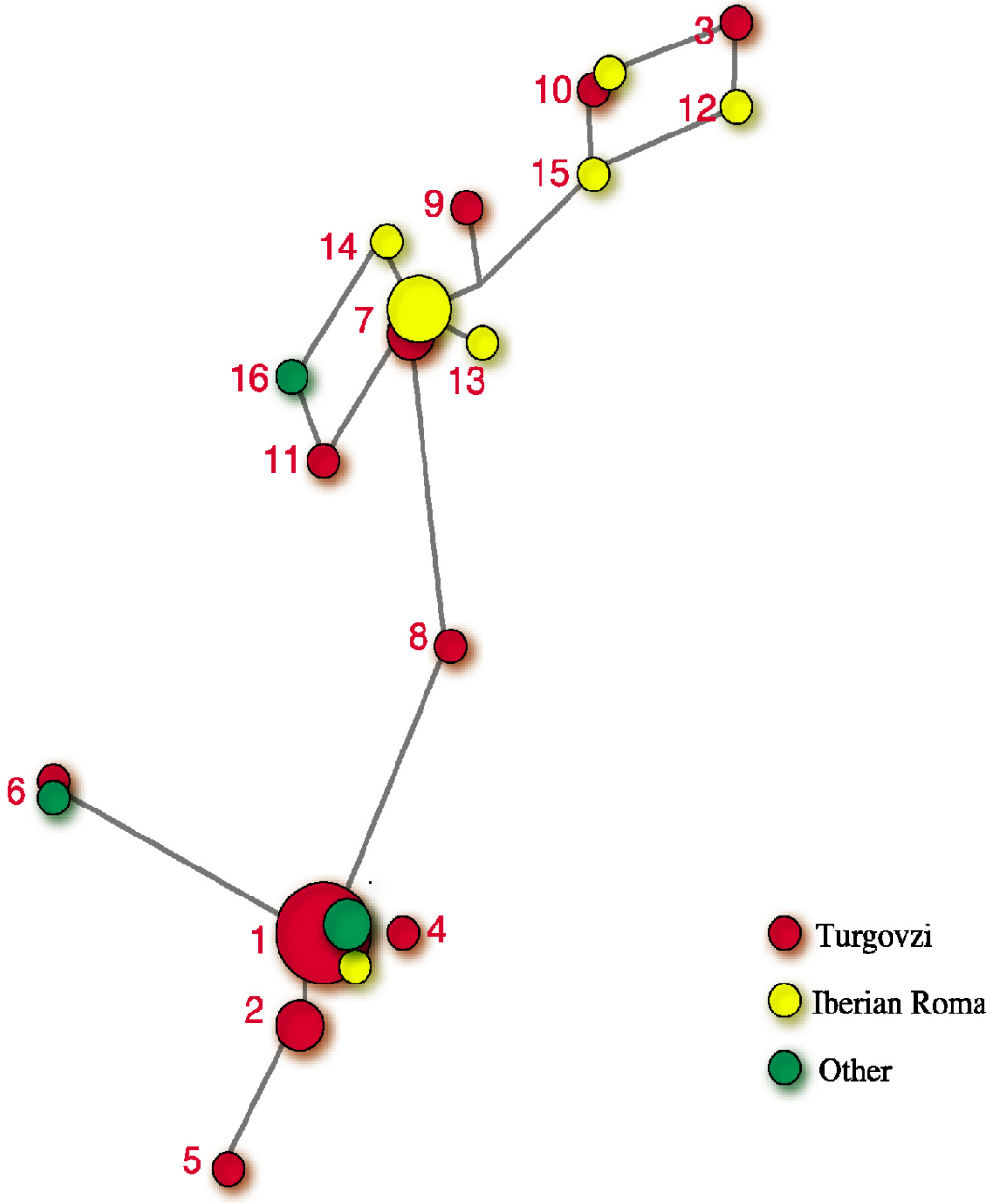

Figure 10-3 Median-joining network of C283Y haplotypes in Romani populations. Branch lengths are proportional to the number of mutations and the number of loci involved in recombination.

10.2.2.2 Age of the C283Y mutation

The overall age of the C283Y mutation in the Roma was calculated, as was the age of the mutation within the Turgovzi and Iberian Roma (table 10-7) using the coalescent method of Stephens et al., (1998). For the C283Y haplotype defined by five microsatellite loci spanning 2.75Mb, the recombination rate was estimated to be 0.0275. Thus $r = 0.0275 + 5(1.2\text{x}10^{-3}) = 0.0335$.

Table 10-7
Age of the C283Y mutation based on the coalescence of haplotypes

Population	$P_{ancestral}$	Generations	Age (20y/gen)
Total	0.459 (haplotype 1)	23	460
Turgovzi	0.604 (haplotype 1)	15	300
Iberian	0.666 (haplotype 7)	12	240

NB The ancestral haplotype in each population is indicated

The age of the C283Y mutation in the Roma was estimated to be 460 years. The mutation age is younger within the Turgovzi than the Iberian Roma with an estimated age difference of 60 years.

The mutation was dated in the Turgovzi based on LD observed at the flanking loci D13S115 and D13S283 (table 10-8). The recombination fraction between D13S115 and the mutation was estimated as 0.007 and between D13S283 and the mutation as 0.013.

Table 10-8
Age of the C283Y mutation in the Turgovzi based on LD

	G_{D3S115}	$G_{D13S283}$	$G_{average}$
Generations	73.6	19.3	46.45
20y/gen	1,472 years	386 years	929 years

Using linkage disequilibrium (Guo & Xiong, 1997; Risch et al., 1995), estimates of the number of generations since the C283Y mutation occurred differ greatly based on the two different loci. An average of the two estimates gives a mutation age of 929 years before present.

10.2.2.3 Linkage disequilibrium around the C283Y mutation

Thirty normal chromosomal haplotypes from the Turgovzi population were constructed from genotyping parents of affected individuals. None of the eleven Turgovzi C283Y haplotypes was found amongst the normal Turgovzi haplotypes. However, haplotype 16, which is homozygous in an affected Italian Roma, was found to occur in the sample of normal chromosomes in the Turgovzi sample.

Allele frequencies were calculated at each locus for normal and affected chromosomes (table 10-9). Linkage disequilibrium at each locus was assessed by calculating the P_{excess} statistic without discriminating between disease-associated alleles that have arisen through recombination from those that have arisen from mutation. The location of the disease causing mutation is correlated with the highest observed value of linkage disequilibrium.

Table 10-9
Chromsome 13q12 microsatellite allele frequencies for disease and normal chromosomes

Locus	D13S115		D13S232		D13S787		D13S292		D13S283	
Allele	C283Y	N	C283Y	N	C283Y	N	C283Y	N	C283Y	N
1				0.033		0.100	0.021	0.367	0.021	
2	0.771	0.433		0.633	0.062	0.467		0.267	0.146	0.690
3	0.167	0.200	1.00	0.233	0.938	0.367	0.979	0.367		
4						0.067				
5	0.042	0.200								
6	0.021								0.021	0.069
7				0.100						
8		0.167								
9										
10									0.792	0.069
11										
12										0.034
13										0.034
14									0.021	0.103
	P_{excess}=0.596		**P_{excess}=1**		**P_{excess}=0.902**		**P_{excess}=0.967**		**P_{excess}=0.776**	

NB The most frequent alleles for each sample are highlighted.

CHAPTER 11

DISCUSSION

11.1 The Distribution of Private Founder Mutations

HMSNL has been reported in Romani individuals from several different populations (Butinar et al., 1999; Colomer et al., 2000; Kalaydjieva et al., 2000; Kalaydjieva et al., 1996; Merlini et al., 1998). Identification of the R148X mutation in *NDRG1* has demonstrated the identical underlying genetic defect in affected individuals (section II). Similarly, LGMD2C has been diagnosed in a number of Romani individuals from disparate populations (Lasa et al., 1998; Merlini et al., 2000; Piccolo et al., 1996). These affected individuals have been shown to be homozygous for the C283Y mutation in *SGCG*. The widespread distributions of these autosomal recessive diseases suggests high disease allele frequencies in the Roma. Carriers of the disease allele are to be expected in populations in which the disease has been identified. In addition, the common genetic heritage of Romani populations (section I) suggests the disease allele may occur in populations with no reported cases of the disorders. Identification of the genome level mutation allows rapid population screening for nonsymptomatic carriers. This estimate provides an indication of the prevalence of the disease mutation and the expected number of affected births.

The R148X mutation in *NDRG1* was identified in six of eight Romani populations. HMSNL had previously been identified in three of these populations; namely the Lom, Kalderash and Spanish Roma. However, the sample of Spanish Roma examined in this study is of unknown relatedness to the Romani individuals included in the HMSNL study. In addition to these populations, the disease allele was identified in the Turgovzi, Kalaidjii, Lingurari and Lithuanian Roma, populations in whom HMSNL has not been reported. Thus, the disease allele is represented in the three major migrational groupings of European Roma and transcends metagroup boundaries in the Bulgarian Roma. The overall carrier rate in the entire Romani sample is 5.4%. However, within-population carrier frequencies are highly variable. The majority of

R148X carriers are found in the Lom population in which a carrier frequency of 19.4% is observed. This carrier rate is dramatically higher than that observed in other populations, which range from 2-5%. The failure to identify the R148X mutation in the Feredjelli and Intreni possibly points to negligible carrier rates in these populations. However, sample sizes were small for both of these populations and should be expanded to confirm this suggestion.

The frequency of the C283Y mutation in *SGCG* was investigated in the same eight Romani populations. The C283Y mutation was identified in just two populations, the Turgovzi and Spanish Roma. LGMD2C-affected individuals have previously been identified in the Spanish Roma (Lasa et al., 1998). The generalised carrier frequency of the C283Y mutation in the Roma is 2.6%; however, this is based on the inclusion of the six populations in which the mutation was not identified. Fourteen of the fifteen C283Y carriers are found in the Turgovzi, which corresponds with a carrier frequency of 6.25% within that population. In the Spanish Roma, the C283Y carrier frequency is 1.6%. The mutation was not identified in a sample of 101 Feredjelli, a Xoroxane Romani population that is co-resident with the Turgovzi. This provides evidence for limited gene flow between these co-resident Romani populations. The C283Y mutation was not identified in 165 individuals from Vlach Romani groups, thus presenting the possibility that the *SGCG* C283Y mutation is absent in the Vlach Roma.

Mutation screening has illustrated the variability of carrier frequencies in Romani populations. This study has determined carrier frequencies almost as high as 20% in one population, whilst in other populations the same disease allele is not found. This variation can be explained by the effect of genetic drift on recessive alleles. This stochastic process can result in high frequencies of a disease allele or conversely, result in the loss of disease alleles. With the genesis of each new Romani population it is possible that some disease alleles are not represented in the nascent populations, or that alleles occurring at very low frequencies are rapidly lost. Alternatively, it is possible that disease alleles are transmitted to the new populations and the same random processes result in inflated gene frequencies. In the French Canadian population it has been shown that a single carrier of a disease allele at the inception of the population some 250 years ago is sufficient to result in present-day carrier frequencies of 5%

(Heyer, 1999). Thus, it is conceivable that in Romani populations a very small number of founders carrying disease causing mutations can explain high carrier frequencies in the extant Romani populations.

High carrier frequencies can be expected to result in a large number of affected births. In populations with high frequencies of particular disease alleles, the practice of endogamy can serve to increase the chance of two carriers forming a union and therefore, the chance of an affected child being born from such a union. For example, if marital partners in the Lom population were selected entirely from within the community, in which the R148X carrier frequency is almost 20%, the expected frequency of HMSNL affected births would be 1 in 100. Similarly, a large number of LGMD2C-affected individuals can be expected in the Turgovzi. These findings represent additional examples of the phenomenon first noted in a Slovakian Romani population where a high incidence of congenital glaucoma (Ferak, Gencik, & Gencikova, 1982) is due to the high frequency of a single disease allele (Plasilova et al., 1999). A high frequency of carriers of deleterious mutations within Romani populations provides justification for the implementation of targeted genetic screening. It is apparent that populations that are at the highest risk for increased frequencies of disease alleles are those in which a large number of affected individuals are found. However, the common genetic heritage of Romani populations and the demonstrated widespread distribution of disease alleles suggest the value of offering carrier testing to Romani individuals from all populations. Whilst the lower frequencies of disease alleles seen in some populations correspond with a reduced population risk, all Romani individuals can be considered as potential carriers of these private mutations.

The high disease allele frequencies observed in Romani populations pose a potential difficulty for gene mapping studies. Gene frequencies play an important role in genetic linkage studies (Terwilliger & Ott, 1994). Commonly, gene frequencies are estimated to be between 0.01 and 0.001. In the Roma, gene frequencies have been found almost an order of magnitude greater than typical estimations. Such high gene frequencies can result in overestimated lod scores (Kruglyak, Daly, & Lander, 1995; Lander & Botstein, 1987). High gene frequencies are also an impediment to searches for large shared segments. This is due to the concomitant reduction in lengths of

conserved ancestral haplotypes with higher gene frequencies (Lander & Botstein, 1987). Therefore, the elevated gene frequencies observed in some Romani populations is an important consideration for the design of gene mapping studies.

11.2 Founder Mutations and Population Histories

The difference in the distribution of the R148X and C283Y mutationss suggests two different scenarios for these disease alleles. Dating of the R148X mutation using a coalescent and linkage disequilibrium method produced estimates of 620 and 1,580 years old respectively. These estimations suggest the existence of the R148X allele in the proto-Roma prior to their entrance into Europe and the process of population fracturing. The occurrence of the R148X mutation in the Vlach, Balkan and Western European populations points to the presence of the mutation at a reasonably high frequency in a common ancestral population. This claim is further supported by the occurrence of HMSNL in other Romani populations in which screening was not performed, such as the Slovenian (Butinar et al., 1999), Italian (Merlini et al., 1998) and Rumanian (Kalaydjieva et al., 2000) Roma. In comparison, in this study the C283Y mutation was only found in two populations; the Turgovzi, a Balkan Romani population and the Spanish Roma, a Western European Romani population. This distribution could be explained by a low frequency of this mutation in a parental population. Dating of the C283Y mutation using the coalescent method of Stephens et al., (1998) produced a mutation age of 460 years before present. An estimation of the allele age based on linkage disequilibrium was determined to be 929 years before present. These estimations are considerably younger than that proposed by Piccolo et al. (1996), who estimated that the C283Y mutation occurred at least 1,200 years BP. Their estimate was based on a small sample of seven disease haplotypes in unrelated consanguineous families from different populations. As demonstrated with the distribution of R148X haplotypes, diversity is predominantly observed between populations. Thus, mutation age estimates that sample a small number of chromosomes from multiple populations can be expected to produce older mutation age estimates than those that sample within a population. So far, the majority of cases of LGMD2C caused by the C283Y mutation have been reported in Western European Roma from Portugal, France, Italy (Piccolo et

al., 1996) and Spain (Lasa et al., 1998). Within Balkan and Vlach groups, the C283Y mutation has only been found in the Turgovzi. This distribuution suggests that the C283Y mutation was more highly represented in populations that migrated into Western Europe than in populations that remained in Eastern Europe. The relatively young age of the mutation and its distribution suggests the mutation may have occurred within Europe after the process of population fission was underway. The high carrier frequency of the C283Y mutation within the Turgovzi can be understood as a localised founder effect similar to that observed for R148X in the Lom population.

The R148X disease haplotype was constructed using 24 polymorphic loci. This provides a highly resolved haplotype in comparison to those typically used (eg. Hollox et al., 2001; Mateu et al., 2001; Morral et al., 1994). Therefore, differentiation between haplotypes will occur much more rapidly. This has clearly occurred in the Roma, as haplotypes appear to be largely restricted to single populations and variation has been generated within populations. Only four of the twenty unique haplotypes are found in more than one population. Shared haplotypes are of interest because they point to a recent common ancestor. The sharing of R148X haplotype M by the Lom, Monteni, Rumanian and Spanish Roma provides evidence of the relatedness of these populations. The Lom, Monteni and Rumanian Roma live in close proximity and possibly share similar histories which might explain identical disease haplotypes occurring in the populations. The presence of R148X haplotype M in the Spanish Roma might indicate a more recent migration from one of these populations. The three other shared haplotypes are found in the Kalderash. The Kalderash practice strict endogamy and have only recently ceased to be nomadic (Marushiakova & Popov, 1997). Therefore, it is interesting to note that the Roma resident in France, which share haplotype D with the Kalderash resident in Bulgaria, identify themselves as Kalderash Roma (L. Kalaydjieva, pers comm). Haplotype sharing between the Kalderash and the Monteni, Italian Roma and German may be indicative of gene flow or more recent population affinities.

The greatest haplotype diversity is observed in the Kalderash in whom ten haplotypes are observed. This is consistent with observations within Bulgaria that the prevalence of HMSNL is highest in the Kalderash (I. Tournev, pers comm). The coalescent age of the mutation of 520 years in the Kalderash indicates that the mutation

has been present within the Kalderash for a considerable time. It is interesting to contrast this population with the Lom. Although only four different haplotypes are found in the Lom, they are more evenly distributed and therefore, haplotype diversity and the age of the mutation are similar to that of the Kalderash. Furthermore, the carrier frequency of HMSNL is extremely high in the Lom (19.4%). This suggests that there was more than one founder in the Lom and some degree of haplotype diversity existed at the point of founding of the population. Therefore, the estimated age of the mutation in the Lom is likely to be an overestimate. This highlights the fallibility of dating methods that do not incorporate allele frequencies into calculations.

HMSNL disease haplotype diversity is extremely low in the Monteni. This is consistent with the investigation of male and female lineages which found the Monteni to have dramatically restricted male-specific genetic diversity and low female-specific genetic diversity (section I). A single predominant disease haplotype is found in the Monteni with very little within-population diversity. This can be explained by the introduction of the mutation into the population very recently. Strict endogamous practices would minimise the opportunity for a different disease allele to enter the population, whilst genetic drift could lead to the predominance of a single disease haplotype.

Population screening has revealed that the C283Y mutation is not widely dispersed amongst Romani populations. Haplotype diversity is greater in the Turgovzi than in the Iberian Roma. A network of disease haplotypes in these two populations illustrates that the Iberian diversity can be considered as a subset of that seen in the Turgovzi. This is consistent with the Iberian Roma representing a population that split from the Turgovzi. Dating of the C283Y allele suggests that the mutation is older in the Turgovzi than in the Iberian Roma, which is consistent with this scenario. It should be noted however, that the dates obtained within these populations appear to be underestimates in light of historical records. It is apparent that the population fissioning process represents genetic bottlenecks, with the emerging haplotype distribution in extant populations is largely a function of stochastic processes. The unique haplotypes found in each population are signatures of diversity generated after the populations have split.

11.3 Fine-Scale Linkage Disequilibrium

Linkage disequilibrium is the nonrandom association between two loci (Kruglyak, 1999). Such associations are often used as evidence for a causally important association between genetic markers and a hypothesised disease locus or to refine the chromosomal location of a gene (Weiss, 1993). Linkage disequilibrium was assessed around the known disease locus for HMSNL. The construction of haplotypes over a 3.3cM region using 27 polymorphic loci allowed an examination of LD variation over short regions of the chromosome. In addition, LD in the HMSNL locus was compared in three Romani populations allowing insights into population-specific phenomena.

LD was assessed using the simple P_{excess} statistic proposed by Hastbacka et al., (1992). This statistic essentially quantifies the difference in disease-associated allele frequencies to the frequency of that same allele in the unaffected population. As the aim of the analysis was to simulate a potential genome scan, all disease associated alleles were included, regardless of whether they differed from the most common allele due to mutation or recombination. Thus, the LD values at each locus do not exclusively reflect haplotype decay due to recombination. In the analysis, linkage disequilibrium was found to be variable across the 3.3cM region. In the whole sample, LD generally increased with proximity to the disease locus. Irregularities were observed in this general trend which can be explained by mutations at loci, and randomly occurring fluctuations in the frequency IBS alleles in the population controls. This deviation from the general trend is exemplified by the extreme reduction in LD observed at the D8S378 locus. Hypermutability of this locus in disease chromosomes from the Kalderash population results in a dramatically reduced association at this proximate locus.

Linkage disequilibrium within each of the three populations displays even greater variation. Within these populations, spurious false positive peaks of linkage disequilibrium are observed. These occur when recombinant disease haplotypes bear identical alleles at a polymorphic locus to the nonrecombinant disease haplotypes. These alleles are IBS, however this method assumes they are IBD. In the Kalderash and Monteni, two such false positives are observed. In the Lom population, the region of maximal linkage disequilibrium extends over 14 loci spaced across approximately 800kb. This is a result of the absence of recombinant chromosomes over this region. As

depicted in the network, recombinations in the Lom are of limited extent which renders the Lom population of limited use in refining the disease locus. Total linkage disequilibrium in all disease haplotypes is only observed over the region including markers 458a13-458b57-369a89. Furthermore, this segment is the only one in which LD maxima in the three populations are coincident. This illustrates the essential inclusion of Roma from disparate populations for linkage disequilibrium mapping and refining candidate regions.

Linkage disequilibrium around mendelian disease alleles is evidently highly variable over short physical distances. However, reasonably high levels of LD are observed over significant portions of the genome. The P_{excess} statistics remain high over a 3.3.cM region around the R148X and a 2.75cM region around the C283Y mutation. This presents the possibility that genome-wide scans for allelic associations with a particular phenotype might be possible in the Roma. Within the Rom, well defined phenotypes of monogenic traits can reasonably be hypothesised to result from founder mutations that can remain in strong linkage disequilibrium with nearby loci. The strongest allelic associations can be expected in populations in which genetic diversity is low and the mutation is young. This approach is similar to the search for shared genomic segments; however, in this case association is sought at a single locus rather than haplotypes constructed from two loci. Further studies are required to examine the extent and strength of allelic associations around disease loci in the Roma. However, these results suggest that controls could be selected from unaffected individuals in the same population rather than relying on pedigree samples and data exclusively.

