PHYLOGENETIC GENOMICS OF THE FOUNDING POPULATION OF NEWFOUNDLAND: INSIGHTS FROM COMPLETE mtDNA SEQUENCES

ANGELA MICHELLE POPE

Your file Votre référence
ISBN: 978-0-494-57485-0
Our file Notre référence
ISBN: 978-0-494-57485-0

Phylogenetic genomics of the founding population of Newfoundland: insights from complete mtDNA sequences

by

Angela Michelle Pope

A dissertation submitted in partial fulfillment of the requirements for the degree of Masters of Science (Medicine – Human Genetics)

Department of Medicine
Memorial University of Newfoundland
St. John's, Newfoundland and Labrador

July 2007

Abstract

The importance of the island of Newfoundland as a place to study genetics stems from the history of its early settlement. Beginning in the early part of the 17th century, the island was settled by fewer than 25 000 founders. The genetic structure of the modern-day descendants of these settlers is expected to be influenced both by the initial founder effects and subsequent inbreeding and genetic drift. Therefore, it is of interest to explore the population genetic structure of Newfoundlanders for evidence of these phenomena.

Complete (16 570 bp) mitochondrial DNA sequences were obtained for 27 individual Newfoundlanders by conventional dideoxy sequencing as well as a novel microarray technology (GeneChip: Affymetrix). A total of 220 SNPs were found; every individual had a unique sequence. Two individuals were sequenced by both methods. In both cases, microarray sequencing was shown to be highly efficient and accurate: 99.99% and 99.97% of bases were called by the algorithm, and 100.0% of SNPs were detected, with no false positives.

In combination with published data from 42 other European, Eurasian, and First Nations individuals, phylogenetic analysis showed that 25 Newfoundland individuals could be associated with one of five major haplogroups (H, J, K, T, and U) previously identified in Europeans from mtDNA Control Region (CR) profiles. Phylogenetic analysis of complete genomes more clearly defines the relationships among these haplogroups than do CR profiles alone, and shows that haplogroups U (with respect to K) and H (with respect to subtypes H3 and H16) are not monophyletic. One individual was assignable to haplogroup A, which has not previously been seen in Europeans but

which is common in northeastern First Nations peoples. Another individual had a signature most similar to haplogroup I, a rare Scandinavian type: the phylogenetic relationships of this lineage to H, (U+K), and (J+T) are unresolved.

In contrast with expectations from the genetics of small populations and the historical settlement patterns, "genome-type" diversity in the Newfoundland samples is high and closely parallels patterns in western Europeans, with no loss of haplotypes or significant shifts in their relative frequencies.

Acknowledgements

First and foremost, I would like to thank Dr. Steven M. Carr without whom this project would not have been possible. His supervision, countless suggestions, and help throughout this project have been beyond measure. Secondly, I would like to thank Dr. Ban Younghusband for his financial support and supervision throughout this experiment. I would also like to acknowledge Dr. H. Dawn Marshall and Dr. Terry Young for all of their support as well as their helpful comments on the manuscript. Many thanks must also go to my fellow lab mates: Mark Coulson, Ana Duggan, Anne-Marie Gale, Kim Johnstone, Sarah Flynn, and Corrine Wilkerson – the help and support they provided was indispensable. I would also like to thank Siobhan Coady and Lynette Peddle of Newfound Genomics Inc. for the use of the Affymetrix unit. Finally, I would like to thank my partner Curtis – his support and encouragement was a major factor in the culmination of this thesis and his computer prowess never ceases to amaze me!

Table of Contents

List of Tables

List of Figures

List of Abbreviations

12S	= 12S rDNA
16S	= 16S rDNA
A	= adenine
ATP6	= ATPase subunit 6
ATP8	= ATPase subunit 8
bp	= base pairs
C	= cytosine
COXI	= cytochrome oxidase I
COXII	= cytochrome oxidase II
COXIII	= cytochrome oxidase III
CR	= control region
Cyt*b*	= cytochrome *b*
DNA	= deoxyribonucleic acid
dNTP	= deoxyribonucleotide triphosphate
dS/N	= differential signal-to-noise ratio
EDTA	= ethylenediamine tetraacetic acid
G	= guanine
mtDNA	= mitochondrial DNA
NADH	= nicotinamide adenine dinucleotide
ND1	= NADH dehydrogenase subunit 1

ND2 = NADH dehydrogenase subunit 2

ND3 = NADH dehydrogenase subunit 3

ND4 = NADH dehydrogenase subunit 4

ND4L = NADH dehydrogenase subunit 4L

ND5 = NADH dehydrogenase subunit 5

ND6 = NADH dehydrogenase subunit 6

nt = nucleotide position

PCR = polymerase chain reaction

PAUP* = Phylogentic analysis using parsimony and other methods

RNA = ribonucleic acid

SNPs = single nucleotide polymorphisms

T = thymine

TBE = Tris-borate-EDTA

tRNA = transfer RNA

List of Appendices

1.0 Introduction

1.1 History of the Settlement of Newfoundland

The importance of Newfoundland as a place to study genetics stems from the history of its early settlement. The island of Newfoundland is thought to have been founded by less than 25 000 settlers beginning in the early part of the 17th century (Prowse, 1972). Historical records indicate that these settlers were of three main ethnic groups: English, Irish, and French. The vast majority of current Newfoundlanders are the descendants of immigrants from the British Isles. They came from highly localized regions of southwest England, including Devon, Dorset and Cornwall (Figure 1), and the southeast of Ireland (Prowse, 1972; Mannion, 1977; Rowe, 1980). Descendants of French settlers are less numerous and many are believed to have subsequently relocated to St. Pierre et Miquelon. Much less is known about the history of the French settlers.

The first settlements of Newfoundland by the English were seasonal. English migrants would venture to Newfoundland during the summer to take advantage of the fishery and return home in the winter (Prowse, 1972; Mannion, 1977; Rowe, 1980). It was not until

the early 17th century that any permanent settlement was attempted. One of the first successful settlements in Newfoundland was in Conception Bay; several families were resolute in their desire to stay, instead of returning to England (Mannion, 1977). Between 1675 and 1677, the English inhabited more than 30 sites, ranging from Trepassey to Salvage on the eastern shore. In consequence, the area from Cape Race to Cape Bonavista is known as the English shore (Figure 2) (Rowe, 1980; Prowse, 1972).

The origin of Irish settlement in Newfoundland is controversial. However, there is clear evidence that there were Irish settlers in St. John's prior to 1675 (Rowe, 1980). Early potato famines induced thousands of Irish people to immigrate to Newfoundland. By 1753, all major communities on the Avalon Peninsula had Irish majorities; they constituted nearly half the total population (Rowe, 1980; Parfrey *et al.*, 2002). The Irish tended to segregate themselves geographically during the settlement of Newfoundland. Almost all the southern shore, Trepassey, St. Mary's, Placentia and the Placentia Bays had settlements that comprised entirely Irish Roman Catholics (Figure 2) (Prowse, 1972; Rowe, 1980). Protestant English and Roman Catholic Irish both inhabited the larger towns of St. John's and Harbour Grace. Although they lived in close proximity to one another, there was

limited intermarriage due to the religious barrier. When the Irish settled away from the Avalon Peninsula in areas such as Bonavista, Notre Dame Bay and White Bay, they tended to keep to themselves and to settle harbours that were not occupied by the Protestant English (Rowe, 1980). This religious barrier maintained such segregation until recent times.

In the early 16th century, the fishery on the south, west, and northeast coasts was a French monopoly (Prowse, 1972; Rowe, 1980). The area north of Cape Bonavista and down the west coast of the island was known as the French Shore (Figure 2), and constituted nearly half of the island's circumference (Rowe, 1980). The majority of the French were seasonal residents. In their absence, the English and Irish encroached on their territory. The French lost their claim to the island during Queen Anne's war in 1767; today they inhabit the two islands of St. Pierre et Miquelon, just off the southern coast of Newfoundland.

In addition to religion, the English, Irish and French settlers are believed to have had limited intermarriage due to language, socioeconomic, and geographic barriers. It is thought that the progeny of the settlers remain near the original settlements, such that current regional populations are thought to be relatively genetically

homogeneous (Young *et al.*, 1999; Parfrey *et al.*, 2002). As an example, a study of three outports found that only 1-8% of the population were immigrants to the area, and 60% of births were to parents originating from the same small community (Bear *et al.*, 1988).

1.2 The Value of Founder Populations for Studying Disease

Subsequent to its founding, the population of Newfoundland grew rapidly, due to large family size. These families were typically isolated from people in other communities, thus consanguineous marriages were not uncommon. This further contributed to the potential for genetic homogeneity and population isolation. Present-day genetic homogeneity can largely be attributed to three main factors: founder effect, genetic drift, and inbreeding. "Founder effect" refers to the situation whereby a population is started or "founded" by a small subset of individuals from a larger population (in the case of the Newfoundland population, this occurred when a small number of individuals from specific areas of Europe settled in the province). Therefore, there would have been limited initial genetic variation and non-random sampling in the new population, and this may have

introduced rare disease alleles at a higher frequency than observed in the larger European population. In turn, this may have provided an opportunity for rare diseases to become prevalent (Woods *et al.*, 1999). The second factor affecting genetic homogeneity is genetic drift. In populations that maintain small sizes over long periods, allele frequencies fluctuate more strongly (they "drift") which tends to reduce genetic variation, and may increase the frequency of rare alleles by chance. The final factor influencing genetic homogeneity is inbreeding. Inbreeding is defined as the mating of individuals that share at least one common ancestor (Griffiths *et al.*, 2000; Hartl & Jones, 2005). Inbreeding increases the probability that disease alleles that are identical by descent will come together in homozygous combinations, so as to increase the incidence of genetic disease to a higher than expected frequency. For example, in offspring of first-cousin matings an additional sixteenth of the variation in DNA is homozygous when compared to offspring of outbred marriages (Sheffield *et al.*, 1998). Consequently, many genetic diseases show high prevalence in the province of Newfoundland, including Bardet-Beidel (Woods *et al.*, 1999), non-polyposis colorectal cancer (Woods *et al.*, 2005), and late infantile neuronal ceroid lipofuscinosis (Andermann *et al.*, 1988).