It is possible that the spurious drops in LD observed in the HMSNL region would be overcome by the use of SNPs in association mapping. Much of the irregularity of LD in the HMSNL region is due to microsatellite mutation. As SNPs are considered to be unique events, every new SNP that was encountered would provide evidence of a recombination. This would ameliorate the apparent absence of allelic association due to locus mutation. However, the relative uninformativeness of SNPs could possibly undermine this benefit.

SECTION IV

PILOT STUDY OF COMMUNITY BASED CARRIER TESTING IN THE ROMA

CHAPTER 12

SUBJECTS AND METHODS

12.1 Introduction and Study Design

12.1.1 Background to the Study

Elucidation of the molecular genetic basis of disorders provides a powerful means for disease diagnosis and predictive testing. By disentangling the genetic defects that underlie disorders with similar clinical presentations, robust and definitive assays can be developed. Gene defects that result in inborn errors of metabolism are most amenable to treatment. However, for the majority of monogenic disorders, treatment strategies following molecular diagnosis are hindered by a lack of knowledge about gene function. In such situations the application of genetic medicine can only entail predictive testing. In nonsymptomatic prospective parents this can take the form of carrier testing for deleterious genes.

Carrier testing has been widely employed for a number of genetic disorders. The impetus for such predictive testing is usually supplied by a disease history in the family. However, for some genetic disorders the population prevalence of a particular disorder is deemed to be high enough to warrant widespread screening. Such is the case for the ΔF508 mutation in *CFTR* in the Caucasian population. In population isolates, limited genetic diversity often results in reduced genetic heterogeneity of inherited disorders. In addition, founder effect can result in single mutations occurring at high frequencies. As a result, population-targeted carrier testing enjoys a greater efficiency in these populations. This scenario has been exploited in genetically isolated populations such as the Ashkenazi Jews and the Finns. In the Roma, a number of disease alleles have been found to occur at high frequencies (Kalaydjieva, Gresham & Calafell, 2001; Kalaydjieva et al., 1999; Plasilova et al., 1999). This suggests that targeted genetic screening may be fruitful in these populations.

In 1996, the C283Y mutation in *SGCG* was identified as the cause of limb girdle muscular dystrophy type 2C, a severe autosomal recessive form of early onset muscular dystrophy (Piccolo et al., 1996). The affected individuals, resident in Spain, Italy and France, were all of Romani ethnicity. Analysis of the disease haplotype through genotyping of closely associated microsatellite markers suggested a founder mutation (Piccolo et al., 1996). This report was followed by the identification of additional affected Roma in Germany (M. Jeanpierre, pers comm) and Portugal (Lasa et al., 1998). It was hypothesised that the same mutation might be found in Romani communities in the Balkans. After an extensive search of Bulgarian hospital records and fieldwork in Romani neighbourhoods by Dr Ivailo Tournev of the Sofia Medical School, thirty-two Limb Girdle Muscular Dystrophy patients were identified. The thirty-two affected individuals were found to belong to a total of nineteen unrelated families. All were Xoroxane Roma, who are Muslims and descendants of early migrants into the Balkans, and most resided in the northeastern part of Bulgaria.

Given that the Roma are known to adhere to strict endogamous practices and exhibit reduced genetic diversity (section I), the identification of a large number of individuals affected by the same genetic disorder suggested an increased frequency of the disease allele. Therefore, a pilot genetic screening project was initiated in the Northeast Bulgarian town of Omurtag during 1998. Carrier testing in the community was considered feasible, based on an urban population of some three to four thousand Romani inhabitants. Moreover, community members were deemed to be in a reasonable economic situation and well informed about the disease, and they demonstrated receptivity to the suggested study.

12.1.2 Research Questions

This study aimed at implementing a pilot community genetics program in a Romani community deemed to be at high-risk for a monogenic disorder, LGMD2C, and assessing factors that might affect the uptake of carrier testing. Specifically, it aimed at answering the following questions:

1. What is the genetic basis of LGMD2C in Xoroxane Romani families?

2. What are the attitudes towards genetic disease and its prevention held by members of the Xoroxane community?
3. What psychological and social factors must be addressed when implementing a community-based genetic program in a Romani community?
4. Is community-wide screening a suitable approach to carrier testing for disease alleles in Romani populations?

12.1.3 Subjects

12.1.3.1 Genetic counselling and testing in affected families

Prior to the offer of genetic testing, LGMD2C patients and members of their families were provided with relevant information about the disease. This included information on the mechanisms of inheritance of the disease, its clinical course, the prospects of treatment for affected individuals, the availability of carrier testing for unaffected relatives, and reproductive options. The consulting physicians of the field team provided the information in an informal setting, during a visit to the family's home.

Thirty-two affected individuals were clinically diagnosed as having LGMD2C. The affected individuals came from 19 different families. These families were resident in northeastern Bulgaria, in the communities of Omurtag and neighbouring villages and in a small area south of the Balkan Mountains. Unaffected family members from 16 families requested the carrier test. Blood samples were collected from a total of 157 unaffected family members on FTA Genecards. Families and individuals were assigned an identification number to ensure anonymity. A record of the name of the individual referred to by the identification number was accessible only to Associate Professor Kalaydjieva and Dr Tournev.

Subsequent to the mutation assay, results were provided to the Bulgarian physicians who were members of the field team. Individuals identified as being non-carriers were informed of their status in writing. All identified carriers of the mutation were informed personally and in writing, and provided with post-test counselling.

12.1.3.2 Collection of samples for community genetic screening

A field team under the leadership of Associate Professor Kalaydjieva and Dr Tournev collected the samples. Blood samples were obtained by a finger prick using a sterile lance. The blood samples were applied to individual FTA Genecards on which name and date of birth of the individual was recorded. All blood samples were forwarded to myself.

Community members were provided with educational information regarding the disease, the mechanisms of inheritance and reproductive options by members of the field team. Testing for the C283Y mutation was offered to members of the community subsequent to and independent of the administration of a questionnaire investigating attitudes to genetic disease and prevention. Individuals were informed of the procedure for taking a blood sample, the time required for the testing to be completed and a tentative timetable for the return of the results and the availability of counselling for identified carriers. If the individual desired the test, a blood sample was requested. A total of 325 members of the community requested the carrier test. They included 71 individuals who had completed the questionnaire.

12.2 Methods

12.2.1 Laboratory Methods

Processing of FTA Genecards was performed as described in section 3.2.1. The C283Y mutation assay was performed using the R*sa*1 restriction endonuclease as described in section 9.2.1.

12.2.2 Questionnaire Design and Analysis

12.2.2.1 Construction of the questionnaire

The questionnaire was designed by Associate Professor Luba Kalaydjieva of the Centre for Human Genetics, Edith Cowan University and Professor Assen Jablensky from the Department of Psychiatry and Behavioural Sciences, The University of

Western Australia. The questionnaire was written in Bulgarian for the purpose of distribution and translated into English for analysis. Throughout, a numerical coding system was used to facilitate recording and collation of data. The questionnaire was composed of four separate sections, each of which was designed to investigate different social and attitudinal aspects of the high-risk community, relevant to genetic disease and predictive testing.

Part I was designed to collect information regarding demographic factors, including age and marital status of the individual, employment and household structure. Investigations of reproductive history included questions on number of children (living and deceased). This section also inquired as to whether the individual wished to be tested for the C283Y mutation.

Part II aimed at obtaining ethnological information regarding language, faith and marriage contract practices.

Part III of the questionnaire aimed to quantify an estimate of the education level of the individual. This was achieved by means of a fifteen-word vocabulary test. Individuals were visually and orally informed of the word and asked to provide a definition. They were then scored as knowing the correct definition, providing a partial answer or clearly not knowing the meaning of the word. The interviewer recorded a numerically coded score.

Part IV of the questionnaire investigated the occurrence of Limb Girdle Muscular Dystrophy Type 2C in the interviewee's immediate and extended family, and asked questions designed to investigate attitudes towards prevention of the disease. In the first section of Part IV, the individual was asked a series of five questions that aimed at ascertaining whether any family members had exhibited symptoms associated with the disease. The second subsection of Part IV aimed at determining relevant attitudes towards predictive genetic testing. Three hypothetical scenarios were presented to the interviewee. These stories were read to the interviewee and followed by questions that pertained to the described scenario. These questions addressed either the reasons behind the actions of the characters in the scenario or the possible outcomes of these actions. For each question that was posed, a series of responses were provided from which the interviewee was asked to choose the one that best described their opinion.

12.2.2.2 Distribution and completion of questionnaires

The questionnaire was presented to members of the high risk Romani community of Omurtag in Northeast Bulgaria during 1998. Questions were posed by means of an oral interview conducted by Dr Ivailo Tournev and members of the field team. The questionnaire was presented prior to and independent of the testing for the C283Y mutation. A total of 102 individuals completed the questionnaire. Testing for the C283Y mutation was offered to all individuals, however, only 71 requested the test.

Answers to questions were recorded on the questionnaire using the numerical coding system. Additional information was recorded on the questionnaire paper by the interviewer in Bulgarian and translated into English by Associate Professor Kalaydjieva and Dr Angelicheva of the Centre for Human Genetics, Edith Cowan University.

Subsequent to the completion of all questionnaires, the original documents were forwarded to myself for processing and analysis.

12.2.2.3 Analysis of questionnaire

Each individual who had completed a questionnaire was assigned an identification number. The suffix to this number indicated gender and marital status (ie. FM = female married, FS = female single etc.). Data and numerically coded answers from the questionnaires were entered into a database created using Microsoft Excel. The names of individuals were not included in any of the analyses. Answers to questions were tabulated using Microsoft Excel. Data analyses were performed using Microsoft Excel and SPSS v10.0.

Nonrandom statistical relationships between answers provided in the questionnaire and different parameters were investigated by means of chi-squared tests, with P-values less than 0.05 considered significant.

CHAPTER 13

RESULTS

13.1 Results of the C283Y Detection Assay

DNA samples were investigated for the C283Y mutation using the R*sa*1 digest. This assay clearly determines the genotypic state at the disease-causing mutation (figure 13-1).

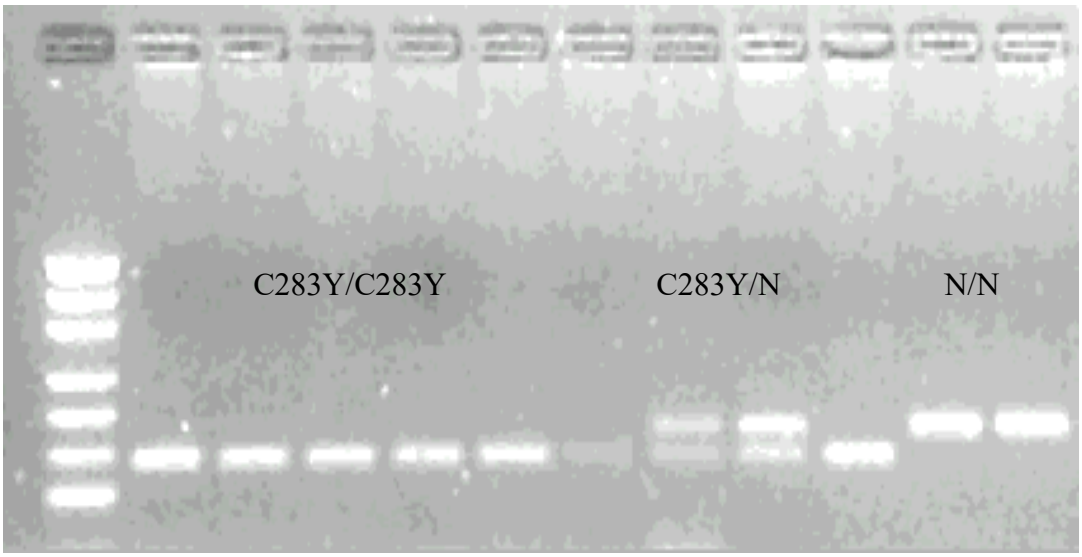

Figure 13-1 Agarose gel containing R*sa*1 digest products from the C283Y assay

13.2 C283Y Status of Affected Individuals

LGMD2C-affected individuals came from 19 different families (Table 13-1). Of the thirty-two clinically diagnosed LGMD2C, thirty were homozygous for the C283Y mutation. Family 4 was found to have one affected individual homozygous for the C283Y mutation and one affected individual heterozygous for the mutation. The affected individual from Family 17 was homozygous for the wild-type state at codon 283 of *SGCG*.

Table 13-1
Number of affected individuals in each family and their C283Y status. [1]C283Y/C283Y indicates homozygosity for the disease causing A→G mutation in codon 283 of the SGCG, [2]C283Y/N indicates heterozygosity for the mutation, [3]N/N indicates the wildtype state at codon 283.

Family	No. of Affected Individuals	C283Y/C283Y[1] Patients	C283Y/N[2] Patients	N/N[3] Patients
1	4	4	0	0
4	2	1	1	0
6	1	1	0	0
7	1	1	0	0
8	1	1	0	0
9	1	1	0	0
10	1	1	0	0
11	2	2	0	0
13	1	1	0	0
14	1	1	0	0
15	2	2	0	0
16	1	1	0	0
17	1	0	0	1
18	1	1	0	0
19	1	1	0	0
20	2	2	0	0
21	3	3	0	0
22	6	6	0	0
23	0	0	0	0
Total	32	30	1	1

13.3 Carrier Testing of Family Members of Affected Individuals

Individuals from sixteen families (table 13-2) requested carrier testing.

Table 13-2
Family members of individuals affected by LGMD2C who requested the test for the C283Y mutation

Family	Number of unaffected family members who requested test	Non-carriers	C283Y carriers	Non-obligate carriers of the C283Y mutation
1	44	23	21	15
4	9	5	4	2
8	1	0	1	1
9	15	7	8	8
10	3	1	2	1
11	8	3	5	3
13	9	3	6	4
14	3	2	1	1
15	4	1	3	2
16	15	8	7	5
17	4	4	0	0
18	3	0	3	2
19	2	1	1	0
20	3	0	3	1
21	11	5	6	2
22	21	13	8	2
23	2	1	1	1
Total	157	77	80	50

The average age of the family members was twenty-eight years and the median age was twenty-four years. Fifty-one percent (80/157) of family members tested were found to be carriers of the C283Y mutation. The other seventy-seven individuals, all of whom were related to affected individuals, were identified as being homozygous normal at codon 283. Of the eighty carriers, thirty were parents of affected individuals. Therefore, thirty-two percent (50/157) of family members were non-obligate carriers. The average age of non-obligate carriers was twenty-nine years and the median age was twenty-three years. Of the fifty non-obligate carriers, twenty-two were already married and fifteen of these reported having at least one unaffected child. Thus, a total of twenty-eight non-obligate family members, of an average age of nineteen years, were identified as carriers prior to marriage. No non-obligate high-risk couples were identified.

13.4 Pilot Public Health Genetics Program

13.4.1 Results of Questionnaire Investigating Knowledge of Disease and Social, Cultural and Attitudinal Factors Relevant to Community-based Genetic Screening

13.4.1.1 Introduction

The questionnaire was administered to a total of 102 people from the Xoroxane Romani population. For the large part, the questionnaire was answered by married individuals of childbearing age. Fifty-two of the individuals were female, of whom three were unmarried and forty-nine were married. Fifty males were interviewed, six were unmarried and forty-four were married.

For the purposes of analyses, two questionnaires, one filled by a female and one by her husband, were excluded due to the incomplete and confused nature of the answers provided. Thus, one hundred questionnaires were analysed. The answers were coded and entered into a Microsoft Excel spreadsheet.

13.4.1.2 Demographic data

Questionnaire respondents were an average of 24 years old (table 13-3). Households contained an average of more than five people and almost two children.

Table 13-3
Summary of demographic data for questionnaire participants

	Maximum	Minimum	Average
Average age N=91	46 years	11 years	24 years
No. of children in family N=91	7	0	1.9
No. of people in household N=90	11	2	5.4

When household composition was queried the data indicated that 21% contained the couple and their children, 68% contained the couple, their children and one of the spouses' parents, and 11% contained additional members of extended family.

Twenty-four individuals, including males and females, reported having experienced a miscarriage or stillbirth. Eleven of these twenty-four reported more than one miscarriage or stillbirth. Two individuals stated that they had a child now deceased.

People were asked how many members of the household had held paying jobs during the last year. Generally, this question was not answered. In those few cases in which an answer was provided, it was stated that members of the household participated in the buying and selling of goods as a primary source of income.

13.4.1.3 Ethnographic information and investigation of marriage practices

All respondents stated that *Tsigane* (Romany) was the preferred language of use in the home. In addition, 90% of people stated that Turkish was also spoken in the house. As the questionnaire was administered in Bulgarian, it is reasonable to state that at least 90% of respondents were trilingual. However, no individual reported Bulgarian as the preferred language of use. When asked which faith they adhered to, all respondents declared themselves to be Muslim.

Four questions were posed regarding marriage norms in the community (table 13-4).

Table 13-4

Questions regarding marriage practices in the Xoroxane Roma community

Question and Responses	Percentage of responses
Q2.1 When a marriage is contracted, what is the role of the man's parents? (%)	
R1. In most cases they identify the right girl	48.5
R2. In most cases it is the man's choice and approval is sought from the parents	45.4
R3. In most cases it is the man's choice and parental approval is not necessary	6.2
R4. None of the above	0.0
Q2.2 When a marriage is contracted, what is the role of the girl's parents? (%)	
R1. In most cases they identify the right man	3.1
R2. In most cases it is the girl's choice and approval is sought from the parents	22.7
R3. In most cases it is the girl's choice and parental approval is not necessary	7.2
R4. None of the above	67.0
Q2.3 When a marriage is contracted, where does the man usually come from? (%)	
R1. From the same neighbourhood and always from the same group	70.3
R2. Always from our group but not necessarily from the same neighbourhood	26.6
R3. Not always from our group	3.1
R4. None of the above	0.0
Q2.4 When a marriage is contracted, where does the girl usually come from? (%)	
R1. From the same neighbourhood an always from the same group	67.2
R2. Always from our group but not necessarily from the same neighbourhood	29.5
R3. Not always from our group	3.3
R4. None of the above	0.0

Responses indicated that the role of the man's parents in marriage contracts is to identify the right girl for their son (48.5%), or to give their approval or disapproval of the man's choice of potential wife (45.4%). When the same question was asked regarding the role of the girl's parents, 22.7% of responses stated that the girl chooses the man and seeks parental approval. The majority of respondents (67%) stated that none of the choices provided was correct.

In marriage contracts, 96.6% of responses indicated that the man always comes from the same group. Seventy per cent of interviewees stated that the man always comes from the same neighbourhood and group and 26.6% said that the man always comes from the same group but may come from a different neighbourhood. Similarly, 96.7% responses indicated that in marriage contracts the female invariably comes from the same group with 67.2% stating that the female always originates from the same neighbourhood and group and 29.5% stating that the female always comes from the same group but may come from a different neighbourhood.

13.4.1.4 Interviewee's knowledge of disease within family

Knowledge of a family history of the disease was investigated by asking interviewees whether they were aware of any members of their immediate and extended family that has/had exhibited any of the symptoms associated with LGMD2C. Five questions regarding common manifestations or outcomes of the disease were asked (table 13-5).

Table 13-5
Answers to questions investigating knowledge of the disease in individual's family

Questions	No. of Response		
	0	1	2
Q3.1 Do you know a child that, at age 5-6 years, started to complain of easy fatigue, stopped running and playing, had difficulty in climbing stairs and in getting up from a kneeling position?	76	7	13
Q3.2 Do you know a child who, at age 10-15 years, became unable to walk?	76	7	13
Q3.3 Do you know young men/women who died at about 20 years of age and were unable to walk before that?	95	0	1
Q3.4 Do you know a child with deformity of the spine who has breathing difficulty and frequent chest infections?	96	0	0
Q3.5 Do you know a child and/or young man/woman who has heart problems?	95	0	1

NB. 0=no affected relatives, 1=yes, in the immediate family (siblings, own children), 2=yes, among more distant relatives (cousins, nephews and nieces, uncles, aunts).

Seven people (7.3%) indicated that they possibly have an immediate family member affected by LGMD2C by answering positively to questions 1 and 2. In addition to these seven people, thirteen interviewees indicated that they had a member of their extended family who exhibited possible symptoms of LGMD2C (indicated in question 3-1). A single individual reported that they had a distant relative with known heart problems. However, this individual did not report that the relative showed any of the other symptoms of LGMD2C. Of the 21 individuals who demonstrated possible awareness of a family member exhibiting possible signs of the disease, 14 (67%) requested to have the test and 7 (33%) declined the test.

13.4.1.5 Attitudes towards disease and predictive testing

Attitudes towards genetic disease and prevention-associated issues were investigated using three anecdotal scenarios. These scenarios were presented to the interviewees and they were asked to answer questions referring to these scenarios.

For the purposes of analysis, the sample was divided on the basis of a number of criteria (table 13-6). Chi-squared 2x2 contingency tests were used to identify nonrandom relationships between responses and the various categorical criteria.

Table 13-6
Criteria used to sub-divide the respondents to the questionnaire

Criteria	Categories	Sample size (N_{total}=100)
Sex	Male	49
	Female	51
Knowledge of an affected family member	Knowledge	21
	No Knowledge	79
Requested Carrier Test	Requested	69
	Not Requested	31
Children	Have children	76
	No children	24
Age*	Above average	41
	Below average	52
Vocabulary test	Above average	46
	Below average	54

*Ages were not provided by all individuals.

The first scenario was followed by Question 1:

> Ivo and Sevda have a sick child who from age 6 has difficulty walking, has heart problems and chest deformity. Sevda is pregnant again and her doctor recommends an investigation of the unborn baby. However the family decided to go ahead with the pregnancy without any testing. Why do you think they made this choice?
>
> Three possible responses were provided (table 13-7).

Table 13-7
Answers to scenario 1 investigating attitudes towards predictive genetic testing

Category	Response[§]		
	1	**2**	**3**
Male	67.4	2.2	30.4
Female	81.6	0.0	18.4
Knowledge of affected family member	94.7	0.0	5.3
No knowledge of affected family member	69.7	1.3	28.9
Requested carrier test	75.0	0.0	22.6
Declined carrier test	74.2	3.2	22.6
Have children	82.2	1.4	16.4
Do not have children	50.0	0.0	50.0
Above average age of respondents	77.5	2.50	20.0
Below average age of respondents	68.0	2.00	30.0
Above average vocabulary score	76.5	2.0	21.6
Below average vocabulary score	72.7	0.0	27.3
Total	74.7	1.1	24.2

Note. All answers are in percentage values. Significant differences are highlighted. [§]*Response 1=Because it is God's will and it would be sinful to interfere. Response 2=Because one shouldn't trust medicine completely. Response 3=Because it is impossible to know whether an unborn child will have the disease.*

The majority of respondents (74.7%) selected answer 1, "because it is God's will and it would be sinful to interfere". Significantly more individuals who had already produced a child chose this answer (82.2%) than those who did not have a child (50%), $\chi^2(1, N=94) = 10.131$, $p=0.01$. Furthermore, significantly more people who know of an affected child in their family chose answer 1 (94.7%) compared to those who do not know of an affected child in their family (69.7%), $\chi^2(1, N=94) = 4.752$, $p=0.029$.

Scenario 2 presented the individual with the following hypothetical situation:

> Ivo and Sevda are newly married and don't have children yet. Sevda has a cousin with the disease but there are no sick people in Ivo's family. Sevda wants both she and Ivo to have a test before deciding to have children. Sevda talks to her mother-in-law about this. The

mother-in-law decides they do not need the test because nobody in Ivo's family is affected.

The interviewees were then asked three questions:

- Question 2.1. Would it have been right for Ivo and Sevda to discuss the matter with Sevda's parents and Ivo's father and decide as a group?
- Question 2.2. Would it have been a good idea if Ivo and Sevda did not discuss the matter with anyone and made their own decision?
- Question 2.3. What would happen if the characters disobeyed the mother-in-law's decision?

The answers to these questions were tabulated (table 13-8)

Table 13-8
Answers to questions investigating decision making about reproductive issues

	Question 2.1*			Question 2.2*			Question 2.3^		
Answer	1	2	3	1	2	3	1	2	3
Male	75.0	10.8	4.2	28.6	67.3	4.1	18.4	34.7	46.9
Female	76.5	19.6	3.9	13.7	82.4	3.9	18.0	40.0	42.0
Knowledge of affected family member	76.2	23.8	0	9.5	90.5	0	19.0	42.9	38.1
No knowledge of affected family member	75.6	19.2	5.1	24.1	70.9	5.1	17.9	35.9	46.2
Requested carrier test	82.6	13.0	4.3	18.8	76.8	4.3	16.2	35.3	48.5
Declined carrier test	60.0	36.7	3.3	25.8	71.0	3.2	22.6	41.9	35.5
Have children	76.3	21.1	2.6	15.8	80.3	3.9	16.0	38.7	45.3
Do not have children	73.9	17.4	8.7	37.5	58.3	4.2	25.0	33.3	41.7
Above average age of respondents	75.6	22.0	2.4	19.5	78.0	2.4	14.6	41.5	43.9
Below average age of respondents	76.5	19.6	3.9	23.1	71.2	5.8	19.6	31.4	49.0
Above average performance on test	81.5	13.0	3.7	24.1	70.4	5.6	16.7	25.9	57.4
Below average performance on test	67.4	28.3	4.3	17.4	80.4	2.2	20.0	51.1	28.9
Total	75.8	20.2	4.0	21.0	75.0	4.0	18.2	37.4	44.4

Note. All answers are in percentage values. Significant differences are highlighted. **For questions 2.1 and 2.2: Answer 1 = yes, Answer 2 = no, Answer 3 = not sure.* *^For question 2.3: Answer 1 = nothing will happen, the mother-in-law will accept their choice, Answer 2 = They will be on bad terms after that, Answer 3 = I don't know, all families are different.*

Of the 99 responses provided for question 2.1, 75.8% agreed that it would have been right for the couple to discuss the matter with their parents. A significantly greater

proportion of individuals who chose to have the carrier test answered that it would be right to discuss the matter with the female's parents and male's father (82.0%) than those individuals who declined the carrier test (60.0%) χ^2 (1, N=95) = 7.155, p=0.007.

Seventy five per cent of respondents replied that it would not be a good idea for the couple to make their own decision without consulting anyone. Over twice the proportion of males (28.6%) than females (13.7%) responded that it would be permitted not to consult anyone about making the decision, however, the difference was not significant. Significantly more people who had produced a child (80.3%) believed that to proceed with the test without discussing the matter with the parents would be a bad idea compared with those who did not have children (58.3%), χ2 (1, N=96) = 5.270, p=0.022.

Question 2.3 inquired about the likely outcome of disobeying the mother-in-law's decision and going ahead with a genetic test. Thirty-seven percent said that they would be on bad terms with the mother-in-law if they took such a course of action. However, the majority of respondents (45%) said that they could not know because all families are different. Individuals who performed below average on the vocabulary test were more likely to say that they would be on bad terms after this (51.1%) than did those who scored above the average (25.9%) χ^2 (2, N=99) =8.807 p=0.012.

A third scenario was presented to individuals on which a single question was asked. Four answers (A3.1, A3.2, A3.3, A3.4) were provided for this question and the individual was asked to respond as to whether they agreed, disagreed or were unsure with each answer. The scenario provided was:

> Tinka is not married yet but she often thinks of the children that she would like to have some day. Tinka has a girlfriend and the friend's brother is affected with the disease. Tinka thinks that it is a terrible disease and is worried by the idea of her own children being affected. When Tinka was offered the test, however, she declined to have it.
> *Why do you think she declined?*

Table 13-9 provides the four answers and the distribution of responses to these answers.

The majority of respondents (86%) disagreed with the answer that the character would decline the test because she doesn't believe a doctor could determine if she was a

carrier for the disease. Similarly, most (78%) disagreed that she would decline the test because she would think that there is nothing that could be done if she were a carrier. Over two-thirds of respondents agreed that if she turned out to be a carrier no-one would marry her and that this would be a reason to decline the test. Significantly more people who had a child (75.0%) than those without children (50.0%) agreed that "if she turned out to be a carrier she will never get married" and that this was a reason not to have the test, χ^2 (1, N=97) = 3.813, p=0.050.

Of the 100 responses, 58% agreed that she would decline a test because "if she turns out to be a carrier, the whole family would lose face" and that this would be a reason to decline the test. Significantly more people (63.3%) who did not have a family member with symptoms of the disease agreed that "the whole family would lose face if she turned out to be a carrier" when compared with those who did have an affected family member (38.1%), χ^2 (1, N=97) = 5.249, p=0.022.

Table 13-9

Reasons an individual might decline a genetic test

	No	Yes	Unsure
A1. Because she doesn't believe that doctors can find out whether she is a carrier (%)			
Male	81.6	16.3	2.0
Female	90.2	7.8	2.0
Knowledge of affected family member	95.2	4.8	0.0
No knowledge of affected family member	83.5	13.9	2.5
Requested carrier test	89.9	8.7	1.4
Declined carrier test	77.4	19.4	3.2
Have children	88.2	10.5	1.3
Do not have children	79.2	16.7	4.2
Above average age of respondents	85.4	14.6	0.0
Below average age of respondents	84.6	11.5	3.8
Above average performance on test	85.2	14.8	0.0
Below average performance on test	87.0	8.7	4.3
Total	86.0	12.0	0.0
A2. Because there is nothing she can do even if she is found to be a carrier (%)			
Male	75.5	18.4	6.1
Female	80.4	19.6	0
Knowledge of an affected family member	85.7	9.5	4.8
No knowledge of an affected family member	75.9	21.5	2.5
Requested carrier test	81.2	15.9	2.9
Declined carrier test	71.0	25.8	3.2
Have children	78.9	19.7	1.3
Do not have children	75.0	16.7	8.3
Above average age of respondents	78.0	19.5	2.4
Below average age of respondents	78.8	17.3	3.8
Above average performance on test	79.6	18.5	1.9
Below average performance on test	76.1	19.6	4.3
Total	78.0	19.0	3.0
A3. Because if she turns out to be carrier she will never get married (nobody would want to marry her) (%)			
Male	34.7	59.2	6.1
Female	21.6	78.4	0.0
Knowledge of an affected family member	31.6	65.8	2.5
No knowledge of an affected family member	14.3	81.0	4.8
Requested carrier test	24.6	71.0	4.3
Declined carrier test	35.5	64.5	0.0
Have children	23.7	75.0	1.3
Do not have children	41.7	50.0	8.3
Above average age of respondents	19.5	78.0	2.4
Below average age of respondents	36.5	59.6	3.8
Above average performance on test	29.6	66.7	3.7
Below average performance on test	26.1	71.7	2.2
Total	28.0	69.0	3.0
A4. Because if she turns out to be a carrier, the whole family will lose face (%)			
Male	44.9	51.0	4.1
Female	33.3	64.7	2.0
Knowledge of an affected family member	32.9	63.3	3.8
No knowledge of an affected family member	61.9	38.1	0.0
Requested carrier test	37.7	59.4	2.9
Declined carrier test	41.9	54.8	3.2
Have children	36.8	61.8	1.3
Do not have children	45.8	45.8	8.3
Above average age of respondents	43.9	56.1	0.0
Below average age of respondents	36.5	57.7	1.9
Above average performance on test	44.4	53.7	1.9
Below average performance on test	32.6	63.0	4.3
Total	39.0	58.0	1.0

NB Significant differences are highlighted.

The responses to the four answers to scenario 4.3, in which respondents were asked to state whether they disagreed with or agreed with the answer given, can be grouped into two categories. Answers 1 and 2 posit mistrust in the ability of medicine to act in an adequate manner as a reason to decline the genetic test. Answers 3 and 4 address the issue of fear of stigmatisation of carriers as a reason not to have the genetic test. Individuals were scored for each of these categories depending on their answers to the two questions. These answers were then considered as strongly agreed with the concern (a response of "yes" to both answers), agreed with the concern (response of "yes" to one answer and "no" to one answer) and did not agree with the concern (a response of "no" to both answers). An "unsure" response to either question was considered as being unsure about the issue.

A total of 23% of responses indicated a lack of faith in medicine's ability to detect carriers or aid them is reason to decline the test (figure 13-2). However, only 4% of responses strongly agree with this concern and believe that it is reason to not have the test. The majority of responses (69%) indicated that they disagree with the answers that doctors are unable to identify carriers or act if one is identified as a carrier. The proportion of answers provided by males and females was virtually identical and answers were not significantly related to any demographic factors.

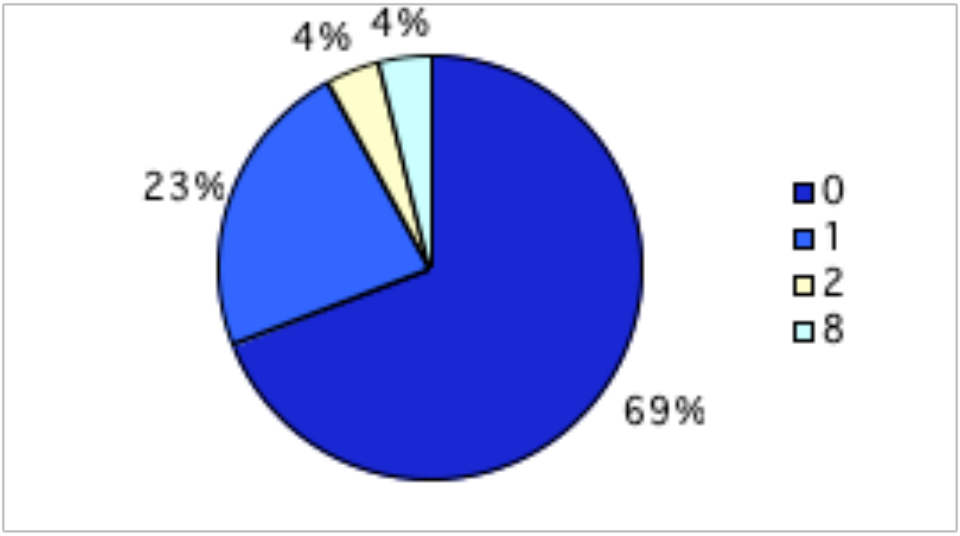

Figure 13-2 Answers indicating that distrust of medicine would be a reason to decline a carrier test. Note: 0 indicates that the individual disagrees with the answers that medicine is unable to act or provide adequate information, 1 indicates that the individual thinks that medicine is unable to help, 2 indicates that the individual strongly believes that medicine is unable to help, 8 indicates that the individual is unsure about the issue.

A total of 47% of responses strongly agree that stigmatisation would result from being identified as a carrier and that this would be a reason to decline the test (figure 13-3). A further thirty-two percent of responses agree that the possibility of stigmatisation would be a reason to decline the test. Only 17% of individuals disagree that stigmatisation is a possible outcome and therefore not a reason to decline the test. Twenty percent more females (57%) than males (37%) indicated that they strongly agreed that the possibility of stigmatisation was suitable reason not to have the test. However, there were no significant relationships between demographic parameters and responses to these answers.

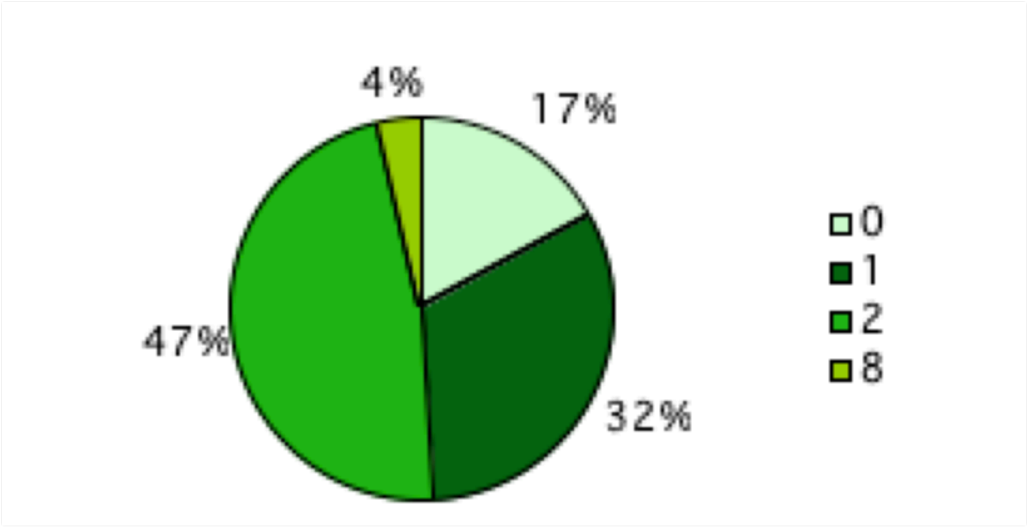

Figure 13-3 Responses indicating fear that positive carrier status may result in stigmatisation. Note: 0 indicates that the individual does not believe that being identified as a carrier will result in stigmatisation and therefore is not a reason to decline the test, 1 indicates that the individual does consider stigmatisation a possibility and therefore reason to decline the test, 2 indicates that the individual strongly believes that being identified as a carrier will result in stigmatisation and therefore the character should not have the test, 8 indicates that the individual is unsure about the issue.

13.4.2 Screening for C283Y Carriers in a High-Risk Community

13.4.2.1 Uptake of genetic test

Carrier testing for the C283Y mutation was requested by 325 members of the high-risk community following the provision of information by the Bulgarian field team regarding the disease and the availability of carrier testing. All individuals originated from a community with LGMD2C patients. Individuals were categorised on the basis of sex and marital status. The participants ranged in age from 8 years to 46 years and were an average 19 years old at the time of the test (table 13-10). Children were tested only when parents requested testing.

Table 13-10
Community members who requested test for C283Y mutation

	Number	Oldest	Youngest	Average Age (years)
Married Females	88	40	15	23
Unmarried Females	75	25	9	14
Married Males	78	46	16	23
Unmarried Males	84	23	8	14
Total	325	46	8	19

13.4.2.2 Carrier test results

Carrier testing for the C283Y mutation was performed using the R*sa*1 assay. Fourteen carriers for the C283Y mutation were identified among the 325 community members. Tests were replicated for these individuals to confirm results. These fourteen individuals comprised four married males and six married females, as well as two unmarried males and two unmarried females. The age of identified carriers ranged from 12 to 30 years old with an average age of 21 years. Seven of the identified carriers had completed the questionnaire. Three carriers of the C283Y mutation had indicated in the questionnaire that they have a family member with possible symptoms of the disorder.

All carriers identified in the community screening were personally contacted by the Bulgarian field team and provided with post-test counselling. Individuals who were negative for the carrier test were informed of the result in writing and in person where possible. There were no high-risk couples identified. The number of carriers identified in the sample suggested a carrier rate of 4.3% and a gene frequency of 0.022 in the population.

Upon returning the results to the community members, the Bulgarian field team became aware that the sample of screened individuals comprised two socially and culturally distinct groups, the Turgovzi and Feredjelli (L. Kalaydjieva, pers. comm.). Members of the community informed the field team that these two groups do not intermarry. Individuals were then assigned to one of these two groups. A total of 233 individuals declared themselves as Turgovzi and 101 declared themselves to be Feredjelli (table 13-11).

Table 13-11
Group affinities of individuals who requested carrier testing

	Turgovzi	Feredjelli
Married Females	69	19
Unmarried Females	42	33
Married Males	62	16
Unmarried Males	51	33
Total	224	101

The fourteen individuals identified as carriers in the community screening belong to the Turgovzi population. The C283Y mutation was absent in the 101 individuals belonging to the Feredjelli. Furthermore, additional investigations have revealed that all families with LGMD2C affected individuals are Turgovzi (L. Kalaydjieva, pers comm). Therefore, the detection of 14 C283Y carriers in a sample of 224 Turgovzi indicates a carrier frequency of 6.2%.

CHAPTER 14

DISCUSSION

14.1 Private Mendelian Disorders and Mutations Among Romani Populations

A number of private Mendelian disorders and/or disease-causing mutations have been reported in the Roma (Abicht et al., 1999; Angelicheva et al., 1999; Kalaydjieva et al., 1996; Kalaydjieva et al., 1999; Plasilova et al., 1999; Rogers et al., 2000). In these cases, complete allelic homogeneity and elevated gene frequencies have been observed due to founder effect. Thus, these mutations provide a rational starting point for molecular diagnosis when an affected individual is of known Romani ethnicity. A systematic survey of Romani populations has revealed high carrier rates for two founder mutations within populations (section III). This suggests that targeted community genetics programs may be justified.

The report of Piccolo et al., (1996) in which a founder C283Y mutation in *SGCG* was identified as a cause of LGMD2C in West European Roma, provided the impetus for investigations of the Bulgarian Romani population for this same disease-causing mutation. This study of the genetic basis of LGMD2C in people of Roma ethnicity has aimed to confirm the occurrence of the private C283Y mutation in geographically separated but historically related populations. The ramifications of this phenomenon to public health genetics has been explored through the initiation of a pilot community-screening program that aimed at providing genetic information to community members, and identifying salient social and psychological variables that may predictably affect the effectiveness of such an endeavour.

14.2 Biological Factors Impacting on Efficiency of Genetic Screening Program

14.2.1 Molecular Genetic Basis of LGMD2C in the Xoroxane Roma

The A→G transition in codon 283 of the *SGCG* was confirmed to be in the homozygous state in thirty of thirty-two clinically diagnosed limb girdle muscular dystrophy type 2C patients. Thus, the founder mutation first identified as the causative gene defect in Roma resident in Italy, France, Spain and Portugal (Lasa et al., 1998; Piccolo et al., 1996) is also present in at least one Balkan Romani population. The common origin of this unique event mutation is validated by disease haplotype analysis, which reveals the same well-conserved haplotype described in Iberian Roma patients (section III).

Whilst this mutation is clearly responsible for the majority of cases of LGMD2C it does not represent the only cause of the disease in the Romani population. Identification of a homozygous normal genotype in one patient and a heterozygous patient indicate an alternative aetiology of the muscular dystrophy in these patients.

14.2.2 Gene Frequencies and Carrier Rates

Genotyping of family members of affected individuals is clearly an efficient means of identifying carriers of the C283Y mutation. This form of carrier testing has been termed cascade testing by Super, Schwarz, & Malone, (1992). The fifty non-obligate carriers identified using this approach represent a carrier identification rate of 1 for every 2.5 family members tested. As carrier testing was offered to all family members regardless of age, a number of individuals were tested for whom reproductive issues may no longer be pertinent. However, the identification of twenty-eight carriers prior to marriage allows these individuals to incorporate this knowledge in future marital and reproductive decisions.

In screening for the mutation in the high-risk community, the identification of 14 carriers was initially understood to correspond with a carrier frequency of 4.3%.