Genetically isolated populations, such as that of Newfoundland, are useful for studying disease genes because inbreeding and founder effects reduce the genetic complexity of the disorder, so that it can be more readily investigated (Sheffield *et al.*, 1998). Linkage disequilibrium (LD), the non-random association between individual marker alleles and disease alleles, is used to identify marker alleles that are identical by descent (Sheffield *et al.*, 1998, Service *et al.*, 1999). Inbreeding and founder effects can increase LD relative to the source population. Founder populations, therefore, may potentially show a higher incidence of single nucleotide polymorphisms (SNPs) associated with genetic conditions, making disease SNPs more easily identifiable. LD can be most effectively utilized for disease mapping in genetically isolated populations, especially those in which consanguineous unions are common (Sheffield *et al.*, 1998). Genetic isolation and inbreeding may serve to reduce the complexity of non-allelic heterogeneity, which can complicate the genetic mapping of some disorders. Furthermore, the average size of a nuclear family is on average larger in Newfoundland (Prowse, 1972; Rowe, 1980), which increases the possibility that multiple affected individuals will be found in a single family and, in turn, facilitates linkage mapping (Sheffield *et al.*, 1998).

1.3 Mitochondrial DNA and Phylogenetic Reconstruction

Prior to using a population to investigate the genes associated with simple or complex genetic condition, an independent determination of the genetic structure of the population must first be performed. This is to ensure that the observed LD is due to close proximity of genes and not another population demographic factor. Mitochondrial DNA (mtDNA) is an extremely useful tool for understanding evolution due to characteristics such as a high copy number, apparent lack of recombination, high substitution rate, and freedom from the effects of positive selection (Ingman *et al.*, 2000; Barbujani & Bertorelle, 2001). Mitochondrial DNA is inherited maternally in vertebrates; thus analysis of mtDNA sequences reflects the history of females. Due to the stochastically constant and "clock like" mutation rate combined with the absence of recombination, SNPs accumulate sequentially over time (Maca-Meyer *et al.*, 2001). Thus, the degree of relatedness between two individuals can be determined through the analysis of shared SNPs. The lack of recombination allows for the construction of a highly accurate phylogenetic tree, because

ancestral DNA sequences are not altered (Kaessmann & Pääbo, 2002). Mitochondrial DNA exhibits rapid evolution, which makes it suitable for comparison of closely related groups such as human populations. Finally, mtDNA exhibits a higher observed substitution rate, because selection pressure, which eliminated mutations in nuclear genes, is reduced (Avise *et al.*, 1994; Russell, 1998; Richards *et al.*, 2000). Another reason for the higher mutation rate is the fact that mtDNA has a less efficient repair system that allows for an increased number of SNPs to be available for selection.

Mitochondrial DNA has been extensively investigated over the past several decades because of its utility as a population and evolutionary genetic marker. Most of these studies have examined only one or a few mitochondrial genes (Allard *et al.*, 2006; Alfonso-Sanchez *et al.*, 2006; Lee *et al.*, 2002; Sigurgardottir *et al.*, 2000). In the present study, the entire mitochondrial genome was examined. Analysis of the whole mitochondrial genome facilitates interpretations based on highly-resolved intraspecific gene trees or pedigrees among individuals. This has been demonstrated in a study of the global diversity in the human population (Ingman *et al.*, 2000). A whole mitochondrial genome study also provides the opportunity to compare rates and patterns of mtDNA evolution within a population of humans.

1.4 Sequencing Protocols

Analysis of human mtDNA sequence variation typically involves identification of restriction fragment length polymorphisms (RFLPs) (Cann and Wilson, 1983; Wallace, 1994) or direct DNA sequencing (Ingman *et al.*, 2000). Most recently, oligonucleotide microarray methods have become available (Chee *et al.*, 1996; Maitra *et al.*, 2004; Carr *et al.*, 2007). As this study involved both dideoxy and microarray sequencing, a review of these techniques is merited.

1.4.1 Dideoxy Sequencing

The dideoxy method was first introduced in 1977 by Sanger and colleagues (Sanger *et al.*, 1977). Starting with a DNA template to be sequenced, the idea was to generate a population of oligonucleotides that correspond in length to each of the nucleotides in the template sequence. The method involves a controlled synthesis of DNA from the starting template using a radiolabeled primer, a polymerase, and a supply of each of the four dNTPs. The generated fragments were

designed to correspond to each nucleotide position in the template by the incorporation of dideoxynucleotides, which terminate DNA synthesis at the position of incorporation because the 3′ hydroxyl group is modified to a non-extensible 3′-OH. A population of such fragments needs to be made for each of the four nucleotides, with the four dideoxynucleotides. In traditional manual sequencing, the population of fragments made with each deoxynucleotide was electrophoresed in four adjacent lanes on a polyacrylamide gel capable of resolving fragments that differ by only one base in length. The result was a stepladder or ladder gel from which the sequence can be read base by base from top to bottom on an autoradiograph.

Automated DNA sequencing relies on the use of fluorescent dyes instead of radiolabeling. Automation involves the excitation of dye molecules by a laser beam, the amplification and detection of fluorescence by a photomultiplier tube or a CCD camera, and computer software to identify each ddNTP-terminated fragment on the basis of fluorescence emission wavelength as it passes the detector and converts it into a sequence (Tamarin, 2002). There are four dyes, one for each nucleotide, each fluorescing at a different wavelength. Dyes can be attached to the primer or the terminator, but dye-terminator labeling has been shown to be more practicable. All four nucleotide

fragment populations are electrophoresed in one lane. Since the fluorescent signal can be continuously detected and processed, gels can be run more efficiently and more data collected per run than for manual gels.

1.4.2 Mitochondrial GeneChip Microarray

The routine sequencing of the complete mitochondrial DNA is labour intensive and error prone. Microarrays are inherently parallel devices that offer both a high-throughput method as well as a minimal input of effort (Hacia, 1999). Chee *et al.* (1996) developed the first mitochondrial-sequencing microarray. This chip comprised tiled oligonucleotide sequencing probes synthesized by standard photolithography and solid phase DNA synthesis. This microarray platform had several limitations. The protocol required generating RNA by *in vitro* transcription of genomic DNA for chip hybridization, only a single strand of the target mitochondrial sequence was tiled onto the chip, and robust genotype assignment software was absent (Maitra *et al.*, 2004).

Maitra and colleagues developed the GeneChip CustomSeq Resequencing microarry as an array-based sequencing platform for

rapid and high-throughput analysis of mtDNA coding region mutations. The MitoChip can sequence 29 366 bp of double-stranded DNA, which includes 980 bp of plasmid DNA sequence as a control for chip hybridization, in a single assay. Both the forward and reverse strands of the entire mitochondrial coding region (15 452 bp) are tiled once onto the array. Both strands of an additional 12 935 bp of the coding region excluding the 12S and 16S RNA sequences are tiled in duplicate (Maitra *et al.*, 2004). The mitochondrial control region was not included on the chip for two reasons. The first is that the control region is particularly GC rich, which often leads to suboptimal hybridization. The second reason is that the most common mutation observed in the control region are indels of a poly C tract known as D310. This type of mutation is poorly detected by current microarray hybridization technology (Maitra *et al.*, 2004). Affymetrix fabricates the MitoChip by photolithography and solid phase DNA synthesis. Each chip contains approximately 300 000 "features", where each feature consists of 10^6 copies of a 25 bp oligonucleotide probe (Maitra *et al.*, 2004). For each 25 bp probe, three variant probes are included that vary the central or 13^{th} base, one for each of the three alternative nucleotides. Mismatch at this position most strongly influences binding, so that a genomic DNA fragment will bind preferentially to

only one of the four oligonucleotide probes at any tiled position. In the process of scanning the MitoChip, the scanner measures the fluorescence intensity for each feature and determines which base at that position has the highest relative intensity (Maitra *et al.*, 2004). The raw data are recorded at each site for both the forward and reverse strands and are presented in a table format. The revised Cambridge reference sequence (Andrews *et al.*, 1999) was used as the reference sequence.

1.5 Objectives of this study

The purpose of this study is to measure the extent of genetic differentiation and degree of relatedness within and among haplogroups present in the Newfoundland population, based on whole mitochondrial DNA sequences. The extent to which the structure of the population is reflected genetically has several important implications for understanding disease in this province. As previously discussed, homogeneous population isolates may influence the local appearance of recessive diseases and other inherited diseases, as shown from the studies on Bardet-Biedl Syndrome (Woods *et al.*, 1999; Young *et al.*,

1999; Moore *et al.*, 2005). They may be valuable in the study of complex diseases, such as diabetes, that have been shown to be multifactorial – that is, involving both several genes as well as environmental factors. A homogeneous population should have fewer of the alleles associated with the genetic disease under consideration, and the environmental conditions may be more uniform among the individuals investigated.

The study of genetic variation in relation to geographic location is known as the "phylogeographic approach" (Avise *et al.*, 1987; Richards *et al.*, 2000). This method has been shown to resolve maternal lineages using mtDNA in the investigation of the European population (Richards *et al.*, 1998; Ingman *et al.*, 2000; Richards *et al.*, 2000; Barbujani and Bertorelle, 2001; Silva *et al.*, 2002, Mishmar *et al.*, 2003). These studies along with others have shown that European mtDNAs fall into several distinct clades or haplogroups. Current literature supports the idea that there is one African "Mitochondrial Eve" from which all living humans are descendants. Subsequently, all non-African individuals have a common ancestor of ~ 30 000 yrs ago (Cann *et al.*, 1987). It has been suggested that there are "Seven Daughters of Eve" which encompass the seven haplogroups of Europeans: U, X, H, V, T, K, and J (Torroni *et al.*, 1996; Sykes, 2001).

It is uncertain whether haplogroups will survive inspection when sequencing the complete mitochondrial genome, as compared to many previous works that rely on sequence data from hypervariable sequences I and II alone. As the Newfoundland individuals investigated in this study were all expected to be of European descent, the analysis of their complete mitochondrial genomes will determine to what extent European genetic patterns have been preserved in Newfoundland, and what effects founder effect, drift, and inbreeding may have had.

Figure 1 Routes of English migration as Postulated by Mannion (1977; Figure 1-4).

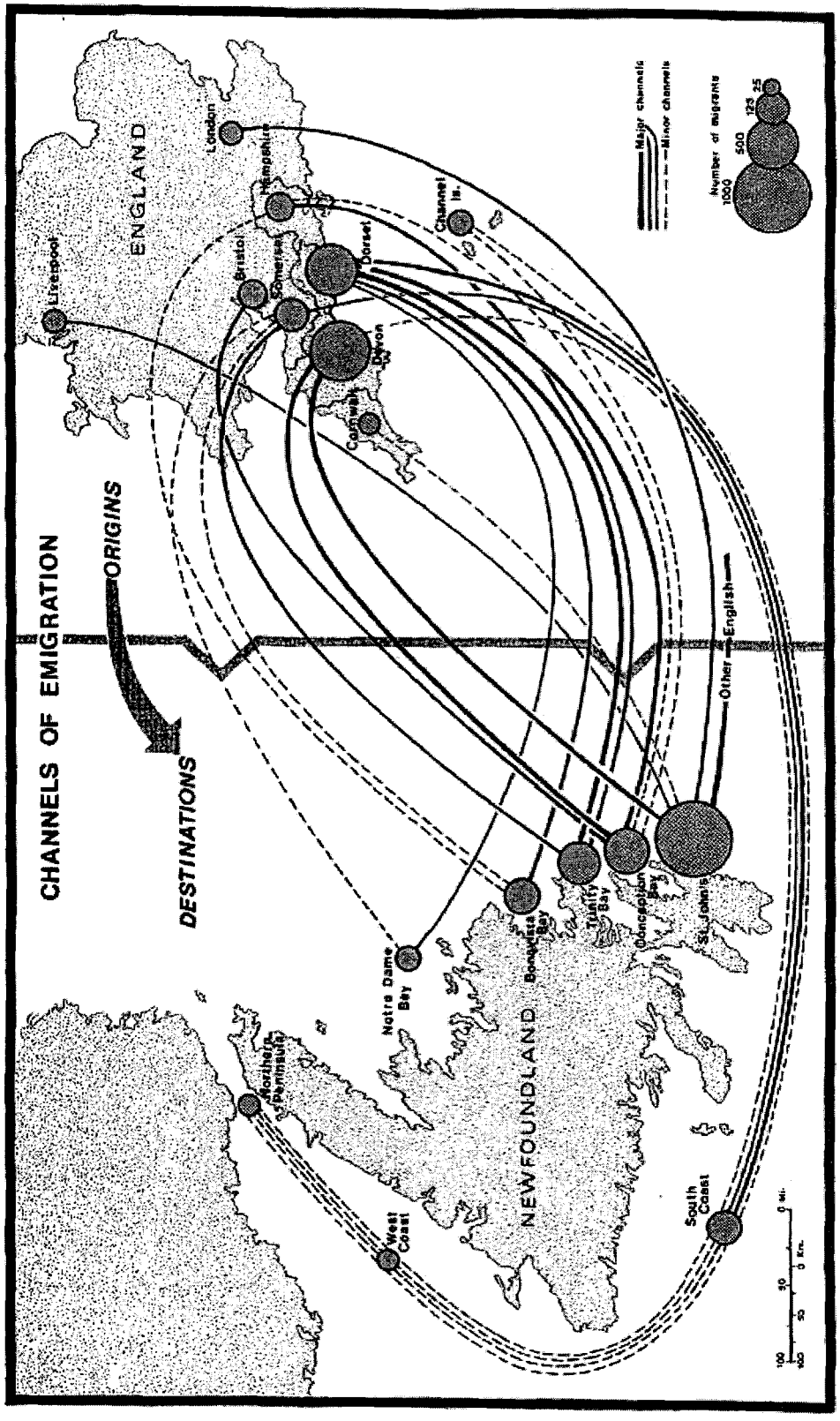
CHANNELS OF EMIGRATION
ORIGINS
DESTINATIONS
ENGLAND
NEWFOUNDLAND
Liverpool
London
Hampshire
Bristol
Somerset
Dorset
Channel Is.
Devon
Cornwall
Major channels
Minor channels
Number of migrants
1000
500
125
25
Other
English
Notre Dame Bay
Bonavista Bay
Trinity Bay
Conception Bay
St. John's
Northern Peninsula
West Coast
South Coast
0 mi.
0 Km.

Figure 1–4

Figure 2 Map of the island of Newfoundland with locations described in the text.


Strait of Belle Isle
St. Anthony
Island of Newfoundland
Northern Peninsula
White Bay
Gulf of St. Lawrence
Notre Dame Bay
Fogo Island
Bay of Islands
Deer Lake
Grand Falls-Windsor
Gander
Bonavista Bay
Port au port Peninsula
Corner Brook
Grand Lake
Gander Lake
Bonavista
Stephenville
Red Indian Lake
Trinity
St. George's Bay
Trinity Bay
Clarenville
Conception Bay
South Coast
Port aux Basques
Burgeo
ST. JOHN'S
Avalon Peninsula
Cabot Strait
Placentia
Grand Bank
Marystown
Burin Peninsula
Placentia Bay
St. Pierre et Miquelon (France)
Atlantic Ocean
St. Mary's Bay
Cape Race

2.0 Materials and Methods

2.1 Choice of Individuals and Collection of DNA Samples

Individuals were selected for this study based on a questionnaire regarding maternal ancestry from a database of individuals whose DNA had already been obtained and stored at Memorial University of Newfoundland (the portion of this questionnaire relevant to the mitochondrial sampling is provided in Appendix A). The Human Investigation Committee (H.I.C) approval had been obtained and DNA extraction were previously performed at the Faculty of Medicine at Memorial University of Newfoundland. Of the 29 DNA samples chosen, 28 samples (16 English, eight Irish, and four French) contained sufficient DNA for mitochondrial analysis.

Since the investigated Newfoundland individuals are of putative European descent, it is useful to compare these individuals with other individuals of known or putative European ancestry. This will help to determine if Europeans continue to be grouped together in their own clades as shown in previous investigations (Ingman *et al.*, 2000; Silva *et al.*, 2002; Mishmar *et al.*, 2003). Therefore, in addition to the 27 individuals sequenced in this study, sequences of 42 supplementary

individuals were included. There are seven European individuals from the Ingman *et al.* (2000) study (Individuals designated Italian, German, Georgian, English, Tatar, Dutch and French: GenBank accession numbers AF346988, AF346983, AF346982, AF346978, AF346974, AF346975, AF346981), the revised Cambridge reference sequence (Andrews *et al.*, 1999: GenBank accession number J01415.1), a Swedish sequence (Arnason *et al.*, 1996: GenBank accession number X93334), and three Newfoundlanders of unspecified ethnicity (NF1, NF2, and NF3: H.D. Marshall, personal communication, 2003). A total of 18 African, Asian, Native American and Indian sequences were included (Ingman *et al.*, 2000) which were the individuals designated Evenki, China1 & China2, Chukchi, Asian Indian, two South American Indians (Warao 1 & 2), Buriat, Khirgiz, Guarani, Japanese1 & Japanese2, Inuit, Uzbek, Piman (North American Indian), Samoan, Korean, and Saami: Genbank accession numbers AF346979, AF346971-2, AF346971, AF346966, AF347012-3, AF346970, AF346991, AF346984, AF346989-90, AF347010, AF347011, AF347001, AF347007, AF346993, and AF347006 respectively. An additional five individuals belonging to haplogroups known to be outside the European clades were included; these individuals were from Morocco, Jordan, Spain, and two different islands of the Spanish

Canary Islands – Tenerife and Hierro (Maca-Meyer *et al.*, 2001: Genbank accession numbers AF381986, AF381999, AF382003, AF382009, and AF382010 respectively). Finally, five sequences from Native American individuals were incorporated (Mishmar *et al.*, 2003: Genbank accession numbers AY195748, AY195749, AY195759, AY195786, and AY195787).

2.2 Protocol for Individuals Processed by Dideoxy Sequencing

2.2.1 Polymerase Chain Reaction (PCR) and DNA Sequencing

The entire mitochondrial genome (16 570 bp) of each of the 20 individuals processed via conventional methods was amplified by polymerase chain reaction (PCR) in twenty-four overlapping segments. The whole mitochondrial genome sequence of six of these individuals was previously obtained (Pope, 2003). The PCR primer sequences used, amplification product length and length of overlap for these amplifications are presented in Appendix B (Rieder *et al.*, 1998).

PCR amplification was performed with a cocktail of 0.25 µL Hotstar *Taq* DNA polymerase (5 U/µL; Qiagen, Inc.: Mississauga, ON),

2.5 µL 10x PCR buffer (Qiagen, Inc.: Piscataway, NJ), 0.5 µL 10 mM dNTPs (Amersham Pharmaceuticals), 2 µL of 10µM forward primer and 2 µL of 10 µM reverse primer (custom primers synthesized by Qiagen Inc.: Mississauga, ON), 19 µL distilled, deionized water (H_2O), and 1 µL of 25 ng/µL of DNA. Following an initial denaturation at 95°C for 15 minutes, the amplification profile consisted of the following steps repeated for 40 cycles: denaturation at 95°C for 30 seconds, anneal at 55°C for 30 seconds, and elongation at 72°C for one minute. A final elongation step following the completion of the 40 cycles was accomplished by maintaining the mixture at 72°C for 10 minutes. PCR reactions were performed in a GeneAmp PCR 9600 thermal cycler (Perkin Elmer).