However, further investigations revealed the restriction of these carriers to one of two distinct populations. The absence of any carriers in the 101 screened Feredjelli and the absence of any affected individuals in this group suggests that the mutation may not be present in this population. Thus, the carrier rate in the Turgovzi is in fact 6.2%. This result contrasts with a study of randomised neonatal screening samples for the C283Y mutation in undefined Romani groups from Northeastern Bulgarian which reported a heterozygote rate of 2.25% (Todorova, Ashikov, Beltcheva, Tournev, & Kremensky, 1999). This discrepant result highlights the necessity of obtaining carrier frequencies for a defined population. Bulgaria is known to be home to some fifty Roma groups who may or may not follow strict endogamous practices (Marushiakova & Popov, 1997). Whilst many of these populations are descended from a common ancestral population, the observation that a mutant allele can occur at a high frequency in one population and be absent in a population living in the immediate geographical vicinity highlights the need for appropriately structured epidemiological studies prior to the implementation of targeted genetic screening in Romani populations.

Whilst the cascade method of testing within families of affected individuals is the most efficient means of identifying carriers, population screening is justified when carrier rates are so high. A carrier frequency of 6.2% represents a significant risk for the members of the community and to restrict the offer of the test to family members would neglect addressing a significant public health concern.

14.2.3 Laboratory Design for Founder Mutations

Founder mutations enable very high detection rates in mutational analyses. The identification of a founder mutation in a homogeneous population presents the logical starting point for identification of carriers of heritable disorders. As such, the design of laboratory methods for this task is greatly simplified with only one known mutation to test.

Detection of all founder alleles does not, however, mean that all carriers of the disorder have been detected. If we examine the mutation detection rates in the thirty-two affected individuals in this study, it is clear that one hundred percent of all C283Y chromosomes were detected. However, the assay failed to detect the mutational basis of

the disease in three disease chromosomes (4.7%) which were found to be normal wildtype at codon 283 of the *SGCG*. Nevertheless, this degree of sensitivity is markedly more favourable than for the majority of monogenic diseases in outbred populations where detection rates are only 60-90% (van Ommen, Bakker, & den Dunnen, 1999). Allelic heterogeneity in large regional populations will result in low detection rates in population based carrier screening.

Clearly, targeted carrier testing in defined subpopulations for founder mutations will identify the greatest proportion of carriers for the lowest cost. Testing for a single mutation reduces laboratory costs and expedites large population screening. However, it must be borne in mind by the provider of genetic services and clearly communicated to the population that carrier detection for a known mutation does not completely correlate with carrier testing for a disorder. From the results obtained in genotyping LGMD2C affected individuals, it is reasonable to consider carrier detection rates of 95.3% for muscular dystrophy in this population.

14.2.4 Expected Trends in Carrier Rates

The social practice of endogamy, whilst fulfilling a social need, results in a greater incidence of autosomal recessive disorders (Bittles & Neel, 1994). Endogamous practices in a population increase the likelihood of two carriers of a recessive mutation forming a union but does not directly cause the proportion of carriers to increase. Factors that increase the number of carriers may be selective advantage and stochastic processes, such as random genetic drift. The effect of drift is more pronounced in small and isolated populations where it can result in the increased frequency of deleterious genes.

In the Turgovzi community investigated in this study, a carrier rate of 6.2% was observed which corresponds with a gene frequency of 0.031. An elevated frequency of a deleterious allele such as the C283Y mutation poses a significant potential risk to the population. If we consider the marital norms identified in the questionnaire, in approximately 97% of cases the marital partner will come from within the group. Therefore, the expected affected birth rate is 1/1072 births. It is likely, however, that the affected birth rate would be greater than this value as calculations of carrier rates in the

population have excluded those carriers identified through testing within families of affected individuals. More correctly, this is the expected incidence of affected births to parents who are unaware of a familial history of the disease.

14.3 Social Factors Relevant to Uptake of Genetic Testing and its Efficiency

14.3.1 Family Structure and Decision Making Regarding Marriage and Reproductive Issues

The majority of individuals investigated in this study live in households that contain at least three generations of family members. Furthermore, it is apparent from the questionnaire that members of the extended family are typically involved in decisions relating to marriage contracts and reproductive issues. In most cases, the parents are intimately involved in the selection and approval of their children's marriage partner. Similarly, decisions regarding genetic testing appear to be an issue that must be addressed by the family group, including the couple and both sets of parents. Therefore, it is clear that the provision of education and counselling for genetic issues should account for this dynamic through the inclusion of potential grandparents in the counselling process. This challenges typically held notions about the right to genetic privacy, which is generally asserted in the western medical setting (BMA, 1998). Clearly, every effort must be made to ensure that all individuals involved in the decision-making process are fully informed whilst meeting the criteria for respecting an individual's autonomy and privacy. Further investigations of culturally specific confidentiality systems would serve to identify the relevant level of privacy that is appropriate for members of this community. Close attention must also be paid to possible shifting family dynamics which may be evidenced by the apparent relationship between education (as ascertained by the vocabulary test) and greater individual control of reproductive decisions.

14.3.2 Major Social Concerns Examined in Community

14.3.2.1 Faith in medical investigations

The majority of community members appear to believe that medical investigations are able to identify carriers and utilise that information for the benefit of the individual. Less than one quarter of the sampled population indicated a lack of confidence in the ability of medicine either to determine carrier status or to do anything if carrier status was found to be positive. This result would seem to indicate that members of this community have a reasonable level of faith in medical practices. However, the answers provided indicate that, for some people, concerns about the efficacy and reliability of medical procedures involved may prevent them accepting a carrier test. These concerns may be addressed through continued education and communal discourse. It is also possible that through first hand observation of the successful provision of relevant genetic information, attitudes in the community may change positively.

14.3.2.2 Religious issues

A belief that the fate of a child's health is in God's hands and that interference is sinful is a commonly held view by the community members. Close to three-quarters of the questionnaire respondents chose this as a reason not to have a prenatal test over the belief that it is impossible to know whether an unborn child will have the disease. This view was even stronger in individuals who already had a child and even more so amongst those who reported knowledge of an affected family member. It is interesting to note that seventy-five percent of individuals who indicated this attitude still requested the carrier test for themselves. This points to a difference in declared attitudes and personal choice.

14.3.2.3 Concerns of stigmatisation

The implications of positive carrier status of a disease-causing mutation are clearly a concern for members of the community. Seventy-nine percent of those interviewed believe that some form of stigmatisation would be an outcome of being

identified as a carrier of the mutation and the majority of these people strongly agree that it would be an outcome. Interestingly, a greater proportion of people who know of an affected family member believe that an individual will never get married if she is identified as a carrier than those who do not know an affected family member. However, a significantly lower proportion of people who know a possibly affected family member believe that the whole family will lose face if an individual is identified as a carrier of the disorder.

Stigmatisation of carriers has been reported as a concern in many populations targeted for screening, including Greeks (Stamatoyannopoulos, 1972), Native Americans (Foster, Bernsten, & Carter, 1998) and Ashkenazi Jews (Rothenberg & Rutkin, 1998). However, follow-up studies of detected carriers have shown that these concerns diminish with time (Zeesman, Clow, Cartier, & Scriver, 1984). Concerns regarding stigmatisation of carriers can only be addressed through effective education and dialogue between the genetic service provider and the community. Information must be transmitted that assures the community of the benign nature of being a carrier for a recessive disorder and serves to empower those individuals who are carriers through active education.

14.3.3 Ameliorating Factors Impacting on Attitudes Towards Genetic Testing

Attitudes towards predictive testing were cross-tabulated with a number of parameters in order to identify factors that may impact on the receptivity of such a public health program. In general, there was a large degree of homogeneity in answers, which suggests well-entrenched cultural and social attitudes. However, the identification of factors that impact on attitudes towards testing may illuminate strategies for the implementation of an effective and useful program.

14.3.3.1 Knowledge of the disease within one's family

Knowledge of the occurrence of the disease within one's family had a strong impact on views held by individuals. Only 66% of people who knew the disease might occur in their families decided to take the test. Furthermore, of those individuals who knew of the disease almost all declared that the fate of a child's health was in God's

hands. Thus, it is apparent that knowing the disease occurs in one's family may be correlated with a greater degree of fatalism. However, this should not be interpreted as a reduced interest in knowing one's carrier status as evidenced by the large number of family members of affected individuals who requested the test. Clearly, a far smaller percentage of these people believe that identification as a carrier for the mutation would confer a loss of face for the whole family. However, fear of identification as a carrier may be increased among family members of affected individuals and impact on their decision to become informed about their carrier status.

14.3.3.2 Children

Having already produced children appears also to result in a greater degree of fatalism regarding the health of a newborn child. People with children also seem to place more emphasis on the role of the extended family in decision-making regarding reproductive issues. As it has been suggested that young adults prior to marriage and child-rearing are the most appropriate targets for carrier testing (BMA, 1998), it is possible that the attitudes of these people are the most important. In general, childless people in the community show a reduced concern with stigmatisation of carriers and are less inclined to believe that the fate of a child is in God's hands. Thus, members of this demographic group may be more receptive to counselling and testing for carrier status.

14.4 Summary of Findings

This study has demonstrated the implementation of a successful pilot carrier-screening program in a Balkan Romani community at increased risk for a disease-causing mutation. Genetic testing of patients has revealed the causative mutation as being identical to that found in Roma elsewhere in Europe. Cascade testing in the families of affected individuals has proved an efficient means of identifying a large number of carriers. A high gene frequency within a population justifies community wide screening for the mutation.

It is apparent that population structure, as revealed by social anthropology, warrants careful consideration in the design of such programs. The co-habitation of socially distinct groups that share a common origin does not necessarily mean that the

deleterious gene is shared between the groups. Thus, targeted population screening must be preceded by epidemiological investigations that pay close attention to the social and genetic structures of populations.

In general, this Romani population appears to be receptive to carrier testing. Education and counselling processes must be sensitive to family structures and decision-making processes and aim to minimise concerns about stigmatisation of carriers. This can only be achieved through participatory education and counselling and demonstration of the reproductive benefits of carrier identification. Further studies are required to answer the question of whether the benefits of such programs outweigh the potential harm. Long-term effectiveness studies should investigate the incorporation of carrier knowledge, a reduction in incidence of affected births and the psychological and social responses to knowledge of carrier status.

CHAPTER 15

CONCLUSION

15.1 Summary and Recapitulation

Historical records first mention the Roma in Europe some 800 years ago (Fraser, 1992). Chronicles from the following centuries allow the reconstruction of their migration into the Balkans and, some 200 years later, into Western Europe. Subsequent migrations of Roma have occurred in Europe, altering the demographic topography of the population. The single major migration of the last 500 years occurred in the 19th century when Roma left Wallachia and Moldavia following the end of their enslavement in those lands. From an early stage, the Roma have led a nomadic existence. To ensure their economic viability and minimise external hostilities, it has been necessary to exist in small groups. Thus, the Roma have fractured into a constellation of populations. The differing historical legacies of these groups have resulted in a mosaic of populations dispersed throughout Europe. These groups have differentiated and are socially and culturally diverse. In many cases, adherence to strict endogamous practices has ensured a strong internal cohesion within each of these groups, and at the same time expedited their divergence from one another.

In the absence of a recorded history, linguistics is able to clarify some of the opacity of the origins and historical relatedness of these groups. The languages of the Roma have been studied for over 200 years, informing us that the dialects of many of these groups stem from a common language of Indian origins (Fraser, 1992; Hancock, 2001). Additional shared cultural practices and traditions buttress this finding (Rishi, 1976). However, culture is known to be a rapidly evolving phenomenon that can easily outpace changes in the biological composition of a population. Furthermore, cultural traits can be acquired, and are not necessarily indicative of biological affinities. Therefore, studies of the genetic composition of a population provide a unique means of investigating population history.

In this thesis, separated Romani populations have been shown to share common maternal and paternal lineages of Asian origin. Whereas previous genetic studies of Romani populations have examined the frequency distribution of common polymorphic variants, these lineages represent discrete variants of demonstrated restriction to Asian populations. Thus, the evidence provides sound support for theories of the Indian origin of the Roma. Limited diversity within the Y chromosomal haplogroup VI-68 and the mitochondrial haplogroup M suggests that Romani populations are descended from a small number of related founders. This points to a pre-European Romani population of a single ethnic identity, rather than a conglomerate of people of different Indian origins. It is likely that the population that exited India was comprised primarily of a socially and ethnically distinct class of people that already had the features of a population isolate. It is unclear when the Roma left India, however linguistic analyses point to a departure after 1000 AD (Hancock, 2001). Coalescent dating of Y chromosome haplogroup VI-68 implies that the ancestral male population existed for some 1,000 years prior to the emigration of the proto-Roma.

The migration of the Roma from India into Europe can be estimated to have taken around 100 to 200 years. Linguistics points to a migration route through Persia and Armenia with possible extended stays in those regions (Fraser, 1992; Hancock, 1999). Genetic analyses in this study show a significant population component of probable Middle Eastern or Central Asian origins. Haplogroup VI-56 and mitochondrial haplogroup U3 are both most common in Middle Eastern populations. In the Roma, these lineages are found in different populations and display limited diversity, which provides evidence for pre-European admixture by a small number of related founders. Disentangling the population origins of the other male and female lineages is difficult given their lack of regional specificity. Certainly, many of these can be attributed to male- and female- mediated admixture by autochthonous European populations. However, unique lineages in the Roma are possibly representative of pre-European admixture. Discerning the history of these lineages in the Roma awaits clarification of their distribution in worldwide populations.

Extant Romani populations are related by common lineages but these lineages are nonrandomly dispersed, providing additional insights into the history of the Roma in

Europe. Male-specific genetic diversity is structured according to the major migrational groupings of Romani populations. Haplogroup VI-68 occurs at the highest frequencies in Vlach-speaking groups. These are groups that have at one time been resident in Wallachia, Moldavia and Rumania and presumably were enslaved to some degree. It is conceivable that slavery did much to prevent intermarriage between the Roma and the society that subjugated them. Similarly, the male components of populations contained within the Balkan and West European Roma are most closely related, which reflects common histories within Europe. In contrast, maternal lineages do not conform to migration groupings. Mitochondrial DNA lineages demonstrate a clear distinction between the Western European Roma and all other Roma. However, there is minimal substructuring within the female component of geographically proximate populations. This is possibly indicative of greater female-mediated gene flow between geographically proximate populations, and greater autochthonous European female admixture in the Balkan and Vlach Roma.

Long-term endogamous practices have evidently resulted in strikingly reduced genetic diversity in Romani populations. The genetic diversity of males is lowest in the Vlach populations, a reflection of the preservation of high frequencies of the founding haplogroup. The most restricted female genetic diversity is observed in the Lithuanian and Spanish Roma. These populations have relatively high male genetic diversity. Therefore, this suggests a stricter adherence to female endogamic practices. Further social anthropological investigations are required to confirm or refute this postulation.

The unique population structure of the Roma was shown to be essential for the identification of the founder mutation in *NDRG1* resulting in HMSNL. Whereas the localisation of the disease gene by Kalaydjieva et al., (1996) was facilitated by the limited diversity of disease haplotypes within a single pedigree, the refinement of the region benefited from the heterogeneity of disease chromosomes in different Romani populations. Reducing the size of the critical interval to 202kb, through the use of multiple historical recombinations, greatly ameliorated the task of searching for a disease-causing mutation. This is an essential step given that positional candidate genes in a chromosomal segment can be numerous. The value of sampling separated Romani groups is highlighted by the HMSNL locus refinement based on 6 historical

recombinations identified in 5 populations. This illustrates the necessity of including individuals from different Romani populations in refined mapping efforts of disease loci. The homogeneity of disease haplotypes observed within endogamous Romani populations would impede efforts at locus refinement that are limited to a single population.

Identification of the HMSNL gene defect contributes an essential step to understanding the cellular pathology of this particular disorder. The function of NDRG1 is poorly understood; however, published findings suggest its role as a signalling or chaperone molecule. Studies investigating the role of this protein in the affected tissues of individuals with HMSNL should elucidate the biochemical or mechanistic dysfunction. In addition, future studies of *NDRG1* should provide important insights into normal Schwann cell-axon interactions and development. This should have applications beyond understanding and treating HMSNL, as the same cellular mechanisms could underlie a number of neurological disorders, both inherited and acquired.

The HMSNL mutation is widely dispersed in the Roma. Screening for R148X heterozygotes revealed carriers in Vlach, Balkan and West European populations. The disorder has been described in Italian (Merlini et al., 1998), Slovenian (Butinar et al., 1999), Spanish (Colomer et al., 2000), French and Rumanian Roma (Chandler et al., 2001), in addition to the initially reported occurrence in Vlach Roma (Kalaydjieva et al., 1996). The prevalence of this disease allele can be explained by its genesis over 1000 years ago, prior to the fracturing of Romani populations. Thus, the R148X disease allele can be classed as a significant health risk in the Roma. A carrier frequency of 19.5% in the Lom population provides the impetus for widespread community screening. In other populations, the frequency of carriers justifies cascade screening and possibly community-wide carrier testing.

In contrast to the R148X disease allele, the C283Y allele is not widely distributed in Roma. This allele possibly arose after the arrival of the Roma in Europe, thereby restricting its distribution in separated populations. The C283Y mutation has been reported in a number of Western European Romani populations (Lasa et al., 1998; Piccolo et al., 1996). Screening of populations in this study identified a single C283Y

carrier in the sample of Spanish Roma. Among the Balkan and Vlach Roma it was only identified in the Turgovzi. In this population it occurs at high frequency, with a carrier rate of 6.25%. The experience of the LGMD2C screening in the Turgovzi population in North East Bulgaria suggests that community screening is suitable when a large number of affected individuals are observed. Limiting access to a carrier test to individuals related to affected persons is not justifiable in light of the demonstrated high carrier rates. Moreover, this pilot public health genetics program has shown the receptivity of a Romani community. However, attitudes towards predictive testing are culturally specific. As Romani populations are culturally heterogeneous, the findings from this pilot study may not be universally applicable.

For genetic research, the high allele frequencies observed in the Roma have profound implications for strategies that may be employed to identify disease genes. Searching for IBD segments, using approaches such as homozygosity mapping and segment-sharing run a high risk of failure. This is because high gene frequencies combined with the old age of a mutation are likely to result in disease haplotypes that cannot be detected using conventional 10cM maps. Although the disease haplotype may not be preserved over large regions, it is possible that significant allelic associations may still be detected at distant marker loci. A recent study has demonstrated the success of a genome-wide search for linkage disequilibrium in mapping a rare form of cytochrome oxidase deficiency in French Canadians (Lee et al., 2001). In this study, the authors used untransmitted parental alleles to estimate allele frequencies in the unaffected population. However, in association studies, parents of affected individuals only provide half the information of unrelated individuals. The limited genetic diversity within Romani populations implies that rare genetic disorders can reasonably be expected to result from founder mutations. In disorders that have been examined thus far complete allelic homogeneity has been observed. Thus, linkage disequilibrium mapping for monogenic disorders need not be confined to pedigree data. True population-based case-control studies can be predicted to be fruitful in the Roma. This approach would be particularly useful for those diseases for which extended pedigrees are uncommon, such as late onset disorders and those that result in early mortality.

15.2 Future Directions

This study has resolved some of the long-standing questions regarding the history and origins of the Roma. Further studies should aim at refining the origin of the Roma within the Indian subcontinent. Searches for the distribution of Y chromosome haplogroup VI-68 and mitochondrial DNA haplogroup M5 is one possible way to seek for related populations within India. These lineages would also prove illuminating in examining other populations of hypothesised relatedness to the Roma, such as the Lom and Dom of Central Asia and the Middle East, and the many non-Romani itinerant groups within Europe. The observed differences in sex-specific histories in the Roma warrant further investigation. It is possible that some of the differences are an artefact of the variation between Y chromosome and mtDNA mutational processes. To overcome this problem, use of the X chromosome should be investigated. As UEPs and microsatellite DNA are found on the X chromosome, this would provide more directly comparable data for males and females. Ideally, sample sizes for populations should be increased in order to provide robust results.

Future gene mapping efforts in the Roma would benefit from the examination of known disease loci. It is proposed that genome-wide scans for linkage disequilibrium within defined Romani populations are a possible approach to mapping of monogenic traits. However, for this to be efficient, it is necessary to know the extent to which disease haplotypes are conserved in different individuals. This would dictate the required marker density of a genome scan for linkage disequilibrium. Knowledge of the extent of background linkage disequilibrium would benefit this endeavour. Furthermore, this investigation is of relevance to the proposed role of linkage disequilibrium mapping for the genes underlying complex traits (Risch, 2000). Association between alleles can result from different demographic histories, genetic drift and admixture in addition to physical linkage,. Linkage disequilibrium varies throughout the genome, and with different population histories and structure, thus necessitating empirical investigations. Population modelling has suggested that linkage disequilibrium around common variants can be expected to only extend over 3 kb genomic regions in heterogeneous populations (Kruglyak, 1999). However, recent studies have shown that large chromosomal blocks are in linkage disequilibrium in north

Europeans (Reich et al., 2001). The extent of linkage disequilibrium in population isolates of different histories is debated. Evidence suggests that linkage disequilibrium is moderately higher in expanded population isolates such as the Finns and Sardinians (Boehnke, 2000) Others have pointed to higher levels of linkage disequilibrium in demographically stable population isolates such as the Saami (Laan & Paabo, 1997). Admixture has been shown to result in long-range linkage disequilibrium in the Lemba (Wilson & Goldstein, 2000) and African Americans (Pfaff et al., 2001). Thus, examination of linkage disequilibrium in the Roma, who have been demonstrated to be *admixed population isolates*, would be of great interest. Moreover, the study of linkage disequilibrium has been demonstrated to provide additional insights into population history (Reich et al., 2001).