Confirmation of successful amplification was achieved by the electrophoresis of 2.5 µL of the PCR product through a 2% Low Electroendosmosis agarose gel (SeaKem) in Tris-Borate-EDTA (TBE) buffer (Sigma-Aldrich) containing 1 ng/mL ethidium bromide (Sigma-Aldrich). DNA was visualized on an ultraviolet light transilluminator at 312 nm (Spectroline, model TC-312A). Comparison to a known molecular weight marker, *ΦX174* digested with *Hae*III (Amersham, Inc.), determined if the correct PCR product had been amplified. Amplification products were then purified with the Qiagen QIAquick

PCR purification kit (Qiagen, Inc.) to ensure the removal of non-specific amplification products and primer concatenation products.

Sequencing reactions were performed using Big Dye Terminator version 3.0 chemistry (Applied Biosystems) with both the forward and reverse primers. Primers used for sequencing reactions were the same as those used for PCR and are listed in Appendix B. Sequencing reactions were carried out by first evaporating 5 µL of the purified DNA sample to dryness in a SpeedVac Concentrator SVC100H vacuum centrifuge in order to increase DNA template concentration. Next, a cocktail of 2.2 µL of deionized nanopure water, 2 µL of Big Dye Terminator (v. 3.0; Applied Biosystems Inc.) and 1.6 µL of 2 µM primer (either forward or reverse) were added to each sample. Sequencing reactions were carried out in a Perkin-Elmer Cetus TC-1/PE 480 thermal cycler as follows: an initial denaturation at 96°C for 2 minutes, followed by 35 cycles consisting of 96°C for 30 seconds, annealing of primers at 50°C for 15 seconds, and extension of products at 60°C for 4 minutes.

Excess reactants were removed from the sequencing reactions by isopropanol precipitation. This purification step was carried out by the addition of 40 µL of 75% isopropanol followed by incubation at room temperature for 30 minutes. This was followed with a 20 minute

centrifugation at 15°C at a speed of 13 000 rpm (17 949 rcf) in a refrigerated Eppendorf 5804R centrifuge. The supernatant was removed via aspiration, taking care not to disturb the pellet. The pellet was then washed with 250 µL of 75% isopropanol and the sample allowed to incubate for 10 minutes at room temperature preceding a 10 minute centrifugation. The supernatant was again removed via aspiration, and the DNA pellet was vacuum-centrifuged for 8 minutes or allowed to air dry for approximately one hour to ensure complete removal of the isopropanol. DNA pellets were resuspended in 5 µL 5:1 formamide/25 mM EDTA containing bromophenol blue dye (Sigma-Aldrich). Samples were then "snap-cooled" in a TC-1/PE 480 thermal cycler by heating to 95°C for 2 minutes followed by rapid cooling to 5°C, in order to ensure that the DNA was single-stranded with no secondary structure.

The samples were then loaded onto porous membrane combs (The Gel Company: San Francisco, CA) and placed in an ABI377 automated DNA Sequencer (Applied Biosystems) for electrophoresis. During the 9-hour electrophoresis run, ABI Prism 377-96 Data Collection Software (v. 2.6) was used to control the electrophoresis and laser detection. The software Sequencing Analysis v. 3.2. was used to extract sequence chromatograms and make a preliminary

determination of each of the 24 PCR products for the twenty individuals.

2.3 Protocol for Individuals Sequenced Using GeneChip Technology

2.3.1 PCR Amplification, Quantitation, and Pooling of PCR Products

The entire mitochondrial genome of ten Newfoundland individuals (the remaining eight individuals [1208, 10354, 10670, 10796, 11269, 11528, 12127, and 13016], and two individuals "re-sequenced" for comparison to the dideoxy methodology [12204 and 13392]) was amplified by PCR in 14 overlapping regions. "Re-sequencing" is an unfortunate term, as it implies that a sequence obtained by microchip sequencing has been obtained previously. A more appropriate term for this process is "iterative sequencing", as this indicates the repetitive production of a series of new homologous sequences from different individuals (Carr *et al.*, 2007). The PCR primer sequences, amplification length, and length of overlap are

indicated in Appendix C. These primers are a subset of those described in Section 2.2. PCR amplification and amplicon purification were performed as previously described.

Quantitation of each amplicon was performed with the Absorption Spectrophotometry Method outlined in the GeneChip CustomSeq Resequencing Array Protocol v. 2.0 (Affymetrix, Inc.). The concentration (in units of ng/μL) was measured in an Eppendorf BioPhotometer and entered into the PCR pooling Excel Spreadsheet downloaded from the Affymetrix website (www.affymetrix.com). Given the concentration and size (bp) of each amplicon, the spreadsheet determined the volume (μL) per PCR amplicon required to add to the pool to ensure equimolar amounts of each fragment.

2.3.2 DNA Fragmentation, Labeling, and Chip Hybridization

DNA fragmentation was performed with the procedure outlined in the Affymetrix GeneChip CustomSeq Resequencing Array Protocol v.2.0. For each experiment, a 100 μL master mix was prepared that contained 10.8 μL 10X Fragmentation Buffer (Affymetrix, Inc.), 88.05 μL of ddH_2O, and 1.05 μL of 3 U/μL Affymetrix Fragmentation Reagent.

Next, 3.7 µL of the Fragmentation cocktail was added to each DNA pool and placed in an Eppendorf Mastercycler thermocycler pre-heated to 37°C. The following thermal profile was then performed: 37°C for 15 minutes, 95°C for 15 minutes, hold at 4°C.

Fragmented DNA was labeled by the addition of 12 µL of 5X TdT buffer (Affymetrix, Inc.), 2 µL of 5 mM GeneChip DNA Labeling Reagent, and 3.4 µL of 30 U/µL Terminal Deoxynucleotidyl Transferase (TdT) (Affymetrix, Inc.) and the following thermal profile was performed: 37°C for 2 hours, 95°C for 15 minutes, hold at 4°C.

Prehybridization, hybridization, washing and scanning of the MitoChips were performed as described in the Affymetrix CustomSeq Resequencing array protocol v.2.0. Prehybridization was accomplished by applying a prehybridization buffer to the array and placing the array in the Affymetrix GeneChip hybridization Oven 640 at 45°C rotating at 60 RPM for 15 minutes. Next, the prehybridization buffer was removed and replaced with 200 µL of the fragmented and labeled DNA pool. The array was returned to the Hybridization Oven and incubated for 16 hours at 45°C at a rotation speed of 60 RPM.

Following hybridization the chips were washed on the Affymetrix GeneChip Fluidics Station with the preprogrammed DNA ARRAY-WS2 protocol (Affymetrix Microarray Suite v.3.0.2). Scanning of the chips

was performed in a 400 Hewlett Packard GeneArray Scanner. Analysis was done with the GeneChip DNA Analysis Software (GDAS) v.3.0.2.

2.4 Dideoxy Sequence Analysis

The complete mitochondrial genome contig was assembled and edited for each individual using Sequencher 4.1.2 software. Alignment with the revised Cambridge Reference Sequence (Andrews *et al.*, 1999) assisted in preliminary detection of SNPs. Consensus sequences for each individual were then exported as ASCII files into Eyeball Sequence Editor Program (ESEE v. 3.2, Cabot and Beckenbach, 1989) for further comparison and SNP analysis. The number of polymorphic sites, their codon position, and their status as transitions or transversions, or synonymous versus replacement sites, were recorded (Tables 1 and 2).

2.5 Microarray Sequence Analysis

Analysis of the complete mtDNA genome for each individual sequenced with the microarray method was performed using GDAS software (Affymetrix, Inc.). GDAS uses cell intensity data to make calls for every base position represented on the resequencing probe array. The algorithim uses the intensity data across multiple data files to improve its calling accuracy, and then computes a quality score for each call (Maitra *et al.*, 2004). The quality score is a representation of the call's statistical accuracy. Identification of SNPs is simplified as the results are presented along with the revised Cambridge Reference Sequence (Andrews *et al.*, 1999).

Although this software program is accurate enough to make the majority of calls, minimal manual revising was still required; many of the bases that were originally identified as "N" could in fact be called. This was accomplished by analysis of the probe intensity data. The probe intensity data displays cell intensity data for all the four possible calls at any one position on the chip. An algorithm was developed to determine which base at each position registered the highest intensity. The algorithm determined the extent of the intensity of each base when compared to the other three possible bases. This was

accomplished with the summation of sense and anti-sense probe intensities to give four base-specific intensity scores for each position. The highest and second-highest scores for each position were identified, along with the sum of intensities across all four bases. The difference between the two highest intensities was divided by the sum, which yielded a value defined as the differential signal-to-noise ratios (dS/N). This value expresses the confidence placed on each call. The approach is similar to that of Hacia *et al.* (1998), with standardization for total probe intensity. A cutoff value was determined for each individual, and calls that did not meet this criterion were left uncalled (N's).

2.6 Phylogenetic Analysis

Once complete sequences were obtained from all samples investigated from both techniques, the sequence information was imported into Phylogenetic Analysis Using Parsimony and other methods (PAUP*) v. 4.0 (Swofford, 2002). The Neighbor-Joining technique is a distance method that uses the absolute number of differences to identify the tree with the smallest sum of branch lengths. This technique was used for all comparisons performed in this

study. When an outgroup was required, the Evenki Sequence (Genbank accession number: AF346979) was used as it is known to fall outside of the European Clade (Ingman *et al.*, 2000). Bootstrap analysis was performed to determine statistical significance of the branch lengths. This technique randomized the data 10 000 times and scored how many times individuals were grouped together.

3.0 Results

3.1 DNA Sequences

A total of 16 570 bp of mtDNA sequence was obtained from each of 28 individuals from the three ethnic groups investigated. Analysis of the complete mtDNA genome identified 220 variable nucleotide sites (SNPs) (Table 1). Each SNP was verified by the sequencing of both strands. Of these, 71 sites were phylogenetically informative (Nei, 1987), that is, the SNP identifies a subset of at least two individuals, such that individuals in that subset are potentially more closely related to each other than to any others. Each of the remaining 149 SNPs is unique to a single individual. The position of these SNPs occurring in the coding region is presented in Table 2.

The number of nucleotide differences between each pair of individuals investigated is presented in Table 3. The mean pairwise difference among all pairs of individuals is 26.12. The smallest observed difference was a single nucleotide change, between two English individuals (10656 and 13016), and the largest was 56 changes, between an Irish and a French individual (12127 and 13392).

Upon examination of these 28 individuals, two samples (12516 and 802) were shown to have identical mtDNA sequences. As these

individuals were designated as different ethnicities (French and English, respectively), it was reasonable to assume that they were neither the same individual, nor maternally related individuals. Since these samples were provided, it is unknown if there was an error in the DNA banking records. These two samples were sequenced on different runs on different days, minimizing the likelihood of mislabeling. However, to address this possibility, new dilutions from the stock tubes were made and analysis of a ~1 500 bp region shown to be highly variable for these samples (containing 6 of the 15 SNPs or 40%) was performed. Again, these sequences were shown to be identical. An investigation is ongoing to determine if these two samples are indeed from the same individual. For the purposes of this study, the DNA samples were still included but as a single sequence named 12516/802 and with the information regarding ethnicity removed.

3.2 Patterns of Molecular Evolution

Similar numbers of variable sites (115 versus 105) were identified in the 13 protein-coding genes combined, as compared with

the non protein-coding regions (the Control Region, 12S RNA and 16S RNA genes. Nine tRNA genes had variable sites ($tRNA^{GLN}$, $tRNA^{TRP}$, $tRNA^{ALA}$, $tRNA^{ASP}$, $tRNA^{GLY}$, $tRNA^{ARG}$, $tRNA^{LEU}$, $tRNA^{GLU}$ and $tRNA^{THR}$) (Table 1). The greatest number of variable sites occurred in the control region (60), followed by cytochrome *b* (18), NADH dehydrogenase subunit 5 (16) and NADH dehydrogenase subunit 2 (14).

An index for the variability of each gene region is the number of variable sites divided by the total length in nucleotides of that gene (Table 1). The Control Region was the most variable (5.36 substitutions per 100 base pairs, or 5.36%), followed by the NADH3 (1.74%). COXIII was the least variable (0.64%). Variation at the remaining genes ranged between 0.65% and 1.59%. Transitions accounted for 189 of the 206 variable positions (91.7%), and transversions at the remaining 17 (8.3%)(Table 1).

Among the 115 identified coding region SNPs, 36 (30.8%), 7 (6.0%), and 72 (61.5%), occurred at the first, second, and third position, respectively (Table 2); among the first position changes, 4 of 36 (11.1%) occurred in leucine-coding codons. Of the 115 SNPs documented in the 27 individuals sequenced, 71 (61.7%) were synonymous changes, while the remaining 44 (38.3%) of these were nonsynonymous (Table 2). The largest number of nonsynonymous

changes was located in the COXI gene, where 7 of 10 substitutions are missense mutations. Two genes, ATP8 and ND4L, exhibited only synonymous changes. A discussion of the patterns of molecular evolution can be found in Appendix D.

3.3 Haplogroup Designation

Haplogroup assignment for the 27 Newfoundland individuals was determined by the presence or absence of specific restriction sites, as well as the particular signature of SNPs present (Torroni *et al.*, 1994, 1996; Richards *et al.*, 1998; Macaulay *et al.*, 1999; Richards *et al.*, 2000; Maca-Meyer *et al.*, 2001). These results are presented in Table 4, and are depicted in Fig. 3. Individuals assigned to haplogroups H and U were able to be further subdivided; Table 5 shows the haplogroup specific SNP sites for the sub-classification of haplogroup H (Brandstätter *et al.*, 2006), and the work by Maca-Meyer and colleagues in 2001 permitted individual 12204 to be further subdivided into haplogroup U6, and individuals 11983 and 11727 into U5b. Once haplogroup assignment was complete, the frequency at which each haplogroup occurred in the Newfoundland population could be

calculated. Comparisons of expected and observed frequencies in the Newfoundland population are presented in Table 6. Similar frequencies were shown to occur in the Newfoundland population when compared to the European population (Richards *et al.*, 2000; Sykes, 2001; Torroni *et al.*, 2006). To determine the statistical significance of these numbers, a Monte Carlo simulation (Roff and Bentzen, 1989) was carried out. Variations in haplogroup frequency were shown to be highly non-significant and these results are presented in Table 7.

3.4 Phylogenetic Analysis

Tree topologies that represent the relationships among the complete mtDNA sequences of the 27 Newfoundlanders investigated in this study are presented in Figure 4. A Neighbor-Joining tree was used, as this algorithm has been shown to reconstruct correct phylogenetic trees with a high probability when analyzing closely related samples (Saitou & Nei, 1987). Neighbor-Joining is a clustering method that attempts to find the tree with the minimal value for S (sum of branch lengths) (Jobling *et al.*, 2004). The Neighbor-Joining tree shown in Figure 4 is rooted with an Evenki sequence known to be outside the

European clades and has an S = 267. Bootstrap analysis was performed with PAUP* 4.0 (Swofford, 2002) to determine statistical support for the branching order. This test resampled 10 000 replicates. The bootstrap confidence levels are also presented in Figure 4.

Several important features are apparent from this analysis. First, there are several phylogenetically distinct clusters of individuals that are grouped together with strong support. Individuals 11785 and 11469 are grouped together with 100% bootstrap support and individuals 12127, 1017, and 11528 are also grouped together with 100% support. Individuals 12204, 13136, 11983, and 11727 are grouped together with reasonably strong support (65), while the remaining 17 individuals are all grouped together with 70% bootstrap support. This phylogram demonstrates the relationships between and among individuals belonging to the same phylogenetic cluster or haplogroup. Individuals 11785 and 11469 (haplogroup J) are grouped together with individuals 12127, 1017, and 11528 (haplogroup T) with strong support (72%). Individuals 12204, 11983, and 11727 (haplogroup U) and individual 13136 (haplogroup K) being grouped together with reasonably strong support (65%) also illustrates a relationship among haplogroups.

In order to compare the efficacy of complete sequences versus the use of the control region only, a Neighbor-joining tree using control region data only is presented in Figure 5. Bootstrap values determined from 10 000 replicates are also included. While some phylogenetic structure is preserved, key limitations are apparent. First of note is that individuals belonging to the same haplogroup are not always grouped together as seen in haplogroups U and H16. Also noteworthy, is the lack of information regarding relationships among haplogroups; we do not see the structure showing that T and J are closely related nor that haplogroup K is actually a subgroup of haplogroup U (Macaulay *et al.*, 1999; Maca-Meyer *et al.*, 2001).

Trees were then constructed with the complete sequences of the above 27 individuals and the addition of the revised Cambridge Reference sequence (Andrews *et al.*, 1999), three additional Newfoundland individuals of unknown ethnic origin (NF1-NF3: H.D. Marshall, personal communication, 2003), a Swedish individual (Arnason *et al.*, 1996), five Native American sequences (Mishmar *et al.*, 2003), five sequences of individuals known to be outside the European haplogroups (Maca-Meyer *et al.*, 2001), and 26 European and non-European individuals described by Ingman *et al.* (2000) (Figure 6). This expanded data set lends support to the

phylogeographic structure illustrated in Figure 4. The additional sequence information allows us to garner further information regarding relationships within and among phylogenetic clusters that were not evident in the data presented relying solely on the control region.

3.5 Comparison Between Microarray and Dideoxy Sequencing Methods

Sequence data with both microarray and automated dideoxy sequencing methods was available for two individuals (13392 and 12204). In comparison with the sequence data from the two sources, the chip confirmed all 25 known coding region SNPs for individual 13392 and all 24 known coding region SNPs for individual 12204. The microarray method also detected a previously unidentified SNP in each individual (Table 8).

Additionally, the microarray method detected several discrepancies upon comparison with the existing dideoxy sequence. For individual 13392, two differences were found. In one case, the dideoxy sequence indicated a SNP where the microarray showed none; in a second, it identified an editing error (Table 8). For individual

12204, the microarray sequence detected two discrepancies – in both instances the dideoxy sequence called a SNP where the microarray showed none. Moreover, the microarray sequence was able to detect a single editing error that was made in the manual sequence. An example of the microarray output is presented in Figure 7.

3.6 Analysis of Microarray platform

A total of 10 Newfoundland individuals were investigated with GeneChip microarrays: 1208, 10354, 10670, 10796, 11269, 11528, 12127, 13016, 12204, and 13392. Since there were two individuals where sequence information was available using both methods of analysis, a comparison was made between the two and the corresponding results were used to determine the efficiency and accuracy of the microarray platform (Tables 9 and 10). This platform was shown to be an extremely efficient method as evidenced by the 99.99% and 99.97% accuracy rate for individuals 13392 and 12204 respectively. There were, however, a number of base positions that demonstrated consistently poor hybridization characteristics (shared N's). These base positions are presented in Table 11. A total of 135 positions of 16 570 bp (0.81%) generated weak signals (N) across all

ten individuals. Of these, the 122 of 135 (90.4%) were C bases, often in regions containing two or more consecutive C bases. The remaining uncalled positions comprised of seven A residues and six T residues.

Table 1: The numbers of transitions and transversions, and the percent variation in mitochondrial genes that contained variable sites among 27 Newfoundland individuals.