Rare genetic disorders provide insights into gene function. In population isolates it is possible to find many individuals affected by the same disorder. This provides a useful resource for studying disease variation and modifying effects. In a study of 40 C283Y homozygous LGMD2C affected Romani individuals, variation in the severity of the phenotype was observed (Merlini et al., 2000). This is a consistent finding for monogenic disorders that challenges the concept of "simple" mendelian disorders (Scriver & Waters, 1999). If sample size is large enough, modifying genes can be investigated using candidate gene association studies (Cazeneuve et al., 2000) or linkage studies within pedigrees for genes of strong effect (Riazuddin et al., 2000). Both of these approaches would be feasible in the Roma. Disease alleles in the Roma also provide an important biological source material for further studies. The R148X mutation in *NDRG1* represents a null mutation, which is biologically equivalent to a gene knockout in model organisms. Expression studies of tissues expressing this gene would prove illuminating in revealing molecular pathways and interactions. This should not be limited to only the tissue believed to be involved in the disease pathology. Recent estimates suggest 30-40 000 genes in the human genome (Lander et al., 2001), however the transcriptome is believed to contain 100 000 unique transcripts. Thus, genes are likely to have multiple and diverse functions in different tissues. This implies that the same gene defect may result in different pathological pathways in different cell types. Therefore, tissue-specific studies of the effect of mutant alleles would prove interesting.

BIBLIOGRAPHY

Abicht, A., Stucka, R., Karcagi, V., Herczegfalvi, A., Horváth, R., Mortier, W., Schara, U., Ramaekers, V., Jost, W., Brunner, J., Janssen, G., Seidel, U., Schlotter, B., Müller-Felber, W., Pongratz, D., Rüdel, R., & Lochmüller, H. (1999). A common mutation (epsilon1267delG) in congenital myasthenic patients of Gypsy ethnic origin. *Neurology, 53*(7), 1564-1569.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res, 25*(17), 3389-3402.

Alves-Silva, J., da Silva Santos, M., Guimaraes, P. E., Ferreira, A. C., Bandelt, H. J., Pena, S. D., & Prado, V. F. (2000). The ancestry of Brazilian mtDNA lineages. *Am J Hum Genet, 67*(2), 444-461.

Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J., Staden, R., & Young, I. G. (1981). Sequence and organization of the human mitochondrial genome. *Nature, 290*(5806), 457-465.

Angelicheva, D., Turnev, I., Dye, D., Chandler, D., Thomas, P. K., & Kalaydjieva, L. (1999). Congenital cataracts facial dysmorphism neuropathy (CCFDN) syndrome: a novel developmental disorder in Gypsies maps to 18qter. *Eur J Hum Genet, 7*(5), 560-566.

Arnason, E., Sigurgislason, H., & Benedikz, E. (2000). Genetic homogeneity of Icelanders: fact or fiction? *Nat Genet, 25*(4), 373-374.

ASHG/ACMG. (1995). Points to consider: ethical, legal, and psychosocial implications of genetic testing in children and adolescents. American Society of Human Genetics Board of Directors, American College of Medical Genetics Board of Directors. *Am J Hum Genet, 57*(5), 1233-1241.

Avcin, M. (1969). Gypsy isolates in Slovenia. *J Biosoc Sci, 1*(3), 221-233.

Awadalla, P., Eyre-Walker, A., & Smith, J. M. (1999). Linkage disequilibrium and recombination in hominid mitochondrial DNA. *Science, 286*(5449), 2524-2525.

Baethmann, M., Gohlich-Ratmann, G., Schroder, J. M., Kalaydjieva, L., & Voit, T. (1998). HMSNL in a 13-year-old Bulgarian girl. *Neuromuscul Disord, 8*(2), 90-94.

Bamshad, M., Kivisild, T., Watkins, W. S., Dixon, M. E., Ricker, C. E., Rao, B. B., Naidu, J. M., Prasad, B. V., Reddy, P. G., Rasanayagam, A., Papiha, S. S., Villems, R., Redd, A. J., Hammer, M. F., Nguyen, S. V., Carroll, M. L., Batzer, M. A., & Jorde, L. B. (2001). Genetic evidence on the origins of indian caste populations. *Genome Res, 11*(6), 994-1004.

Bamshad, M. J., Watkins, W. S., Dixon, M. E., Jorde, L. B., Rao, B. B., Naidu, J. M., Prasad, B. V., Rasanayagam, A., & Hammer, M. F. (1998). Female gene flow stratifies Hindu castes. *Nature, 395*(6703), 651-652.

Bandelt, H. J., Forster, P., Sykes, B. C., & Richards, M. B. (1995). Mitochondrial portraits of human populations using median networks. *Genetics, 141*(2), 743-753.

Barhoumi, C., Amouri, R., Ben Hamida, C., Ben Hamida, M., Machghoul, S., Gueddiche, M., & Hentati, F. (2001). Linkage of a new locus for autosomal recessive axonal form of Charcot- Marie-Tooth disease to chromosome 8q21.3. *Neuromuscul Disord, 11*(1), 27-34.

Bartsocas, C. S., Karayanni, C., Tsipouras, P., Baibas, E., Bouloukos, A., & Papadatos, C. (1979). Genetic structure of the Greek gypsies. *Clin Genet, 15*(1), 5-10.

Baas, F., van Ommen, G. J., Bikker, H., Arnberg, A. C., & de Vijlder, J. J. (1986). The human thyroglobulin gene is over 300 kb long and contains introns of up to 64 kb. *Nucleic Acids Res, 14*(13), 5171-5186.

Beckman, L., & Takman, J. (1965). On the anthropology of a Swedish Gypsy population. *Hereditas, 53*(1), 272-280.

Bekker, H., Modell, M., Denniss, G., Silver, A., Mathew, C., Bobrow, M., & Marteau, T. (1993). Uptake of cystic fibrosis testing in primary care: supply push or demand pull? *Brit Med J, 306*(6892), 1584-1586.

Ben Othmane, K., Hentati, F., Lennon, F., Ben Hamida, C., Blel, S., Roses, A. D., Pericak-Vance, M. A., Ben Hamida, M., & Vance, J. M. (1993). Linkage of a locus (CMT4A) for autosomal recessive Charcot-Marie-Tooth disease to chromosome 8q. *Hum Mol Genet, 2*(10), 1625-1628.

Ben Othmane, K., Speer, M. C., Stauffer, J., Blel, S., Middleton, L., Ben Hamida, C., Etribi, A., Loeb, D., Hentati, F., Roses, A. D., & et al. (1995). Evidence for linkage disequilibrium in chromosome 13-linked Duchenne- like muscular dystrophy (LGMD2C). *Am J Hum Genet, 57*(3), 732-734.

Bernasovsky, I., Halko, N., Biros, I., Sivakova, D., & Jurickova, J. (1994). Some genetic markers in Valachian (Olachian) Gypsies in Slovakia. *Gene Geogr, 8*(2), 99-107.

Bernasovsky, I., Suchy, J., Bernasovska, K., & Vargova, T. (1976). Blood groups of Roms (Gypsies) in Czechoslovakia. *Am J Phys Anthropol, 45*(2), 277-280.

Bernhardt, B. A. (1997). Empirical evidence that genetic counseling is directive: where do we go from here? *Am J Hum Genet, 60*(1), 17-20.

Bertranpetit, J. (2000). Genome, diversity, and origins: the Y chromosome as a storyteller. *Proc Natl Acad Sci U S A, 97*(13), 6927-6929.

Bertranpetit, J., & Calafell, F. (1996). Genetic and geographical variability in cystic fibrosis: evolutionary considerations. *Ciba Found Symp, 197*, 97-114.

Bertranpetit, J., Calafell, F., Comas, D., Pérez-Lezaun, A., & Mateu, E. (1996). Mitochondrial DNA sequences in Europe: an insight into population history. In A.J. Boyce & C.G.N. Mascie-Yalor (Eds.), *Molecular Biology and Human Diversity* (pp. 112-129) Cambridge: Cambridge University Press.

Bhattacharyya, N. P., Basu, P., Das, M., Pramanik, S., Banerjee, R., Roy, B., Roychoudhury, S., & Majumder, P. P. (1999). Negligible male gene flow across ethnic boundaries in India, revealed by analysis of Y-chromosomal DNA polymorphisms. *Genome Res, 9*(8), 711-719.

Bittles, A. H., & Neel, J. V. (1994). The costs of human inbreeding and their implications for variations at the DNA level. *Nat Genet, 8*(2), 117-121.

Bittles, A. H., Mason, W. M., Greene, J., & Rao, N. A. (1991). Reproductive behavior and health in consanguineous marriages. *Science, 252*(5007), 789-794.

Black, M. (1999). *Analysis of the population genetics of the Han and Hui of Liaoning province, Peoples Republic of China.* Unpublished MSc, Edith Cowan University, Joondalup

Blaney, G (2000). *A Short History of the World.* Maryborough. Penguin Books Australia Ltd.

BMA. (1998). *Human Genetics: choice and responsibilities*. Oxford: Oxford University Press.

Boehnke, M. (2000). A look at linkage disequilibrium. *Nat Genet, 25*(3), 246-247.

Bolino, A., Brancolini, V., Bono, F., Bruni, A., Gambardella, A., Romeo, G., Quattrone, A., & Devoto, M. (1996). Localization of a gene responsible for autosomal recessive demyelinating neuropathy with focally folded myelin sheaths to chromosome 11q23 by homozygosity mapping and haplotype sharing. *Hum Mol Genet, 5*(7), 1051-1054.

Bolino, A., Muglia, M., Conforti, F. L., LeGuern, E., Salih, M. A., Georgiou, D. M., Christodoulou, K., Hausmanowa-Petrusewicz, I., Mandich, P., Schenone, A., Gambardella, A., Bono, F., Quattrone, A., Devoto, M., & Monaco, A. P. (2000). Charcot-Marie-Tooth type 4B is caused by mutations in the gene encoding myotubularin-related protein-2. *Nat Genet, 25*(1), 17-19.

Bonné-Tamir, B., Nystuen, A., Seroussi, E., Kalinsky, H., Kwitek-Black, A. E., Korostishevsky, M., Adato, A., & Sheffield, V. C. (1997). Usher syndrome in the Samaritans: strengths and limitations of using inbred isolated populations to identify genes causing recessive disorders. *Am J Phys Anthropol, 104*(2), 193-200.

Boretzky, N. (1995). Interdialectal Interference in Romani. In Y. Matras (Ed.), *Romani in Contact: the history, structure and sociology of a language*. (pp. 27-35)Amsterdam: John Benjamins Publishing Company.

Bosch, E., Calafell, F., Comas, D., Oefner, P. J., Underhill, P. A., & Bertranpetit, J. (2001). High-resolution analysis of human Y-chromosome variation shows a sharp discontinuity and limited gene flow between northwestern Africa and the Iberian Peninsula. *Am J Hum Genet, 68*(4), 1019-1029.

Bosch, E., Calafell, F., Santos, F. R., Pérez-Lezaun, A., Comas, D., Benchemsi, N., Tyler-Smith, C., & 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., White, R. L., Skolnick, M., & Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet, 32*(3), 314-331.

Bouchard, J. P., Richter, A., Mathieu, J., Brunet, D., Hudson, T. J., Morgan, K., & Melancon, S. B. (1998). Autosomal recessive spastic ataxia of Charlevoix-Saguenay. *Neuromuscul Disord, 8*(7), 474-479.

Bouzekri, N., Taylor, P. G., Hammer, M. F., & Jobling, M. A. (1998). Novel mutation processes in the evolution of a haploid minisatellite, MSY1: array homogenization without homogenization. *Hum Mol Genet, 7*(4), 655-659.

Bowles Biesecker, B., & Marteau, T. M. (1999). The future of genetic counselling: an international perspective. *Nat Genet, 22*(2), 133-137.

Brinkmann, B., Klintschar, M., Neuhuber, F., Huhne, J., & Rolf, B. (1998). Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *Am J Hum Genet, 62*(6), 1408-1415.

Brock, D. J. (1994). Carrier screening for cystic fibrosis. *Prenat Diagn, 14*(13), 1243-1252.

Brock, D. J. (1995). Carrier screening in the community. *J Inherit Metab Dis, 18*(4), 525-532.

Brown, D. T., Samuels, D. C., Michael, E. M., Turnbull, D. M., & Chinnery, P. F. (2001). Random genetic drift determines the level of mutant mtDNA in human primary oocytes. *Am J Hum Genet, 68*(2), 533-536.

Butinar, D., Zidar, J., Leonardis, L., Popovic, M., Kalaydjieva, L., Angelicheva, D., Sininger, Y., Keats, B., & Starr, A. (1999). Hereditary auditory, vestibular, motor, and sensory neuropathy in a Slovenian Roma (Gypsy) kindred. *Ann Neurol, 46*(1), 36-44.

Calafell, F., Underhill, P., Tolun, A., Angelicheva, D., & Kalaydjieva, L. (1996). From Asia to Europe: mitochondrial DNA sequence variability in Bulgarians and Turks. *Ann Hum Genet, 60*(1), 35-49.

Calloway, C. D., Reynolds, R. L., Herrin, G. L., Jr., & Anderson, W. W. (2000). The frequency of heteroplasmy in the HVII region of mtDNA differs across tissue types and increases with age. *Am J Hum Genet, 66*(4), 1384-1397.

Cann, R. L., Stoneking, M., & Wilson, A. C. (1987). Mitochondrial DNA and human evolution. *Nature, 325*(6099), 31-36.

Carvajal-Carmona, L. G., Soto, I. D., Pineda, N., Ortiz-Barrientos, D., Duque, C., Ospina-Duque, J., McCarthy, M., Montoya, P., Alvarez, V. M., Bedoya, G., & Ruiz-Linares, A. (2000). Strong Amerind/white sex bias and a possible Sephardic contribution among the founders of a population in northwest Colombia. *Am J Hum Genet, 67*(5), 1287-1295.

Carvalho-Silva, D. R., Santos, F. R., Hutz, M. H., Salzano, F. M., & Pena, S. D. (1999). Divergent human Y-chromosome microsatellite evolution rates. *J Mol Evol, 49*(2), 204-214.

Carvalho-Silva, D. R., Santos, F. R., Rocha, J., & Pena, S. D. (2001). The phylogeography of Brazilian Y-chromosome lineages. *Am J Hum Genet, 68*(1), 281-286.

Casana, P., Martinez, F., Haya, S., Lorenzo, J. I., Espinos, C., & Aznar, J. A. (2000). Q1311X: a novel nonsense mutation of putative ancient origin in the von Willebrand factor gene. *Br J Haematol, 111*(2), 552-555.

Casanova, M., Leroy, P., Boucekkine, C., Weissenbach, J., Bishop, C., Fellous, M., Purrello, M., Fiori, G., & Siniscalco, M. (1985). A human Y-linked DNA polymorphism and its potential for estimating genetic and evolutionary distance. *Science, 230*(4732), 1403-1406.

Cavalli-Sforza, L. L. (1998). The DNA revolution in population genetics. *Trends Genet, 14*(2), 60-65.

Cavelier, L., Jazin, E., Jalonen, P., & Gyllensten, U. (2000). MtDNA substitution rate and segregation of heteroplasmy in coding and noncoding regions. *Hum Genet, 107*(1), 45-50.

Cazal, P., Graafland, R., & Mathieu, M. (1951). *Les groupes sanguins chez les gitans de Frances.* Paper presented at the 4th Inst. Congress Blood Transfusion, Lisbon.

Cazeneuve, C., Ajrapetyan, H., Papin, S., Roudot-Thoraval, F., Genevieve, D., Mndjoyan, E., Papazian, M., Sarkisian, A., Babloyan, A., Boissier, B., Duquesnoy, P., Kouyoumdjian, J. C., Girodon-Boulandet, E., Grateau, G., Sarkisian, T., & Amselem, S. (2000). Identification of MEFV-independent modifying genetic factors for familial Mediterranean fever. *Am J Hum Genet, 67*(5), 1136-1143.

Ceu Moreira, M., Barbot, C., Tachi, N., Kozuka, N., Mendonca, P., Barros, J., Coutinho, P., Sequeiros, J., & Koenig, M. (2001). Homozygosity mapping of Portuguese and Japanese forms of ataxia-oculomotor apraxia to 9p13, and evidence for genetic heterogeneity. *Am J Hum Genet, 68*(2), 501-508.

Chakravarti, A. (2001). To a future of genetic medicine. *Nature, 409*(6822), 822-823.

Chang, M. L., Eddy, R. L., Shows, T. B., & Lau, J. T. (1995). Three genes that encode human beta-galactoside alpha 2,3- sialyltransferases. Structural analysis and chromosomal mapping studies. *Glycobiology, 5*(3), 319-325.

Charbonneau, M., Laberge, C., Scriver, C. R., Dussault, J. H., Lemieux, B., & Melancon, S. (1987). The Quebec Network of Genetic Medicine. *Can J Public Health, 78*(2), 79-83.

Chen, Y. S., Torroni, A., Excoffier, L., Santachiara-Benerecetti, A. S., & Wallace, D. C. (1995). Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. *Am J Hum Genet, 57*(1), 133-149.

Chiba-Falek, O., Nissim-Rafinia, M., Argaman, Z., Genem, A., Moran, I., Kerem, E., & Kerem, B. (1998). Screening of CFTR mutations in an isolated population: identification of carriers and patients. *Eur J Hum Genet, 6*(2), 181-184.

Chinnery, P. F., Thorburn, D. R., Samuels, D. C., White, S. L., Dahl, H. M., Turnbull, D. M., Lightowlers, R. N., & Howell, N. (2000). The inheritance of mitochondrial DNA heteroplasmy: random drift, selection or both? *Trends Genet, 16*(11), 500-505.

Clarke, A. (1991). Is non-directive genetic counselling possible? *Lancet, 338*(8773), 998-1001.

Clarke, V. A. (1973). Genetic factors in some British Gypsies. In D. F. Roberts & E. Sunderland (Eds.), *Genetic Variation in Britain*. London: Taylor and Francis.

Collins, A. (2000). Mapping in the sequencing era. *Hum Hered, 50*(1), 76-84.

Collins, F. S. (1992). Positional cloning: let's not call it reverse anymore. *Nat Genet, 1*(1), 3-6.

Collins, F. S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R., & Walters, L. (1998). New goals for the U.S. Human Genome Project: 1998-2003. *Science, 282*(5389), 682-689.

Colomer, J., Iturriaga, C., Kalaydjieva, L., Angelicheva, D., King, R. H., & Thomas, P. K. (2000). Hereditary motor and sensory neuropathy-Lom (HMSNL) in a Spanish family: clinical, electrophysiological, pathological and genetic studies. *Neuromuscul Disord, 10*(8), 578-583.

Comas, D., Calafell, F., Mateu, E., Péréz-Lezaun, A., Bosch, E., & Bertranpetit, J. (1997). Mitochondrial DNA variation and the origin of the Europeans. *Hum Genet, 99*(4), 443-449.

Comas, D., Calafell, F., Mateu, E., Péréz-Lezaun, A., Bosch, E., Martinez-Arias, R., Clarimon, J., Facchini, F., Fiori, G., Luiselli, D., Pettener, D., & Bertranpetit, J. (1998). Trading genes along the silk road: mtDNA sequences and the origin of central Asian populations. *Am J Hum Genet, 63*(6), 1824-1838.

Cotton, R. G., & Scriver, C. R. (1998). Proof of "disease causing" mutation. *Hum Mutat, 12*(1), 1-3.

Crowe, D. (1991). The Gypsy Historical Experience in Romania. In D. Crowe & J. Kolsti (Eds.), *The Gypsies of Eastern Europe*. (pp. 61-80)London: M.E. Sharpe, Inc.

de Knijff, P. (2000). Messages through bottlenecks: on the combined use of slow and fast evolving polymorphic markers on the human Y chromosome. *Am J Hum Genet, 67*(5), 1055-1061.

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., Péréz-Lezaun, A., Piccinini, A., Prinz, M., Roewer, L., & et al. (1997). Chromosome Y microsatellites: population genetic and evolutionary aspects. *Int J Legal Med, 110*(3), 134-149.

de la Chapelle, A. (1993). Disease gene mapping in isolated human populations: the example of Finland. *J Med Genet, 30*(10), 857-865.

de Pablo, R., Vilches, C., Moreno, M. E., Rementeria, M., Solis, R., & Kreisler, M. (1992). Distribution of HLA antigens in Spanish Gypsies: a comparative study. *Tissue Antigens, 40*(4), 187-196.

Delghandi, M., Utsi, E., & Krauss, S. (1998). Saami mitochondrial DNA reveals deep maternal lineage clusters. *Hum Hered, 48*(2), 108-114.

Desviat, L. R., Perez, B., & Ugarte, M. (1997). Phenylketonuria in Spanish Gypsies: prevalence of the IVS10nt546 mutation on haplotype 34. *Hum Mutat, 9*(1), 66-68.

Di Rienzo, A., Peterson, A. C., Garza, J. C., Valdes, A. M., Slatkin, M., & Freimer, N. B. (1994). Mutational processes of simple-sequence repeat loci in human populations. *Proc Natl Acad Sci U S A, 91*(8), 3166-3170.

Efron, B. (1982). *The jacknife, the bootstrap and other resampling plans.* Paper presented at the Regional conference series in applied mathematics, Philadelphia

Ellegren, H. (2000). Heterogeneous mutation processes in human microsatellite DNA sequences. *Nat Genet, 24*(4), 400-402.

Elson, J. L., Andrews, R. M., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M., & Howell, N. (2001). Analysis of European mtDNAs for recombination. *Am J Hum Genet, 68*(1), 145-153.

Ely, B. (1961). Les groupes sanguin de 47 Tsiganes de la region parisienne. *Bull. Soc. Anthropol. Paris, 2*, 233-237.