Gene	Length in Base Pairs	Variable Sites	Percent Variation	Transitions	Transversions
CR	1120	60	5.36	58	2
12S	953	8	0.84	7	1
16S	1558	13	0.83	11	2
ND1	956	8	0.84	8	
$t^{RNA\text{-}Gln}$	71	1	1.41	1	
ND2	1041	14	1.34	12	2
$t^{RNA\text{-}Trp}$	67	1	1.49	1	
$t^{RNA\text{-}Ala}$	68	1	1.47	1	
COXI	1541	10	0.65	8	2
COXII	683	7	1.02	7	
ATP8	206	2	0.97	2	
ATP6	680	8	1.18	7	1
COXIII	780	5	0.64	4	1
$t^{RNA\text{-}Gly}$	67	1	1.49	1	
ND3	345	6	1.74	6	
$t^{RNA\text{-}Arg}$	64	2	3.13	1	1
ND4L	296	2	0.68	2	
ND4	1377	12	0.87	11	1
$t^{RNA\text{-}Leu}$	70	1	1.43	1	
ND5	1811	16	0.88	15	1
ND6	524	7	1.34	6	1
$t^{RNA\text{-}Glu}$	68	1	1.47	1	
Cyt*b*	1134	18	1.59	16	2
$t^{RNA\text{-}Thr}$	65	2	3.07	2	
Total		**206 (+ 14 indels)**		**189**	**17**

Table 2. Pattern of amino acid substitution in the complete mitochondrial genome of 27 Newfoundland humans.

Gene	Nucleotide Position	Change in Codon	Amino Acid Substitution
ND1	3315	GCC to ACC	Ala to Thr
	3347	CTA to CTG	N/A
	3479	AAA to AAG	N/A
	3506	ACC to ACT	N/A
	3743	CTA to CTG	N/A
	3756	CTA to TTA	N/A
	4215	TAT to CAT	Tyr to His
	4247	ATT to ATC	N/A
ND2	4528	ACA to ACT	N/A
	4687	GCT to GCC	N/A
	4768	ATA to ATG	N/A
	4819	GAG to GAA	N/A
	4823	ACC to GCC	Thr to Ala
	4916	AAC to GAC	Asn to Asp
	5093	ATT to TTT	Ile to Phe
	5119	CTA to CTG	N/A
	5146	ACG to ACA	N/A
	5197	TTA to TTG	N/A
	5276	TTC to CTC	Phe to Leu
	5299	ATC to ATT	N/A
	5436	ACC to ATC	Thr to Ile
	5470	ACG to ACA	N/A
COXI	5944	GAC to GAT	N/A
	5996	GCT to ACT	Ala to Thr
	6059	ATC to GTC	Ile to Val
	6075	GTC to GGC	Val to Gly
	6479	GTC to ATC	Val to Ile
	6488	CTC to ATC	Leu to Ile
	6775	CAT to CAC	N/A
	7027	GCC to GCT	N/A
	7244	ACC to GCC	Thr to Ala
	7268	GTA to ATA	Val to Met
COXII	7767	ATA to ATG	N/A
	7804	GTC to ATC	Val to Ile
	7896	TGG to TGA	N/A
	7911	GAG to GAA	N/A

Table 2 Cont'd

	7940	AAC to AGC	Asn to Ser
	8026	GCC to ACC	Ala to Thr
	8250	GGG to GGA	N/A
ATP8	8472	CCT to CCC	N/A
	8571	TAG to TAA	N/A
ATP6	8571	GGC to AGC	Gly to Ser
	8587	GTA to GAA	Val to Glu
	8696	ATG to ATA	N/A
	8793	CAC to TAC	His to Tyr
	8859	ACA to GCA	Thr to Ala
	8980	CAA to CGA	Gln to Arg
	9054	GCC to ACC	Ala to Thr
	9149	TTA to TTG	N/A
COXIII	9340	CTA to CTT	N/A
	9390	ACA to ATA	Thr to Met
	9476	GTT to ATT	Val to Ile
	9911	GAA to AAA	Glu to Lys
	9950	TGA to CGA	Trp to Arg
ND3	10083	ATC to ACC	Ile to Thr
	10191	TCC to TTC	Ser to Phe
	10237	ATT to ATC	N/A
	10252	TTT to TTC	N/A
	10393	GAC to GAT	N/A
	10397	ACC to GCC	Thr to Ala
ND4L	10549	ATA to ATG	N/A
	10597	ATA to ATG	N/A
ND4	10971	TGA to TGG	N/A
	11250	CTA to CTG	N/A
	11298	ACT to ACC	N/A
	11466	TTA to TTG	N/A
	11589	CTA to CTG	N/A
	11718	GGG to GGA	N/A
	11787	ATC to ATT	N/A
	11811	CTA to CTG	N/A
	11928	AAT to AAC	N/A
	12006	TGG to TGA	N/A
	12091	CTC to ATC	Leu to Ile
	12135	TCT to TCC	N/A
ND5	12371	CTG to CTA	N/A
	12500	ATG to ATA	N/A

Table 2 Cont'd

	12611	GTA to GTG	N/A
	12704	ATC to ATT	N/A
	12713	ATT to ATC	N/A
	12786	TCC to ACC	Ser to Thr
	12939	GCC to ACC	Ala to Thr
	13190	ACT to ACC	N/A
	13367	GGG to GGA	N/A
	13433	ATA to ATG	N/A
	13515	CAC to TAC	His to Tyr
	13616	ATT to ATC	N/A
	13707	GCA to ACA	Ala to Thr
	13779	ATC to GTC	Ile to Val
	13946	ATC to ATT	N/A
	14128	ACC to ATC	Thr to Ile
ND6	14166	CTC to CTT	N/A
	14178	TAT to TAC	N/A
	14181	GTA to GTG	N/A
	14232	GAT to GAC	N/A
	14271	TTG to TTC	Leu to Phe
	14318	TTA to CTA	N/A
	14586	GGT to GGC	N/A
Cyt*B*	14765	ACT to ATT	Thr to Ile
	14797	TTC to CTC	Phe to Leu
	14904	ATG to ATA	N/A
	14910	TAC to TAT	N/A
	14977	ATC to GTC	Ile to Val
	15027	CTC to CTA	N/A
	15042	GGG to GGA	N/A
	15325	ACA to GCA	Thr to Ala
	15451	CTT to ATT	Leu to Ile
	15529	TTA to CTA	N/A
	15606	AAA to AAG	N/A
	15630	TTA to TTG	N/A
	15631	CTA to TTA	N/A
	15654	ATA to ATG	N/A
	15669	CAT to CAC	N/A
	15720	TAT to TAC	N/A
	15757	ATC to GTC	Ile to Val
	15849	ACT to ACC	N/A

Table 3. Pairwise absolute distance matrix of nucleotide substitutions in the complete mitochondrial genome among 27 Newfoundland mtDNAs.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
1																											
2	14																										
3	10	18																									
4	34	42	38																								
5	28	38	30	50																							
6	6	14	10	36	30																						
7	4	18	14	38	30	10																					
8	9	17	11	39	33	7	13																				
9	12	22	18	42	38	12	16	15																			
10	28	34	30	48	38	28	32	31	36																		
11	34	44	38	56	48	32	38	37	40	46																	
12	9	17	13	37	31	9	13	12	17	31	39																
13	26	34	30	44	34	28	30	31	36	34	48	31															
14	24	34	30	46	38	28	28	31	34	40	40	31	38														
15	28	36	32	48	36	30	32	33	38	36	50	33	12	40													
16	2	16	12	36	30	8	6	11	14	30	36	11	28	26	30												
17	30	38	34	52	46	30	34	35	40	44	16	35	44	36	46	32											
18	27	34	30	44	38	28	28	31	34	40	40	31	38	6	40	26	36										
19	9	17	13	39	29	5	13	12	17	31	37	10	31	31	33	11	33	31									
20	6	14	8	36	30	6	10	9	14	30	36	9	28	28	30	8	32	28	9								
21	13	21	13	39	33	11	17	14	21	31	37	16	31	31	33	15	33	31	14	11							
22	12	14	16	38	36	12	16	15	20	34	42	15	32	32	34	14	36	30	15	12	19						
23	23	31	25	45	41	23	29	26	33	35	45	28	39	37	39	25	43	37	26	23	28	29					
24	5	13	9	35	29	1	9	6	11	29	33	8	27	27	29	7	31	27	6	5	12	11	24				
25	7	17	13	39	33	7	11	12	13	31	37	12	31	29	33	9	33	29	10	9	14	15	26	8			
26	27	35	31	49	43	29	31	32	37	43	19	32	41	33	43	29	3	33	32	29	32	33	42	28	32		
27	6	16	12	38	32	8	10	11	12	32	38	11	30	28	32	8	34	28	11	8	15	14	27	7	3	31	

Note: Individual assignments are as follows: **1**=157, **2**=1524JL, **3**=1351, **4**=13392, **5**=13136, **6**=13016, **7**=12765, **8**=12516/802, **9**=12218, **10**=12204, **11**=12127, **12**=1208, **13**=11983, **14**=11785, **15**=11727, **16**=11645, **17**=11528, **18**=11469, **19**=11269, **20**=10799, **21**=10796, **22**=10744, **23**=10670, **24**=10656, **25**=10354, **26**=1017, **27**=01MG1402

Table 4. Haplogroup-specific polymorphic sites in 27 Newfoundland mtDNAs as indicated by Torroni *et al*. (1996). Nucleotide positions have been modified in order to correspond with the revised Cambridge reference sequence (Andrews *et al*., 1999). Haplogroups indicated in brackets have been assigned according to SNP data as well as phylogenetic data.