Engert, J. C., Berube, P., Mercier, J., Dore, C., Lepage, P., Ge, B., Bouchard, J. P., Mathieu, J., Melancon, S. B., Schalling, M., Lander, E. S., Morgan, K., Hudson, T. J., & Richter, A. (2000). ARSACS, a spastic ataxia common in northeastern Quebec, is caused by mutations in a new gene encoding an 11.5-kb ORF. *Nat Genet, 24*(2), 120-125.

Excoffier, L., Smouse, P. E., & Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics, 131*(2), 479-491

Excoffier, L., & Yang, Z. (1999). Substitution rate variation among sites in mitochondrial hypervariable region I of humans and chimpanzees. *Mol Biol Evol, 16*(10), 1357-1368.

Eyre-Walker, A. (2000). Do mitochondria recombine in humans? *Philos Trans R Soc Lond B Biol Sci, 355*(1403), 1573-1580.

Felsenstein, J. (1989). PHYLIP - Phylogeny inference package (version 3.2). *Cladistics, 5*, 164-166.

Ferak, V., Gencik, A., & Gencikova, A. (1982). Population genetic aspects of primary congenital glaucoma. II. Fitness, parental consanguinity, founder effect. *Hum Genet, 61*(3), 198-200.

Ferak, V., Sivakova, D., & Sieglova, Z. (1987). [The Slovak gypsies (Romany)--a population with the highest coefficient of inbreeding in Europe]. *Bratisl Lek Listy, 87*(2), 168-175.

Finnilä, S., Lehtonen, M. S., & Majamaa, K. (2001). Phylogenetic Network for European mtDNA. *Am J Hum Genet, 68*(6), 1475-1484.

Forrai, G., Tauszik, T., Tauszik, N., Mohr, T., Tunyogi, M. C., Holics, C., Bankovi, G., & Gal, I. (1989). A high incidence of PKD in a large geographic area of south-western Hungary: a medical genetic study. *Prog Clin Biol Res, 305*, 89-94.

Forster, P., Kayser, M., Meyer, E., Roewer, L., Pfeiffer, H., Benkmann, H., & Brinkmann, B. (1998). Phylogenetic resolution of complex mutational features at Y-STR DYS390 in aboriginal Australians and Papuans. *Mol Biol Evol, 15*(9), 1108-1114.

Forster, P., Rohl, A., Lunnemann, P., Brinkmann, C., Zerjal, T., Tyler-Smith, C., & Brinkmann, B. (2000). A short tandem repeat-based phylogeny for the human Y chromosome. *Am J Hum Genet, 67*(1), 182-196.

Foster, M. W., Bernsten, D., & Carter, T. H. (1998). A model agreement for genetic research in socially identifiable populations. *Am J Hum Genet, 63*(3), 696-702.

Furedi, S., Woller, J., Padar, Z., & Angyal, M. (1999). Y-STR haplotyping in two Hungarian populations. *Int J Legal Med, 113*(1), 38-42

Fraser, A. (1992). *The Gypsies*. Oxford: Blackwell Publishers.

Galikova, J., Vilimova, M., Ferak, V., & Mayerova, A. (1969). Haptoglobin types in gypsies from Slovakia (Czechoslovakia). *Hum Hered, 19*(5), 480-485.

Gatliff, T. (1993). Latcho Drom (T. Gatliff, Director). In M. Ray (Producer). Waterville, Maine: Shadow Distribution.

Gilbert, F. (1998). Establishing criteria for a carrier detection panel: lessons from the Ashkenazi Jewish model. *Genet Test, 2*(4), 301-304.

Goldstein, D. B., & Pollock, D. D. (1997, 21/5/97). *A Review of Mutation Processes and Methods of Phylogenetic Inference*, [Web Page]. Available: http://lotka.stanford.edu/microdist.html [1998, 13/11/98].

Goldstein, D. B., Reich, D. E., Bradman, N., Usher, S., Seligsohn, U., & Peretz, H. (1999). Age estimates of two common mutations causing factor XI deficiency: recent genetic drift is not necessary for elevated disease incidence among Ashkenazi Jews. *Am J Hum Genet, 64*(4), 1071-1075.

Gomez, P. S., Parks, S., Ries, R., Tran, T. C., Gomez, P. F., & Press, R. D. (1999). Polymorphism in intron 4 of HFE does not compromise haemochromatosis mutation results. *Nat Genet, 23*(3), 272.

Goodman, R. M. (Ed.). (1979). *Genetic Disorders Among the Jewish People*. Baltimore: The Johns Hopkins University Press.

Guglielmino, C. R., & Beres, J. (1996). Genetic structure in relation to the history of Hungarian ethnic groups. *Hum Biol, 68*(3), 335-355.

Guo, S. W., & Xiong, M. (1997). Estimating the age of mutant disease alleles based on linkage disequilibrium. *Hum Hered, 47*(6), 315-337.

Gyllensten, U., Wharton, D., Josefsson, A., & Wilson, A. C. (1991). Paternal inheritance of mitochondrial DNA in mice. *Nature, 352*(6332), 255-257.

Gyodi, E., Tauszik, T., Petranyi, G., Kotvasz, A., Palffy, G., Takacs, I., Nemak, P., & Hollan, S. R. (1981). The HLA antigen distribution in the Gipsy population in Hungary. *Tissue Antigens, 18*(1), 1-12.

Hack, A. A., Ly, C. T., Jiang, F., Clendenin, C. J., Sigrist, K. S., Wollmann, R. L., & McNally, E. M. (1998). Gamma-sarcoglycan deficiency leads to muscle membrane defects and apoptosis independent of dystrophin. *J Cell Biol, 142*(5), 1279-1287.

Hammer, M. F. (1994). A recent insertion of an alu element on the Y chromosome is a useful marker for human population studies. *Mol Biol Evol, 11*(5), 749-761.

Hammer, M. F., & Horai, S. (1995). Y chromosomal DNA variation and the peopling of Japan. *Am J Hum Genet, 56*(4), 951-962.

Hammer, M. F., Redd, A. J., Wood, E. T., Bonner, M. R., Jarjanazi, H., Karafet, T., Santachiara-Benerecetti, S., Oppenheim, A., Jobling, M. A., Jenkins, T., Ostrer, H., & Bonné-Tamir, B. (2000). Jewish and Middle Eastern non-Jewish populations share a common pool of Y-chromosome biallelic haplotypes. *Proc Natl Acad Sci U S A, 97*(12), 6769-6774.

Hancock, I. (1987). *The Pariah Syndrome*. Ann Arbor: Karoma Publishrs Inc.

Hancock, I. (1991). Gypsy History in Germany and Neighbouring Lands: a chronology leading to the holocaust and beyond. In D. Crowe & J. Kolsti (Eds.), *The Gypsies of Eastern Europe* (pp. 11-31). London: M.E. Sharpe Inc.

Hancock, I. (1999a). *The Indian Origin and Westward Migration of the Romani People*.Unpublished manuscript.

Hancock, I. (1999b). *Origins of the Romani People*, [Web page]. The Patrin Web Journal. Available: http://www.geocities.com/Paris/5121/history.htm [accessed 1999].

Hancock, I. (2000). The Emergence of Romani as a Koine Outside of India. In T. Acton (Ed.), *Scholarship and Gypsy Struggle: commitment in Romani studies* (Vol. Essays in Honour of Donald Kenrick on the Occassion of his Seventieth Birthday) (pp.1-13), Hatfield: University of Hertfordshire Press.

Hancock, I. (2001). On Ascertaining the Date of Departure of Pre-Romani Speakers from India. To appear in *Dzaniben*.

Harper, P. S., Williams, E. M., & Sunderland, E. (1977). Genetic markers in Welsh gypsies. *J Med Genet, 14*(3), 177-182.

Hastbacka, J., de la Chapelle, A., Kaitila, I., Sistonen, P., Weaver, A., & Lander, E. (1992). Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nat Genet, 2*(3), 204-211.

Hastbacka, J., de la Chapelle, A., Mahtani, M. M., Clines, G., Reeve-Daly, M. P., Daly, M., Hamilton, B. A., Kusumi, K., Trivedi, B., Weaver, A., Coloma, A., Lovett, M., Buckler, A., Kaitila, I., & Lander, E. S. (1994). The diastrophic dysplasia gene encodes a novel sulfate transporter: positional cloning by fine-structure linkage disequilibrium mapping. *Cell, 78*(6), 1073-1087.

Hastbacka, J., Kaitila, I., Sistonen, P., & de la Chapelle, A. (1990). Diastrophic dysplasia gene maps to the distal long arm of chromosome 5. *Proc Natl Acad Sci U S A, 87*(20), 8056-8059.

Hastbacka, J., Kerrebrock, A., Mokkala, K., Clines, G., Lovett, M., Kaitila, I., de la Chapelle, A., & Lander, E. S. (1999). Identification of the Finnish founder mutation for diastrophic dysplasia (DTD). *Eur J Hum Genet, 7*(6), 664-670.

Helgason, A., Sigurethardottir, S., Gulcher, J. R., Ward, R., & Stefansson, K. (2000). mtDNA and the origin of the Icelanders: deciphering signals of recent population history. *Am J Hum Genet, 66*(3), 999-1016.

Helgason, A., Sigurethardottir, S., Nicholson, J., Sykes, B., Hill, E. W., Bradley, D. G., Bosnes, V., Gulcher, J. R., Ward, R., & Stefansson, K. (2000). Estimating Scandinavian and Gaelic ancestry in the male settlers of Iceland. *Am J Hum Genet, 67*(3), 697-717.

Hey, J. (2000). Human mitochondrial DNA recombination: can it be true? *Tree, 15*(5), 181-182.

Heyer, E. (1999). One founder/one gene hypothesis in a new expanding population: Saguenay (Québec, Canada). *Hum Biol, 71*(1), 99-109.

Heyer, E., Puymirat, J., Dieltjes, P., Bakker, E., & de Knijff, P. (1997). Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. *Hum Mol Genet, 6*(5), 799-803.

Heyer, E., & Tremblay, M. (1995). Variability of the genetic contribution of Quebec population founders associated to some deleterious genes. *Am J Hum Genet, 56*(4), 970-978.

Heyer, E., Tremblay, M., & Desjardins, B. (1997). Seventeenth-century European origins of hereditary diseases in the Saguenay population (Québec, Canada). *Hum Biol, 69*(2), 209-225.

Hollox, E. J., Poulter, M., Zvarik, M., Ferak, V., Krause, A., Jenkins, T., Saha, N., Kozlov, A. I., & Swallow, D. M. (2001). Lactase haplotype diversity in the Old World. *Am J Hum Genet, 68*(1), 160-172.

Holtzman, N. A. (1989). *Proceed With Caution: Predicting Genetic Risks in the Recombinant DNA Era*. Baltimore: The Johns Hopkins University Press.

Holtzman, N. A., Murphy, P. D., Watson, M. S., & Barr, P. A. (1997). Predictive genetic testing: from basic research to clinical practice. *Science, 278*(5338), 602-605.

Horai, S., Hayasaka, K., Kondo, R., Tsugane, K., & Takahata, N. (1995). Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. *Proc Natl Acad Sci U S A, 92*(2), 532-536.

Houwen, R. H., Baharloo, S., Blankenship, K., Raeymaekers, P., Juyn, J., Sandkuijl, L. A., & Freimer, N. B. (1994). Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. *Nat Genet, 8*(4), 380-386.

Howell, N., & Smejkal, C. B. (2000). Persistent heteroplasmy of a mutation in the human mtDNA control region: hypermutation as an apparent consequence of simple-repeat expansion/contraction. *Am J Hum Genet, 66*(5), 1589-1598.

Hunter, M., Angelicheva, D., Levy, H. L., Pueschel, S. M., & Kalaydjieva, L. (2001). Novel mutations in the GALK1 gene in patients with galactokinase deficiency. *Hum Mutat, 17*(1), 77-78.

Hurles, M. E., Veitia, R., Arroyo, E., Armenteros, M., Bertranpetit, J., Pérez-Lezaun, A., Bosch, E., Shlumukova, M., Cambon-Thomsen, A., McElreavey, K., Lopez De Munain, A., Rohl, A., Wilson, I. J., Singh, L., Pandya, A., Santos, F. R., Tyler-Smith, C., & Jobling, M. A. (1999). Recent male-mediated gene flow over a linguistic barrier in Iberia, suggested by analysis of a Y-chromosomal DNA polymorphism. *Am J Hum Genet, 65*(5), 1437-1448.

Jeffrey, G. P., Chakrabarti, S., Hegele, R. A., & Adams, P. C. (1999). Polymorphism in intron 4 of HFE may cause overestimation of C282Y homozygote prevalence in haemochromatosis. *Nat Genet, 22*(4), 325-326.

Jobling, M. A., Bouzekri, N., & Taylor, P. G. (1998). Hypervariable digital DNA codes for human paternal lineages: MVR-PCR at the Y-specific minisatellite, MSY1 (DYF155S1). *Hum Mol Genet, 7*(4), 643-653.

Jobling, M. A., & Tyler-Smith, C. (1995). Fathers and sons: the Y chromosome and human evolution. *Trends Genet, 11*(11), 449-456.

Jobling, M. A., & Tyler-Smith, C. (2000). New uses for new haplotypes: the human Y chromosome, disease and selection. *Trends Genet, 16*(8), 356-362.

Jobling, M. A., Williams, G. A., Schiebel, G. A., Pandya, G. A., McElreavey, G. A., Salas, G. A., Rappold, G. A., Affara, N. A., & Tyler-Smith, C. (1998). A selective difference between human Y-chromosomal DNA haplotypes. *Curr Biol, 8*(25), 1391-1394.

Jorde, L. B., & Bamshad, M. (2000). Questioning evidence for recombination in human mitochondrial DNA. *Science, 288*(5473), 1931.

Jorde, L. B., Watkins, W. S., Kere, J., Nyman, D., & Eriksson, A. W. (2000). Gene mapping in isolated populations: new roles for old friends? *Hum Hered, 50*(1), 57-65.

Juengst, E. T. (1998). Group identity and human diversity: keeping biology straight from culture. *Am J Hum Genet, 63*(3), 673-677.

Kaback, M., Lim-Steele, J., Dabholkar, D., Brown, D., Levy, N., & Zeiger, K. (1993). Tay-Sachs disease--carrier screening, prenatal diagnosis, and the molecular era. An international perspective, 1970 to 1993. The International TSD Data Collection Network. *J Amer Med Ass, 270*(19), 2307-2315.

Kalanin, J., Takarada, Y., Kagawa, S., Yamashita, K., Ohtsuka, N., & Matsuoka, A. (1994). Gypsy phenylketonuria: a point mutation of the phenylalanine hydroxylase gene in Gypsy families from Slovakia. *Am J Med Genet, 49*(2), 235-239.

Kalaydjieva, L., Calafell, F., Jobling, M. A., Angelicheva, D., de Knijff, P., Rosser, Z. H., Hurles, M. E., Underhill, P., Tournev, I., Marushiakova, E., & Popov, V. (2001). Patterns of inter- and intra-group genetic diversity in the Vlax Roma as revealed by Y chromosome and mitochondrial DNA lineages. *Eur J Hum Genet, 9*(2), 97-104.

Kalaydjieva, L., Dworniczak, B., Kremensky, I., Koprivarova, K., Radeva, B., Milusheva, R., Aulehla-Scholz, C., & Horst, J. (1992). Heterogeneity of mutations in Bulgarian phenylketonuria haplotype 1 and 4 alleles. *Clin Genet, 41*(3), 123-128.

Kalaydjieva, L., Gresham, D., & Calafell, F. (2001). Genetic studies of the Roma (Gypsies): a review. *BMC Med Genet, 2*(1), 5.

Kalaydjieva, L., Gresham, D., Gooding, R., Heather, L., Baas, F., de Jonge, R., Blechschmidt, K., Angelicheva, D., Chandler, D., Worsley, P., Rosenthal, A., King, R. H., & Thomas, P. K. (2000). N-myc downstream-regulated gene 1 is mutated in hereditary motor and sensory neuropathy-Lom. *Am J Hum Genet, 67*(1), 47-58.

Kalaydjieva, L., Hallmayer, J., Chandler, D., Savov, A., Nikolova, A., Angelicheva, D., King, R. H., Ishpekova, B., Honeyman, K., Calafell, F., Shmarov, A., Petrova, J., Turnev, I., Hristova, A., Moskov, M., Stancheva, S., Petkova, I., Bittles, A. H., Georgieva, V., Middleton, L., & Thomas, P. K. (1996). Gene mapping in Gypsies identifies a novel demyelinating neuropathy on chromosome 8q24. *Nat Genet, 14*(2), 214-217.

Kalaydjieva, L., Nikolova, A., Turnev, I., Petrova, J., Hristova, A., Ishpekova, B., Petkova, I., Shmarov, A., Stancheva, S., Middleton, L., Merlini, L., Trogu, A., Muddle, J. R., King, R. H., & Thomas, P. K. (1998). Hereditary motor and sensory neuropathy--Lom, a novel demyelinating neuropathy associated with deafness in gypsies. Clinical, electrophysiological and nerve biopsy findings. *Brain, 121*(3), 399-408.

Kalaydjieva, L., Péréz-Lezaun, A., Angelicheva, D., Onengut, S., Dye, D., Bosshard, N. U., Jordanova, A., Savov, A., Yanakiev, P., Kremensky, I., Radeva, B., Hallmayer, J., Markov, A., Nedkova, V., Tournev, I., Aneva, L., & Gitzelmann, R. (1999). A founder mutation in the GK1 gene is responsible for galactokinase deficiency in Roma (Gypsies). *Am J Hum Genet, 65*(5), 1299-1307.

Kaplan, F. (1998). Tay-Sachs disease carrier screening: a model for prevention of genetic disease. *Genet Test, 2*(4), 271-292.

Kaufman, T. (1984). *Explorations in Proto-Gypsy Phonology and Classifications.* Paper presented at the Sixth South Asian Languages Analysis Round-Table, Austin.

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., Péréz-Lezaun, A., Piccinini, A., Prinz, M., Schmitt, C., & Roewer, L., (1997). Evaluation of Y-chromosomal STRs: a multicenter study. *Int J Legal Med, 110*(3), 125-133.

Kayser, M., Roewer, L., Hedman, M., Henke, L., Henke, J., Brauer, S., Kruger, C., Krawczak, M., Nagy, M., Dobosz, T., Szibor, R., de Knijff, P., Stoneking, M., & Sajantila, A. (2000). Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. *Am J Hum Genet, 66*(5), 1580-1588.

King, R. H., Tournev, I., Colomer, J., Merlini, L., Kalaydjieva, L., & Thomas, P. K. (1999). Ultrastructural changes in peripheral nerve in hereditary motor and sensory neuropathy-Lom. *Neuropathol Appl Neurobiol, 25*(4), 306-312.

Kittles, R. A., Bergen, A. W., Urbanek, M., Virkkunen, M., Linnoila, M., Goldman, D., & Long, J. C. (1999). Autosomal, mitochondrial, and Y chromosome DNA variation in Finland: evidence for a male-specific bottleneck. *Am J Phys Anthropol, 108*(4), 381-399.

Kittles, R. A., Perola, M., Peltonen, L., Bergen, A. W., Aragon, R. A., Virkkunen, M., Linnoila, M., Goldman, D., & Long, J. C. (1998). Dual origins of Finns revealed by Y chromosome haplotype variation. *Am J Hum Genet, 62*(5), 1171-1179.

Kivisild, T., & Villems, R. (2000). Questioning evidence for recombination in human mitochondrial DNA. *Science, 288*(5473), 1931.

Kivisild, T., Bamshad, M. J., Kaldma, K., Metspalu, M., Metspalu, E., Reidla, M., Laos, S., Parik, J., Watkins, W. S., Dixon, M. E., Papiha, S. S., Mastana, S. S., Mir, M. R., Ferak, V., & Villems, R. (1999). Deep common ancestry of Indian and western-Eurasian mitochondrial DNA lineages. *Curr Biol, 9*(22), 1331-1334.

Kohn, M. (1996). *The Race Gallery*. London: Vintage.

Kokame, K., Kato, H., & Miyata, T. (1996). Homocysteine-respondent genes in vascular endothelial cells identified by differential display analysis. GRP78/BiP and novel genes. *J Biol Chem, 271*(47), 29659-29665.

Kolodny, E. H. (1992). Carrier testing for lysosomal diseases: problems and prospects. In B. Bonné-Tamir & A. Adams (Eds.), *Genetic Diversity Among Jews: diseases and markers at the DNA level*. (pp 333-345) Oxford: Oxford University Press.

Kondo, R., Matsuura, E. T., & Chigusa, S. I. (1992). Further observation of paternal transmission of *Drosophila* mitochondrial DNA by PCR selective amplification method. *Genet Res, 59*(2), 81-84.

Krauter-Canham, R., Bronner, R., Evrard, J.-L., Hahne, G., Friedt, W., & Steinmetz, A. (1997). A transmitting tissue- and pollen-expressed protein from sunflower with sequence similarity to the human RTP protein. *Plant Science, 129*, 191-202.

Kruglyak, L. (1999a). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet, 22*(2), 139-144.

Kruglyak, L. (1999b). Genetic isolates: separate but equal? *Proc Natl Acad Sci U S A, 96*(4), 1170-1172.

Kruglyak, L., Daly, M. J., & Lander, E. S. (1995). Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping. *Am J Hum Genet, 56*(2), 519-527.

Kumar, S., Hedrick, P., Dowling, T., & Stoneking, M. (2000). Questioning evidence for recombination in human mitochondrial DNA. *Science, 288*(5473), 1931.

Kurdistani, S. K., Arizti, P., Reimer, C. L., Sugrue, M. M., Aaronson, S. A., & Lee, S. W. (1998). Inhibition of tumor cell growth by RTP/rit42 and its responsiveness to p53 and DNA damage. *Cancer Res, 58*(19), 4439-4444.

Laan, M., & Paabo, S. (1997). Demographic history and linkage disequilibrium in human populations. *Nat Genet, 17*(4), 435-438.

Labuda, M., Labuda, D., Korab-Laskowska, M., Cole, D. E., Zietkiewicz, E., Weissenbach, J., Popowska, E., Pronicka, E., Root, A. W., & Glorieux, F. H. (1996). Linkage disequilibrium analysis in young populations: pseudo-vitamin D-deficiency rickets and the founder effect in French Canadians. *Am J Hum Genet, 59*(3), 633-643.

Lahn, B. T., & Page, D. C. (1997). Functional coherence of the human Y chromosome. *Science, 278*(5338), 675-680.

Lander, E. S., & Botstein, D. (1987). Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science, 236*(4808), 1567-1570.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., & Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature, 409*(6822), 860-921.

Lasa, A., Piccolo, F., de Diego, C., Jeanpierre, M., Colomer, J., Rodriguez, M. J., Urtizberea, J. A., Baiget, M., Kaplan, J., & Gallano, P. (1998). Severe limb girdle muscular dystrophy in Spanish gypsies: further evidence for a founder mutation in the gamma-sarcoglycan gene. *Eur J Hum Genet, 6*(4), 396-399.

Lee, N., Daly, M. J., Delmonte, T., Lander, E. S., Xu, F., Hudson, T. J., Mitchell, G. A., Morin, C. C., Robinson, B. H., & Rioux, J. D. (2001). A genomewide linkage-disequilibrium scan localizes the Saguenay-Lac-Saint-Jean cytochrome oxidase deficiency to 2p16. *Am J Hum Genet, 68*(2), 397-409.

Lee, R. (1998). *The Roma: origin and diaspora*, [Webpage]. Romani.org. Available: http://romani.org/toronto/diaspora_rl.html [accessed 1999].

LeGuern, E., Guilbot, A., Kessali, M., Ravise, N., Tassin, J., Maisonobe, T., Grid, D., & Brice, A. (1996). Homozygosity mapping of an autosomal recessive form of demyelinating Charcot-Marie-Tooth disease to chromosome 5q23-q33. *Hum Mol Genet, 5*(10), 1685-1688.

Lehesjoki, A. E., Koskiniemi, M., Norio, R., Tirrito, S., Sistonen, P., Lander, E., & de la Chapelle, A. (1993). Localization of the EPM1 gene for progressive myoclonus epilepsy on chromosome 21: linkage disequilibrium allows high resolution mapping. *Hum Mol Genet, 2*(8), 1229-1234.

Lehesjoki, A. E., Koskiniemi, M., Sistonen, P., Miao, J., Hastbacka, J., Norio, R., & de la Chapelle, A. (1991). Localization of a gene for progressive myoclonus epilepsy to chromosome 21q22. *Proc Natl Acad Sci U S A, 88*(9), 3696-3699.

Leonard, C. O., Chase, G. A., & Childs, B. (1972). Genetic counseling: a consumers' view. *N Engl J Med, 287*(9), 433-439.

Li, M., Dickson, D. W., & Spiro, A. J. (1998). Sarcolemmal defect and subsarcolemmal lesion in a patient with gamma- sarcoglycan deficiency. *Neurology, 50*(3), 807-809.

Liégeois, J.-P. (1994). *Roma, Gypsies, Travellers*. Strasbourg: Council of Europe Press.

Lucotte, G., & Ngo, N. Y. (1985). p49f, A highly polymorphic probe, that detects Taq1 RFLPs on the human Y chromosome. *Nucleic Acids Res, 13*(22), 8285.

Lupski, J. R. (2000). Recessive Charcot-Marie-tooth disease. *Ann Neurol, 47*(1), 6-8.

Macaulay, V., Richards, M., Hickey, E., Vega, E., Cruciani, F., Guida, V., Scozzari, R., Bonné-Tamir, B., Sykes, B., & Torroni, A. (1999). The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. *Am J Hum Genet, 64*(1), 232-249.

Malyarchuk, B. A., & Derenko, M. V. (1999). Molecular instability of the mitochondrial haplogroup T sequences at nucleotide positions 16292 and 16296. *Ann Hum Genet, 63*(6), 489-497.

Malaspina, P., Cruciani, F., Ciminelli, B. M., Terrenato, L., Santolamazza, P., Alonso, A., Banyko, J., Brdicka, R., Garcia, O., Gaudiano, C., Guanti, G., Kidd, K. K., Lavinha, J., Avila, M., Mandich, P., Moral, P., Qamar, R., Mehdi, S. Q., Ragusa, A., Stefanescu, G., Caraghin, M., Tyler-Smith, C., Scozzari, R., & Novelletto, A. (1998). Network analyses of Y-chromosomal types in Europe, northern Africa, and western Asia reveal specific patterns of geographic distribution. *Am J Hum Genet, 63*(3), 847-860.

Malyarchuk, B. A., & Derenko, M. V. (1999). Molecular instability of the mitochondrial haplogroup T sequences at nucleotide positions 16292 and 16296. *Ann Hum Genet, 63*(6), 489-497.

Martinez, G., Garcia-Lozano, J. R., Ribes, A., Maldonado, M. D., Baldellou, A., de Pablo, R., & Nunez-Roldan, A. (1998). High risk of medium chain acyl-coenzyme A dehydrogenase deficiency among gypsies. *Pediatr Res, 44*(1), 83-84.

Marushiakova, E., & Popov, V. (1997). *Gypsies (Roma) in Bulgaria*. Frankfurt am Main: Peter Lang.

Mastana, S. S., & Papiha, S. S. (1992). Origin of the Romany gypsies--genetic evidence. *Z Morphol Anthropol, 79*(1), 43-51.

Mateu, E., Calafell, F., Lao, O., Bonné-Tamir, B., Kidd, J. R., Pakstis, A., Kidd, K. K., & Bertranpetit, J. (2001). Worldwide genetic analysis of the CFTR region. *Am J Hum Genet, 68*(1), 103-117.

Mateu, E., Comas, D., Calafell, F., Péréz-Lezaun, A., Abade, A., & Bertranpetit, J. (1997). A tale of two islands: population history and mitochondrial DNA sequence variation of Bioko and Sao Tome, Gulf of Guinea. *Ann Hum Genet, 61*(6), 507-518.

McNally, E. M., Passos-Bueno, M. R., Bonnemann, C. G., Vainzof, M., de Sa Moreira, E., Lidov, H. G., Othmane, K. B., Denton, P. H., Vance, J. M., Zatz, M., & Kunkel, L. M. (1996). Mild and severe muscular dystrophy caused by a single gamma-sarcoglycan mutation. *Am J Hum Genet, 59*(5), 1040-1047.

Meijerink, P. H., Yanakiev, P., Zorn, I., Grierson, A. J., Bikker, H., Dye, D., Kalaydjieva, L., & Baas, F. (1998). The gene for the human Src-like adaptor protein (hSLAP) is located within the 64-kb intron of the thyroglobulin gene. *Eur J Biochem, 254*(2), 297-303.

Merlini, L., Kaplan, J. C., Navarro, C., Barois, A., Bonneau, D., Brasa, J., Echenne, B., Gallano, P., Jarre, L., Jeanpierre, M., Kalaydjieva, L., Leturcq, F., Levi-Gomes, A., Toutain, A., Tournev, I., Urtizberea, A., Vallat, J. M., Voit, T., & Warter, J. M. (2000). Homogeneous phenotype of the gypsy limb-girdle MD with the gamma-sarcoglycan C283Y mutation. *Neurology, 54*(5), 1075-1079.

Merlini, L., Villanova, M., Sabatelli, P., Trogu, A., Malandrini, A., Yanakiev, P., Maraldi, N. M., & Kalaydjieva, L. (1998). Hereditary motor and sensory neuropathy Lom type in an Italian Gypsy family. *Neuromuscul Disord, 8*(3-4), 182-185.

Merryweather-Clarke, A. T., Pointon, J. J., Shearman, J. D., Robson, K. J., Jouanolle, A. M., Mosser, A., David, V., Le Gall, J. Y., Halsall, D. J., Elsey, T. S., Kelly, A., Cox, T. M., Clare, M., Bomford, A., Vandwalle, J. L., Rochette, J., Borot, N., Coppin, H., Roth, M. P., Ryan, E., Crowe, J., Totaro, A., Gasparini, P., Roetto, A., Camaschella, C., Darke, C., Wallace, D. F., Saeb-Parsy, K., Dooley, J. S., Worwood, M., & Walker, A. P. (1999). Polymorphism in intron 4 of HFE does not compromise haemochromatosis mutation results. The European Haemochromatosis Consortium. *Nat Genet, 23*(3), 271.

Mesa, N. R., Mondragon, M. C., Soto, I. D., Parra, M. V., Duque, C., Ortiz-Barrientos, D., Garcia, L. F., Velez, I. D., Bravo, M. L., Munera, J. G., Bedoya, G., Bortolini, M. C., & Ruiz-Linares, A. (2000). Autosomal, mtDNA, and Y-chromosome diversity in Amerinds: pre- and post-Columbian patterns of gene flow in South America. *Am J Hum Genet, 67*(5), 1277-1286.

Meyer, S., Weiss, G., & von Haeseler, A. (1999). Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA. *Genetics, 152*(3), 1103-1110.

Miano, M. G., Jacobson, S. G., Carothers, A., Hanson, I., Teague, P., Lovell, J., Cideciyan, A. V., Haider, N., Stone, E. M., Sheffield, V. C., & Wright, A. F. (2000). Pitfalls in homozygosity mapping. *Am J Hum Genet, 67*(5), 1348-1351.

Michie, S., Bron, F., Bobrow, M., & Marteau, T. M. (1997). Nondirectiveness in genetic counseling: an empirical study. *Am J Hum Genet, 60*(1), 40-47.

Modell, B., & Kuliev, A. M. (1998). The history of community genetics: the contribution of the haemoglobin disorders. *Community Genetics, 1*, 3-11.

Morral, N., Bertranpetit, J., Estivill, X., Nunes, V., Casals, T., Gimenez, J., Reis, A., Varon-Mateeva, R., Macek, M., Jr., Kalaydjieva, L., Angelicheva, D., Dancheva, R., Romeo, G., Russo, M. P., Garnerone, S., Restagno, G., Ferrari, M., Magnani, C., Claustres, M., Desgeorges, M., Schwartz, M., Schwarz, M., Dallapiccola, B., Novelli, G., Ferec, C., de Arce, M., Nemeti, M., Kere, J., Anvret, M., Dahl, N., & Kadasi, L. (1994). The origin of the major cystic fibrosis mutation (delta F508) in European populations. *Nat Genet, 7*(2), 169-175.

Morton, N. E. (1955). Sequential Tests for the Detection of Linkage. *Am J Hum Genet, 7*, 277-318.

Motulsky, A. G. (1995). Jewish diseases and origins. *Nat Genet, 9*(2), 99-101.

Mountain, J. L., Hebert, J. M., Bhattacharyya, S., Underhill, P. A., Ottolenghi, C., Gadgil, M., & Cavalli-Sforza, L. L. (1995). Demographic history of India and mtDNA-sequence diversity. *Am J Hum Genet, 56*(4), 979-992.

Mourant, A. E., Kopec, A. C., & Domaniewska-Sobczak, K. (1976). *The Distribution of the Human Blood Groups and Other Polymorphisms* (second ed.). London: Oxford University Press.

Muller-Hill, B. (1998). *Murderous Science: elimination by scientific selection of Jews, Gypsies and Others in Germany, 1933-1945* (G. R. Fraser, Trans.). New York: Cold Spring Harbour Laboratory Press.

Nakahori, Y., Hamano, K., Iwaya, M., & Nakagome, Y. (1991). Sex identification by polymerase chain reaction using X-Y homologous primer. *Am J Med Genet, 39*(4), 472-473.

Nei, M. (1987). *Molecular Evolutionary Genetics*. New York: Columbia University Press.

Noguchi, S., McNally, E. M., Ben Othmane, K., Hagiwara, Y., Mizuno, Y., Yoshida, M., Yamamoto, H., Bonnemann, C. G., Gussoni, E., Denton, P. H., Kyriakides, T., Middleton, L., Hentati, F., Ben Hamida, C., Nonaka, I., Vance, J. M., Kunkel, L. M., & Ozawa, E. (1995). Mutations in the dystrophin-associated protein gamma-sarcoglycan in chromosome 13 muscular dystrophy. *Science, 270*(5237), 819-822.

Noll, W. W., Belloni, D. R., Stenzel, T. T., & Grody, W. W. (1999). Polymorphism in intron 4 of HFE does not compromise haemochromatosis mutation results. *Nat Genet, 23*(3), 271-272.

Norio, R., Nevanlinna, H. R., & Perheentupa, J. (1973). Hereditary diseases in Finland; rare flora in rare soul. *Ann Clin Res, 5*(3), 109-141.

Okely, J. (1983). *The traveller-gypsies*. Cambridge: Cambridge University Press.

OMIM. (2001). *Online Mendelian Inheritance in Man*. McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University. Available: http://www.ncbi.nlm.nih.gov/omim/ [accessed 2001].

Ott, J. (1991). *Analysis of Human Genetic Linkage* (Revised ed.). Baltimore: The Johns Hopkins University Press.

Ott, J., & Hoh, J. (2000). Statistical approaches to gene mapping. *Am J Hum Genet, 67*(2), 289-294.

Page, R. D. (1996). TREEVIEW: An application to display phylogenetic trees on personal computers. *Comp Apps Biosci, 12*, 357-358.

Parsons, T. J., & Irwin, J. A. (2000). Questioning evidence for recombination in human mitochondrial DNA. *Science, 288*(5473), 1931.

Parsons, T. J., Muniec, D. S., Sullivan, K., Woodyatt, N., Alliston-Greiner, R., Wilson, M. R., Berry, D. L., Holland, K. A., Weedn, V. W., Gill, P., & Holland, M. M. (1997). A high observed substitution rate in the human mitochondrial DNA control region. *Nat Genet, 15*(4), 363-368.

Passarino, G., Semino, O., Bernini, L. F., & Santachiara-Benerecetti, A. S. (1996). Pre-Caucasoid and Caucasoid features of the Indian population revealed by mtDNA polymorphisms. *Am J Hum Genet, 59*(4), 927-934.

Passarino, G., Semino, O., Quintana-Murci, L., Excoffier, L., Hammer, M., & Santachiara-Benerecetti, A. S. (1998). Different genetic components in the Ethiopian population, identified by mtDNA and Y-chromosome polymorphisms. *Am J Hum Genet, 62*(2), 420-434.

Patrin. (1999). *Timeline of Romani History*. The Patrin Web Journal. Available: http://www.geocities.com/Paris/5121/timeline.htm.

Pekkarinen, P., Hovatta, I., Hakola, P., Jarvi, O., Kestila, M., Lenkkeri, U., Adolfsson, R., Holmgren, G., Nylander, P. O., Tranebjaerg, L., Terwillinger, J. D., Lonnqvist, J., & Peltonen, L. (1998). Assignment of the locus for PLO-SL, a frontal-lobe dementia with bone cysts, to 19q13. *Am J Hum Genet, 62*(2), 362-372.

Pekkarinen, P., Kestila, M., Paloneva, J., Terwillign, J., Varilo, T., Jarvi, O., Hakola, P., & Peltonen, L. (1998). Fine-scale mapping of a novel dementia gene, PLOSL, by linkage disequilibrium. *Genomics, 54*(2), 307-315.

Peltonen, L. (2000). Positional cloning of disease genes: advantages of genetic isolates. *Hum Hered, 50*(1), 66-75.

Peltonen, L., Jalanko, A., & Varilo, T. (1999). Molecular genetics of the Finnish disease heritage. *Hum Mol Genet, 8*(10), 1913-1923.

Peltonen, L., Palotie, A., & Lange, K. (2000). Use of population isolates for mapping complex traits. *Nat Rev Genet, 1*(3), 182-190.

Pennica, D., Swanson, T. A., Welsh, J. W., Roy, M. A., Lawrence, D. A., Lee, J., Brush, J., Taneyhill, L. A., Deuel, B., Lew, M., Watanabe, C., Cohen, R. L., Melhem, M. F., Finley, G. G., Quirke, P., Goddard, A. D., Hillan, K. J., Gurney, A. L., Botstein, D., & Levine, A. J. (1998). WISP genes are members of the connective tissue growth factor family that are up-regulated in wnt-1-transformed cells and aberrantly expressed in human colon tumors. *Proc Natl Acad Sci U S A, 95*(25), 14717-14722.

Pérez-Lezaun, A., Calafell, F., Seielstad, M., Mateu, E., Comas, D., Bosch, E., & Bertranpetit, J. (1997). Population genetics of Y-chromosome short tandem repeats in humans. *J Mol Evol, 45*(3), 265-270.

Pfaff, C. L., Parra, E. J., Bonilla, C., Hiester, K., McKeigue, P. M., Kamboh, M. I., Hutchinson, R. G., Ferrell, R. E., Boerwinkle, E., & Shriver, M. D. (2001). Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. *Am J Hum Genet, 68*(1), 198-207.

"Petulengro". (1915-16). Report on the Gypsy Tribes of North-East Bulgaria. *J Gyp Lor Soc, 9*(1), 1-109.

Piccolo, F., Jeanpierre, M., Leturcq, F., Dodé, C., Azibi, K., Toutain, A., Merlini, L., Jarre, L., Navarro, C., Krishnamoorthy, R., Tome, F. M., Urtizberea, J. A., Beckmann, J. S., Campbell, K. P., & Kaplan, J. -C. (1996). A founder mutation in the gamma-sarcoglycan gene of gypsies possibly predating their migration out of India. *Hum Mol Genet, 5*(12), 2019-2022.

Piquemal, D., Joulia, D., Balaguer, P., Basset, A., Marti, J., & Commes, T. (1999). Differential expression of the RTP/Drg1/Ndr1 gene product in proliferating and growth arrested cells. *Biochim Biophys Acta, 1450*(3), 364-373.

Plasilova, M., Stoilov, I., Sarfarazi, M., Kadasi, L., Ferakova, E., & Ferak, V. (1999). Identification of a single ancestral CYP1B1 mutation in Slovak Gypsies (Roms) affected with primary congenital glaucoma. *J Med Genet, 36*(4), 290-294.

Puffenberger, E. G., Kauffman, E. R., Bolk, S., Matise, T. C., Washington, S. S., Angrist, M., Weissenbach, J., Garver, K. L., Mascari, M., Ladda, R., & et al. (1994). Identity-by-descent and association mapping of a recessive gene for Hirschsprung disease on human chromosome 13q22. *Hum Mol Genet, 3*(8), 1217-1225.

Quintana-Murci, L., Semino, O., Bandelt, H. J., Passarino, G., McElreavey, K., & Santachiara-Benerecetti, A. S. (1999). Genetic evidence of an early exit of *Homo sapiens sapiens* from Africa through eastern Africa. *Nat Genet, 23*(4), 437-441.

Rawlinson, H. G. (1975). Early contacts between India and Europe. In A. L. Basham (Ed.), *A Cultural History of Europe* (pp. 425-441). Delhi: Oxford University Press.

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R., & Lander, E. S. (2001). Linkage disequilibrium in the human genome. *Nature, 411*(6834), 199-204.

Rex-Kiss, B., Szabo, L., Szabo, S., & Hartmann, E. (1973). ABO, MN, Rh blood groups, Hp types and Hp level, Gm(1) factor investigations on the Gypsy population of Hungary. *Hum Biol, 45*(1), 41-61.

Riazuddin, S., Castelein, C. M., Ahmed, Z. M., Lalwani, A. K., Mastroianni, M. A., Naz, S., Smith, T. N., Liburd, N. A., Friedman, T. B., Griffith, A. J., & Wilcox, E. R. (2000). Dominant modifier DFNM1 suppresses recessive deafness DFNB26. *Nat Genet, 26*(4), 431-434.

Richards, M., Corte-Real, H., Forster, P., Macaulay, V., Wilkinson-Herbots, H., Demaine, A., Papiha, S., Hedges, R., Bandelt, H. J., & Sykes, B. (1996). Paleolithic and neolithic lineages in the European mitochondrial gene pool. *Am J Hum Genet, 59*(1), 185-203.

Richards, M., Macaulay, V., Hickey, E., Vega, E., Sykes, B., Guida, V., Rengo, C., Sellitto, D., Cruciani, F., Kivisild, T., Villems, R., Thomas, M., Rychkov, S., Rychkov, O., Rychkov, Y., Golge, M., Dimitrov, D., Hill, E., Bradley, D., Romano, V., Cali, F., Vona, G., Demaine, A., Papiha, S., Triantaphyllidis, C., & Stefanescu, G. (2000). Tracing European founder lineages in the Near Eastern mtDNA pool. *Am J Hum Genet, 67*(5), 1251-1276.

Richards, M. B., Macaulay, V. A., Bandelt, H. J., & Sykes, B. C. (1998). Phylogeography of mitochondrial DNA in western Europe. *Ann Hum Genet, 62*(3), 241-260.

Richter, A., Rioux, J. D., Bouchard, J. P., Mercier, J., Mathieu, J., Ge, B., Poirier, J., Julien, D., Gyapay, G., Weissenbach, J., Hudson, T. J., Melancon, S. B., & Morgan, K. (1999). Location score and haplotype analyses of the locus for autosomal recessive spastic ataxia of Charlevoix-Saguenay, in chromosome region 13q11. *Am J Hum Genet, 64*(3), 768-775.