Individual	Nucleotide Position																	Haplogroup
	+1715	+4532	+4579	-7025	-8248	+8994	+9055	-10028	-10393	-10397	-12307	-13365	+13704	-15606	+15924	+16064	-16388	
157 (EN)	+	+	+	-	-	+	+	-	-	-	-	-	+	-	+	+	-	H
1524JL (EN)	+	+	+	-	-	+	+	-	-	-	-	-	+	-	+	+	-	H
1351 (IR)	+	+	+	-	-	+	+	-	-	-	-	-	+	-	+	+	-	H
13016 (EN)	+	+	+	-	-	+	+	-	-	-	-	-	+	-	+	+	-	H
12765 (EN)	+	+	+	-	-	+	+	-	-	-	-	-	+	-	+	+	-	H
12516/802*	+	+	+	-	-	+	+	-	-	-	-	-	+	-	+	+	-	H
12218 (FR)	+	+	+	-	-	+	+	-	-	-	-	-	+	-	+	+	-	H
1208 (EN)	+	+	+	-	-	+	+	-	-	-	-	-	+	-	+	+	-	H
11645 (EN)	+	+	+	-	-	+	+	-	-	-	-	-	+	-	+	+	-	H
11269 (EN)	+	+	+	-	-	+	+	-	-	-	-	-	+	-	+	+	-	H
10799 (IR)	+	+	+	-	-	+	+	-	-	-	-	-	+	-	+	+	-	H
10744 (IR)	+	+	+	-	-	+	+	-	-	-	-	-	+	-	+	+	-	H
10656 (EN)	+	+	+	-	-	+	+	-	-	-	-	-	+	-	+	+	-	H
10354 (EN)	+	+	+	-	-	+	+	-	-	-	-	-	+	-	+	+	-	H
01MG1402 (EN)	+	+	+	-	-	+	+	-	-	-	-	-	+	-	+	+	-	H
10796 (IR)	-	-	+	-	-	+	+	-	-	-	-	-	+	-	+	+	-	? (H)
10670 (EN)	+	+	+	+	-	+	+	-	+	-	-	-	+	-	+	+	+	(I)

Table 4 Cont'd

13136 (IR)	+	+	+	+	-	+	-	-	+	-	+	-	+	-	+	+	-	K
11983 (IR)	+	+	+	+	-	+	+	-	-	-	+	-	+	-	+	+	-	U
12204 (FR)	+	+	+	+	-	+	+	-	-	-	+	-	+	-	+	+	-	U
11727 (IR)	+	+	+	+	-	+	+	-	-	-	+	-	+	-	+	+	-	U
12127 (IR)	+	+	+	+	-	+	+	-	-	-	-	+	+	+	-	+	-	T
11528 (EN)	+	+	+	+	-	+	+	-	-	-	-	+	+	+	-	+	-	T
1017 (EN)	+	+	+	+	-	+	+	-	-	-	-	+	+	+	-	+	-	T
11469 (EN)	+	+	+	+	-	+	+	-	+	-	-	-	-	-	+	-	-	J
11785 (EN)	+	+	+	+	-	+	+	-	+	-	-	-	-	-	+	-	-	J
13392 (FR)	+	+	+	+	-	+	+	-	-	-	-	-	+	-	+	+	-	Other (A)

* Ethnicity information removed.

Figure 3 An adaptation of figure 6 from "The Seven Daughters of Eve" (Sykes, 2001) to illustrate which of the seven haplogroups of Europeans the Newfoundland individuals belong. Each of the circles represents a particular mtDNA sequence, and the area of the circle is proportional to the number of people who share this sequence. The lines joining the circles represent mutations in the mtDNA sequence, and the longer the line between two circles, the more mutations separate those sequences. The figure demonstrates not only the relationships among sequences in the same haplogroup, but also the relationships between haplogroups.

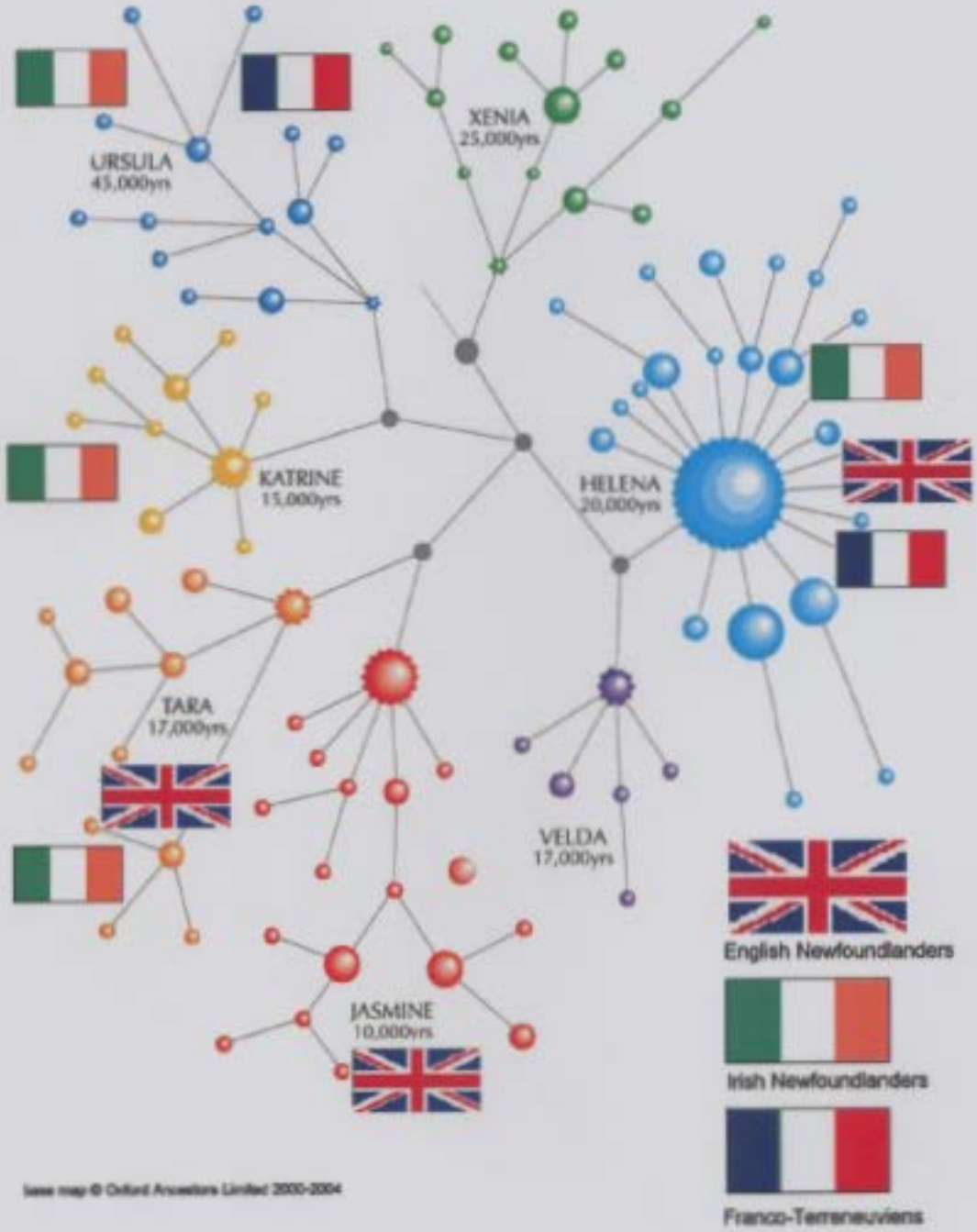

Table 5. Haplogroup-specific polymorphic sites in 16 Newfoundland mtDNAs used to further subdivide haplogroup H as indicated by Brandstätter *et al*. (2006). Three individuals were not able to be sub-classified and therefore are represented as H.

	Nucleotide Position											
Individual	73	456	477	2706	3010	4336	6776	7028	10394	16162	16304	**Haplogroup**
11645 (EN)	G	C	T	A	A	T	T	C	C	G	T	H1a
12765 (EN)	G	C	T	A	A	T	T	C	C	G	T	H1a
157 (EN)	G	C	T	A	A	T	T	C	C	G	T	H1a
01MG1402(EN)	A	C	C	A	A	T	T	C	C	A	T	H1c
10354 (EN)	A	C	C	A	A	T	T	C	C	A	T	H1c
12218 (FR)	A	C	C	A	A	T	T	C	C	A	T	H1c
10796 (IR)	A	C	T	A	G	T	C	C	C	A	T	H3
12516/802*	A	C	T	A	G	T	C	C	C	A	T	H3
1524 (EN)	A	T	T	A	G	T	T	C	C	A	C	H5
10744 (IR)	A	T	T	A	G	C	T	C	C	A	C	H5a
13016 (EN)	A	C	T	A	G	T	T	C	T	A	T	H16
10656 (EN)	A	C	T	A	G	T	T	C	T	A	T	H16
11269 (EN)	A	C	T	A	G	T	T	C	T	A	T	H16
1208 (EN)	A	C	T	A	G	T	T	C	C	A	T	H
1351 (IR)	A	C	T	A	G	T	T	C	C	A	T	H
10799 (IR)	A	C	T	A	G	T	T	C	C	A	T	H

* Ethnicity information removed

Table 6. A comparison of haplogroup frequencies between 27 Newfoundland individuals and modern Europeans as referenced by Sykes, 2001.

	EN	IR	FR	Ethnicity Unknown*	% of NF individuals	% of modern Europeans
H	10	4	1	1	59	47
T	1	2	0	0	11	9
I	1	0	0	0	4	0[1]
J	0	2	0	0	7	17
U	0	2	1	0	11	11
K	0	1	0	0	4	6
V	0	0	0	0	0	5
X	0	0	0	0	0	6
A	0	0	1	0	4	0

* Individual 12516/802 also belongs to Haplogroup H but ethnicity information has been removed for reasons described in Chapter 2.

[1] Although Sykes, 2001 does not mention the frequency of Haplogroup I, previous work looking at three European populations, has determined the frequency to be approximately 2% (Torroni *et al.*, 1996).

Table 7. A comparison of the Newfoundland haplogroup population structure to the European population using a Monte Carlo Simulation (Roff and Bentzen, 1989) to determine statistical significance.