Risch, N. J. (2000). Searching for genetic determinants in the new millennium. *Nature, 405*(6788), 847-856.

Risch, N., de Leon, D., Ozelius, L., Kramer, P., Almasy, L., Singer, B., Fahn, S., Breakefield, X., & Bressman, S. (1995). Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population. *Nat Genet, 9*(2), 152-159.

Rishi, P. W. R. (1976). *Roma: The Panjabi Emigrants in Europe, Central and Middle Asia, The USSR and The Americas*. Patiala: Punjabi University.

Rolf, B., Meyer, E., Brinkmann, B., & de Knijff, P. (1998). Polymorphism at the tetranucleotide repeat locus DYS389 in 10 populations reveals strong geographic clustering. *Eur J Hum Genet, 6*(6), 583-588

Rogers, T., Chandler, D., Angelicheva, D., Thomas, P. K., Youl, B., Tournev, I., Gergelcheva, V., & Kalaydjieva, L. (2000). A novel locus for autosomal recessive peripheral neuropathy in the EGR2 region on 10q23. *Am J Hum Genet, 67*(3), 664-671.

Rosenberg, N. A., Woolf, E., Pritchard, J. K., Schaap, T., Gefel, D., Shpirer, I., Lavi, U., Bonné-Tamir, B., Hillel, J., & Feldman, M. W. (2001). Distinctive genetic signatures in the Libyan Jews. *Proc Natl Acad Sci U S A, 98*(3), 858-863.

Rothenberg, K. H., & Rutkin, A. B. (1998). Toward a Framework of Mutualism: the Jewish community in genetic research. *Community Genetics, 1*, 148-153.

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

Salas, A., Comas, D., Lareu, M. V., Bertranpetit, J., & Carracedo, A. (1998). mtDNA analysis of the Galician population: a genetic edge of European variation. *Eur J Hum Genet, 6*(4), 365-375

Sandland, R. (1996). The real, the simulcram, and the construction of 'gypsy' in law. *J Law Soc, 23*(3), 383-405.

Santachiara Benerecetti, A. S., Semino, O., Passarino, G., Torroni, A., Brdicka, R., Fellous, M., & Modiano, G. (1993). The common, Near-Eastern origin of Ashkenazi and Sephardi Jews supported by Y-chromosome similarity. *Ann Hum Genet, 57*(Pt 1), 55-64.

Santos, F. R., Pandya, A., Kayser, M., Mitchell, R. J., Liu, A., Singh, L., Destro-Bisol, G., Novelletto, A., Qamar, R., Mehdi, S. Q., Adhikari, R., de Knijff, P., & Tyler-Smith, C. (2000). A polymorphic L1 retroposon insertion in the centromere of the human Y chromosome. *Hum Mol Genet, 9*(3), 421-430.

Schegel, N., Gayet, O., Morel-Kopp, M. C., Wyler, B., Hurtaud-Roux, M. F., Kaplan, C., & MacGregor, J. (1994). The molecular basis of Glanzmann thrombasthenia (GT) in a Gypsy population in France. Identification of a new mutation of the IIb gene. *Blood, 84*, 477a.

Scherer, S. (1999). Axonal pathology in demyelinating diseases. *Ann Neurol, 45*(1), 6-7.

Schneider, S., Kueffer, J. M., Roessli, D., & Excoffier, L. (1996). Arlequin: a software environment for the analysis of population genetic data: Genetics and Biometry Lab, University of Geneva, Switzerland.

Scriver, C. R. (1992). What are genes like that doing in a place like this? Human history and molecular prosopography. In B. Bonné-Tamir & A. Adam (Eds.), *Genetic Diversity Among Jews: disease and markers at the DNA level*. (pp 319-332) Oxford: Oxford University Press.

Scriver, C. R., Bardanis, M., Cartier, L., Clow, C. L., Lancaster, G. A., & Ostrowsky, J. T. (1984). Beta-thalassemia disease prevention: genetic medicine applied. *Am J Hum Genet, 36*(5), 1024-1038.

Scriver, C. R., & Fujiwara, T. M. (1992). Cystic fibrosis genotypes and views on screening are both heterogeneous and population related. *Am J Hum Genet, 51*(5), 943-950.

Scriver, C. R., & Waters, P. J. (1999). Monogenic traits are not simple: lessons from phenylketonuria. *Trends Genet, 15*(7), 267-272.

Seielstad, M. T., Minch, E., & Cavalli-Sforza, L. L. (1998). Genetic evidence for a higher female migration rate in humans. *Nat Genet, 20*(3), 278-280.

Semino, O., Passarino, G., Oefner, P. J., Lin, A. A., Arbuzova, S., Beckman, L. E., De Benedictis, G., Francalacci, P., Kouvatsi, A., Limborska, S., Marcikiae, M., Mika, A., Mika, B., Primorac, D., Santachiara-Benerecetti, A. S., Cavalli-Sforza, L. L., & Underhill, P. A. (2000). The genetic legacy of Paleolithic *Homo sapiens sapiens* in extant Europeans: a Y chromosome perspective. *Science, 290*(5494), 1155-1159.

Sheffield, V. C., Nishimura, D. Y., & Stone, E. M. (1995). Novel approaches to linkage mapping. *Curr Opin Genet Dev, 5*(3), 335-341.

Sigurgardottir, S., Helgason, A., Gulcher, J. R., Stefansson, K., & Donnelly, P. (2000). The mutation rate in the human mtDNA control region. *Am J Hum Genet, 66*(5), 1599-1609.

Simoni, L., Calafell, F., Pettener, D., Bertranpetit, J., & Barbujani, G. (2000a). Geographic patterns of mtDNA diversity in Europe. *Am J Hum Genet, 66*(1), 262-278.

Simoni, L., Calafell, F., Pettener, D., Bertranpetit, J., & Barbujani, G. (2000b). Reconstruction of prehistory on the basis of genetic data. *Am J Hum Genet, 66*(3), 1177-1179.

Siren, M. K., Sareneva, H., Lokki, M. L., & Koskimies, S. (1996). Unique HLA antigen frequencies in the Finnish population. *Tissue Antigens, 48*(6), 703-707.

Sivakova, D. (1983). Distribution of three red-cell enzyme polymorphisms (ACP, PGM1 and AK) in gypsies from Slovakia (Czechoslovakia). *Ann Hum Biol, 10*(5), 449-452.

Sivakova, D., Sieglova, Z., Lubyova, B., & Novakova, J. (1994). A genetic profile of Romany (Gypsy) subethnic group from a single region in Slovakia. *Gene Geogr, 8*(2), 109-116.

Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics, 139*(1), 457-462.

Spurdle, A. B., Hammer, M. F., & Jenkins, T. (1994). The Y Alu polymorphism in southern African populations and its relationship to other Y-specific polymorphisms. *Am J Hum Genet, 54*(2), 319-330.

Stamatoyannopoulos, G. (1972). The molecular basis of hemoglobin disease. *Annu Rev Genet, 6*, 47-70.

Stephens, J. C., Reich, D. E., Goldstein, D. B., Shin, H. D., Smith, M. W., Carrington, M., Winkler, C., Huttley, G. A., Allikmets, R., Schriml, L., Gerrard, B., Malasky, M., Ramos, M. D., Morlot, S., Tzetis, M., Oddoux, C., di Giovine, F. S., Nasioulas, G., Chandler, D., Aseev, M., Hanson, M., Kalaydjieva, L., Glavac, D., Gasparini, P., Kanavakis, E., Claustres, M., Kambouris, M., Ostrer, H., Duff, G., Baranov, V., Sibul, H., Metspalu, A., Goldman, D., Martin, N., Duffy, D., Schmidtke, J., Estivil, X., O'Brien, S. J., & Dean, M. (1998). Dating the origin of the CCR5-Delta32 AIDS-resistance allele by the coalescence of haplotypes. *Am J Hum Genet, 62*(6), 1507-1515.

Stoneking, M. (2000). Hypervariable sites in the mtDNA control region are mutational hotspots. *Am J Hum Genet, 67*(4), 1029-1032.

Stoneking, M. (2001). Single nucleotide polymorphisms. From the evolutionary past. *Nature, 409*(6822), 821-822.

Su, B., Xiao, J., Underhill, P., Deka, R., Zhang, W., Akey, J., Huang, W., Shen, D., Lu, D., Luo, J., Chu, J., Tan, J., Shen, P., Davis, R., Cavalli-Sforza, L., Chakraborty, R., Xiong, M., Du, R., Oefner, P., Chen, Z., & Jin, L. (1999). Y-Chromosome evidence for a northward migration of modern humans into Eastern Asia during the last Ice Age. *Am J Hum Genet, 65*(6), 1718-1724.

Super, M., Schwarz, M. J., & Malone, G. (1992). Screening for cystic fibrosis carriers. *Lancet, 340*(8817), 490-491.

Suter, U., & Snipes, G. J. (1995). Biology and genetics of hereditary motor and sensory neuropathies. *Annu Rev Neurosci, 18*, 45-75.

Sykes, B. C., Corte-Real, H., & Richards, M. B. (1998). Palaeolithic and neolithic contributions to the European mitochondrial gene pool. In A.J. Boyce & C.G.N. Mascie-Yalor (Eds.), *Molecular Biology and Human Diversity* (pp. 130-140) Cambridge: Cambridge University Press.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics, 105*(2), 437-460.

Tauszik, T., Forrai, G., & Hollan, S. R. (1987). Taste sensitivity to phenylthiocarbamide in a sample of Hungarian Gipsies. *Gene Geogr, 1*(2), 103-107.

Tauszik, T., Friss, A., Gyodi, E., Santora, Z., Takacs, S., Kotvasz, A., Toth, A. M., Horvath, M., Tarjan, L., Petranyi, G., Hollan, S. R., & Simonovitis, I. (1985). Genetic polymorphism of the Gypsy population in Hungary as based on studies of red blood cell antigens. *Haematologia, 18*(3), 205-217.

Tcherenkov, L., & Laederich, S. (unpublished manuscript). The Nordic Metadialect.

ten Kate, L. P. (1998). Editorial. *Community Genetics, 1*, 1-2.

Terwilliger, J., & Ott, J. (1994). *Handbook of Human Genetic Linkage*. Baltimore: The Johns Hopkins University Press.

Terwilliger, J. D., Zollner, S., Laan, M., & Paabo, S. (1998). Mapping genes through the use of linkage disequilibrium generated by genetic drift: 'drift mapping' in small populations with no demographic expansion. *Hum Hered, 48*(3), 138-154.

Thibodeu, S. N., Bren, G., & Schaid, D. (1993) Microsatellite instability in cancer of the proximal colon. *Science*, 260 (5109), 751.

Thomas, P. K. (2000). Autosomal recessive hereditary motor and sensory neuropathy. *Curr Opin Neurol, 13*(5), 565-568.

Thomas, M. G., Parfitt, T., Weiss, D. A., Skorecki, K., Wilson, J. F., le Roux, M., Bradman, N., & Goldstein, D. B. (2000). Y chromosomes traveling south: the Cohen modal haplotype and the origins of the Lemba--the "Black Jews of Southern Africa". *Am J Hum Genet, 66*(2), 674-686.

Thomas, M. G., Skorecki, K., Ben-Ami, H., Parfitt, T., Bradman, N., & Goldstein, D. B. (1998). Origins of Old Testament priests. *Nature, 394*(6689), 138-140.

Timmerman, V., Nelis, E., de Jonge, P., Martin, J.-J., & van Broeckhoven, C. (1998). Hereditary Neuropathies. In A. E. H. Emery (Ed.), *Neuromuscular Disorders; clinical and molecular genetics*. Chichester: John Wiley & Sons. Ltd.

Todorova, A., Ashikov, A., Beltcheva, O., Tournev, I., & Kremensky, I. (1999). C283Y mutation and other C-terminal nucleotide changes in the gamma-sarcoglycan gene in the Bulgarian Gypsy population. *Hum Mutat, 14*(1), 40-44.

Torroni, A., Richards, M., Macaulay, V., Forster, P., Villems, R., Norby, S., Savontaus, M. L., Huoponen, K., Scozzari, R., & Bandelt, H. J. (2000). mtDNA haplogroups and frequency patterns in Europe. *Am J Hum Genet, 66*(3), 1173-1177.

Tournev, I., Aneva, L., Kamenov, O., Ishpekova, B., Katzarova, V., Guerguelcheva, V., Angelicheva, D., & Kalaydjieva, L. (1998) Gamma-sarcoglycan deficiency in Bulgarian Gypsies. *Muscle Nerve*, Supplement 7, 136.

Tully, L. A., Parsons, T. J., Steighner, R. J., Holland, M. M., Marino, M. A., & Prenger, V. L. (2000). A sensitive denaturing gradient-Gel electrophoresis assay reveals a high frequency of heteroplasmy in hypervariable region 1 of the human mtDNA control region. *Am J Hum Genet, 67*(2), 432-443.

Underhill, P. A., Jin, L., Lin, A. A., Mehdi, S. Q., Jenkins, T., Vollrath, D., Davis, R. W., Cavalli-Sforza, L. L., & Oefner, P. J. (1997). Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. *Genome Res, 7*(10), 996-1005.

Underhill, P. A., Jin, L., Zemans, R., Oefner, P. J., & Cavalli-Sforza, L. L. (1996). A pre-Columbian Y chromosome-specific transition and its implications for human evolutionary history. *Proc Natl Acad Sci USA, 93*(1), 196-200.

Underhill, P. A., Passarino, G., Lin, A. A., Marzuki, S., Oefner, P. J., Cavalli-Sforza, L. L., & Chambers, G. K. (2001). Maori origins, Y-chromosome haplotypes and implications for human history in the Pacific. *Hum Mutat, 17*(4), 271-280.

Underhill, P. A., Shen, P., Lin, A. A., Jin, L., Passarino, G., Yang, W. H., Kauffman, E., Bonné-Tamir, B., Bertranpetit, J., Francalacci, P., Ibrahim, M., Jenkins, T., Kidd, J. R., Mehdi, S. Q., Seielstad, M. T., Wells, R. S., Piazza, A., Davis, R. W., Feldman, M. W., Cavalli-Sforza, L. L., & Oefner, P. J. (2000). Y chromosome sequence variation and the history of human populations. *Nat Genet, 26*(3), 358-361.

Vainzof, M., Passos-Bueno, M. R., Canovas, M., Moreira, E. S., Pavanello, R. C., Marie, S. K., Anderson, L. V., Bonnemann, C. G., McNally, E. M., Nigro, V., Kunkel, L. M., & Zatz, M. (1996). The sarcoglycan complex in the six autosomal recessive limb-girdle muscular dystrophies. *Hum Mol Genet, 5*(12), 1963-1969.

van Belzen, N., Dinjens, W. N., Diesveld, M. P., Groen, N. A., van der Made, A. C., Nozawa, Y., Vlietstra, R., Trapman, J., & Bosman, F. T. (1997). A novel gene which is up-regulated during colon epithelial cell differentiation and down-regulated in colorectal neoplasms. *Lab Invest, 77*(1), 85-92.

van Holst Pellekaan, S., Frommer, M., Sved, J., & Boettcher, B. (1998). Mitochondrial control-region sequence variation in aboriginal Australians. *Am J Hum Genet, 62*(2), 435-449.

van Loghem, E., Tauszik, T., Hollan, S., & Nijenhuis, L. E. (1985). Immunoglobulin allotypes in Hungarian Gypsies. Relationship to populations from India. *J Immunogenet, 12*(3), 131-137.

van Ommen, G. J., Bakker, E., & den Dunnen, J. T. (1999). The human genome project and the future of diagnostics, treatment, and prevention. *Lancet, 354 Suppl 1*, SI5-10.

Varilo, T., Nikali, K., Suomalainen, A., Lonnqvist, T., & Peltonen, L. (1996a). Tracing an ancestral mutation: genealogical and haplotype analysis of the infantile onset spinocerebellar ataxia locus. *Genome Res, 6*(9), 870-875

Varilo, T., Savukoski, M., Norio, R., Santavuori, P., Peltonen, L., & Jarvela, I. (1996b). The age of human mutation: genealogical and linkage disequilibrium analysis of the CLN5 mutation in the Finnish population. *Am J Hum Genet, 58*(3), 506-512.

Veldhuisen, B., Saris, J. J., de Haij, S., Hayashi, T., Reynolds, D. M., Mochizuki, T., Elles, R., Fossdal, R., Bogdanova, N., van Dijk, M. A., Coto, E., Ravine, D., Norby, S., Verellen-Dumoulin, C., Breuning, M. H., Somlo, S., & Peters, D. J. (1997). A spectrum of mutations in the second gene for autosomal dominant polycystic kidney disease (PKD2). *Am J Hum Genet, 61*(3), 547-555.

Verzar, F., & Weszeczky, O. (1921). Rassenbiologische Untersuchungen mittels Isohamagglutininen. *Biochemische Zeitschrift, 126*, 33-39.

Visapaa, I., Fellman, V., Varilo, T., Palotie, A., Raivio, K. O., & Peltonen, L. (1998). Assignment of the locus for a new lethal neonatal metabolic syndrome to 2q33-37. *Am J Hum Genet, 63*(5), 1396-1403.

Voet, D., & Voet, J. G. (1995). *Biochemistry* (Second ed.). New York: John Wiley & Sons, Inc.

Wald, N. J. (1991). Couple screening for cystic fibrosis. *Lancet, 338*(8778), 1318-1319.

Wald, N. J., George, L. M., Wald, N. M., & Mackenzie, I. (1993). Couple screening for cystic fibrosis. *Lancet, 342*(8882), 1307-1308.

Ward, R. H. (1998). Linguistic divergence and genetic evolution: a molecular perspective from the New World. In A.J. Boyce & C.G.N. Mascie-Yalor (Eds.), *Molecular Biology and Human Diversity* (pp. 205-224) Cambridge: Cambridge University Press.

Watson, E. K., Mayall, E. S., Lamb, J., Chapple, J., & Williamson, R. (1992). Psychological and social consequences of community carrier screening programme for cystic fibrosis. *Lancet, 340*(8813), 217-220.

Weber, J. L., & Wong, C. (1993). Mutation of human short tandem repeats. *Hum Mol Genet, 2*(8), 1123-1128.

Weir, B. S. (1990). *Genetic Data Analysis: methods for discrete population genetic data*. Sunderland, Massachusetts: Sinauer Associates, Inc. Publishers.

Weir, B. S. (1996). *Genetic Data Analysis II: methods for discrete population genetic data*. Sunderland, Massachusetts: Sinauer Associates, Inc.

Weiss, K. M. (1993). *Genetic Variation and Human Disease:principles and evolutionary approaches*. Cambridge: Cambridge University Press.

Weiss, K. M. (1998). Coming to terms with human variation, *Annual Review of Anthropology* (Vol. 27, pp. 273-300).

Wertz, D. C., & Fletcher, J. C. (1988). Attitudes of genetic counselors: a multinational survey. *Am J Hum Genet, 42*(4), 592-600.

Wexler, P. (1997). *Could there be a Rotwelsch origin for the Romani lexicon?* Paper presented at the Third International Conference on Romani Linguistics, Prague.

White, P. S., Tatum, O. L., Deaven, L. L., & Longmire, J. L. (1999). New, male-specific microsatellite markers from the human Y chromosome. *Genomics, 57*(3), 433-437.

Wilfond, B. S., & Fost, N. (1990). The cystic fibrosis gene: medical and social implications for heterozygote detection. *J Amer Med Ass, 263*(20), 2777-2783.

Wilson, J. F., & Goldstein, D. B. (2000). Consistent long-range linkage disequilibrium generated by admixture in a Bantu-Semitic hybrid population. *Am J Hum Genet, 67*(4), 926-935.

Wolff, G., & Jung, C. (1995). Nondirectiveness and genetic counseling. *J Genet Couns, 4*, 3-25.

Wright, S. (1965). The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution, 19*, 395-420.

Xu, X., Peng, M., & Fang, Z. (2000). The direction of microsatellite mutations is dependent upon allele length. *Nat Genet, 24*(4), 396-399.

Yang, X., & Griffiths, A. J. (1993). Male transmission of linear plasmids and mitochondrial DNA in the fungus *Neurospora*. *Genetics, 134*(4), 1055-1062.

Zeesman, S., Clow, C. L., Cartier, L., & Scriver, C. R. (1984). A private view of heterozygosity: eight-year follow-up study on carriers of the Tay-Sachs gene detected by high school screening in Montreal. *Am J Med Genet, 18*(4), 769-778

Zerjal, T., Dashnyam, B., Pandya, A., Kayser, M., Roewer, 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., & Tyler-Smith, C. (1997). Genetic relationships of Asians and Northern Europeans, revealed by Y- chromosomal DNA analysis. *Am J Hum Genet, 60*(5), 1174-1183.

Zhou, D., Salnikow, K., & Costa, M. (1998). Cap43, a novel gene specifically induced by Ni2+ compounds. *Cancer Res, 58*(10), 2182-2189.

Zlotogora, J. (1994). High frequencies of human genetic diseases: founder effect with genetic drift or selection? *Am J Med Genet, 49*(1), 10-13.

Zlotogora, J. (1998). Selection for carriers of recessive diseases: a common phenomenon? *Am J Med Genet, 80*(3), 266-268.

Zlotogora, J., Zeigler, M., & Bach, G. (1988). Selection in favor of lysosomal storage disorders? *Am J Hum Genet, 42*(2), 271-273.

Zouros, E., Freeman, K. R., Ball, A. O., & Pogson, G. H. (1992). Direct evidence for extensive paternal mitochondrial DNA inheritance in the marine mussel *Mytilus*. *Nature, 359*(6394), 412-414.