	H	T	J	U	K	V	X
1	16	3	3	3	1	0	0
2	12	2	4	3	2	1	2

Calculated χ^2 value from original matrix: 4.25

RESULTS OF SIMULATIONS

Number of replicates in simulation :5000
Number of replicates which exceed original :3681
Probability of exceeding original χ^2 by chance :0.7362 +-0.0062

Figure 4 The phylogenetic relationships among the whole mitochondrial genomes of 27 Newfoundlanders. The neighbor-joining tree presented has been rooted with an Evenki sequence known to be outside of the European Clade. The branch lengths (S = 267) and bootstrap values (10 000 replicates) are indicated.

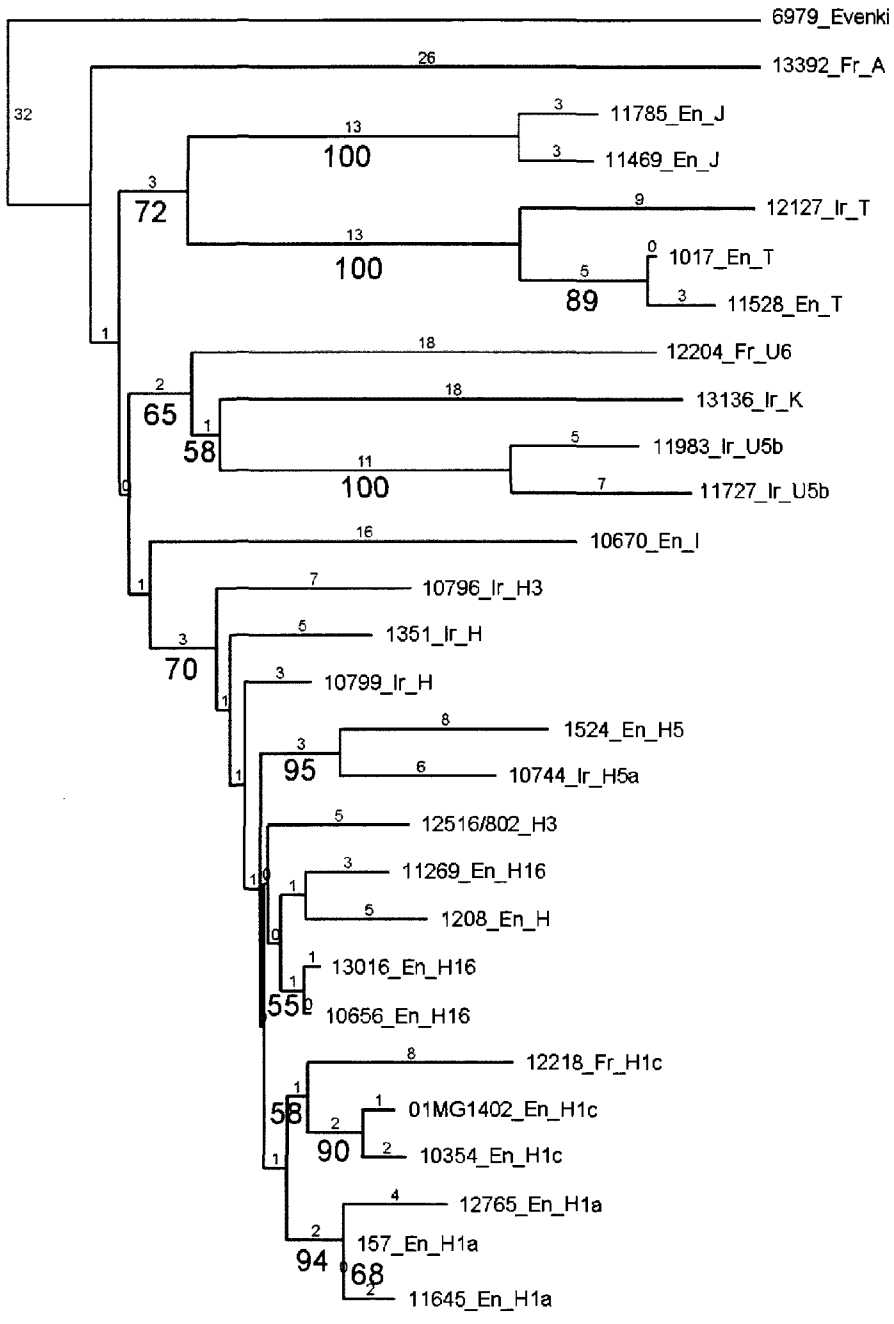
6979_Evenki
13392_Fr_A
11785_En_J
11469_En_J
12127_Ir_T
1017_En_T
11528_En_T
12204_Fr_U6
13136_Ir_K
11983_Ir_U5b
11727_Ir_U5b
10670_En_I
10796_Ir_H3
1351_Ir_H
10799_Ir_H
1524_En_H5
10744_Ir_H5a
12516/802_H3
11269_En_H16
1208_En_H
13016_En_H16
10656_En_H16
12218_Fr_H1c
01MG1402_En_H1c
10354_En_H1c
12765_En_H1a
157_En_H1a
11645_En_H1a

Figure 5 The phylogenetic relationships among 27 Newfoundlanders using mitochondrial control region sequence data only. The neighbor-joining tree presented indicates branch lengths (S = 74) and bootstrap values garnered from 10 000 replicates.

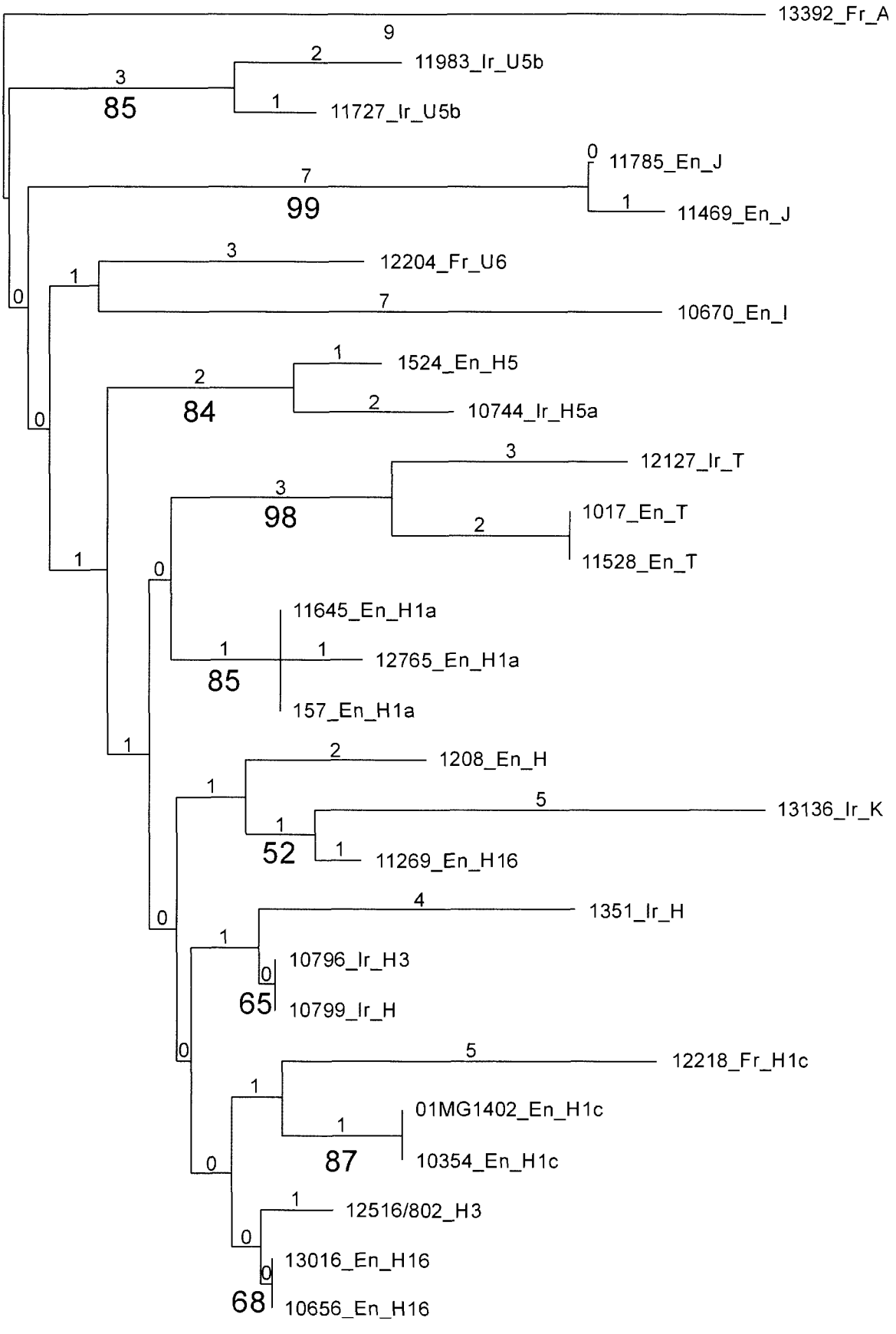
13392_Fr_A
9
2
11983_Ir_U5b
3
85
1
11727_Ir_U5b
0
11785_En_J
7
99
1
11469_En_J
3
12204_Fr_U6
1
0
7
10670_En_I
1
1524_En_H5
2
84
2
10744_Ir_H5a
0
3
12127_Ir_T
3
98
1017_En_T
2
11528_En_T
1
0
11645_En_H1a
1
1
12765_En_H1a
85
157_En_H1a
1
2
1208_En_H
1
5
13136_Ir_K
1
52
1
11269_En_H16
4
1351_Ir_H
0
1
10796_Ir_H3
0
65
10799_Ir_H
0
5
12218_Fr_H1c
1
01MG1402_En_H1c
1
87
0
10354_En_H1c
1
12516/802_H3
0
13016_En_H16
0
68
10656_En_H16

Figure 6 The phylogenetic relationships among 69 individuals of both European and non-European descent. The tree presented is a neighbor-joining phylogram with the Evenki, Buriat, and Khirgiz sequences defined as the outgroup. Bootstrap analysis (50% majority-rule) using the neighbor-joining method and 10 000 replicates are indicated.

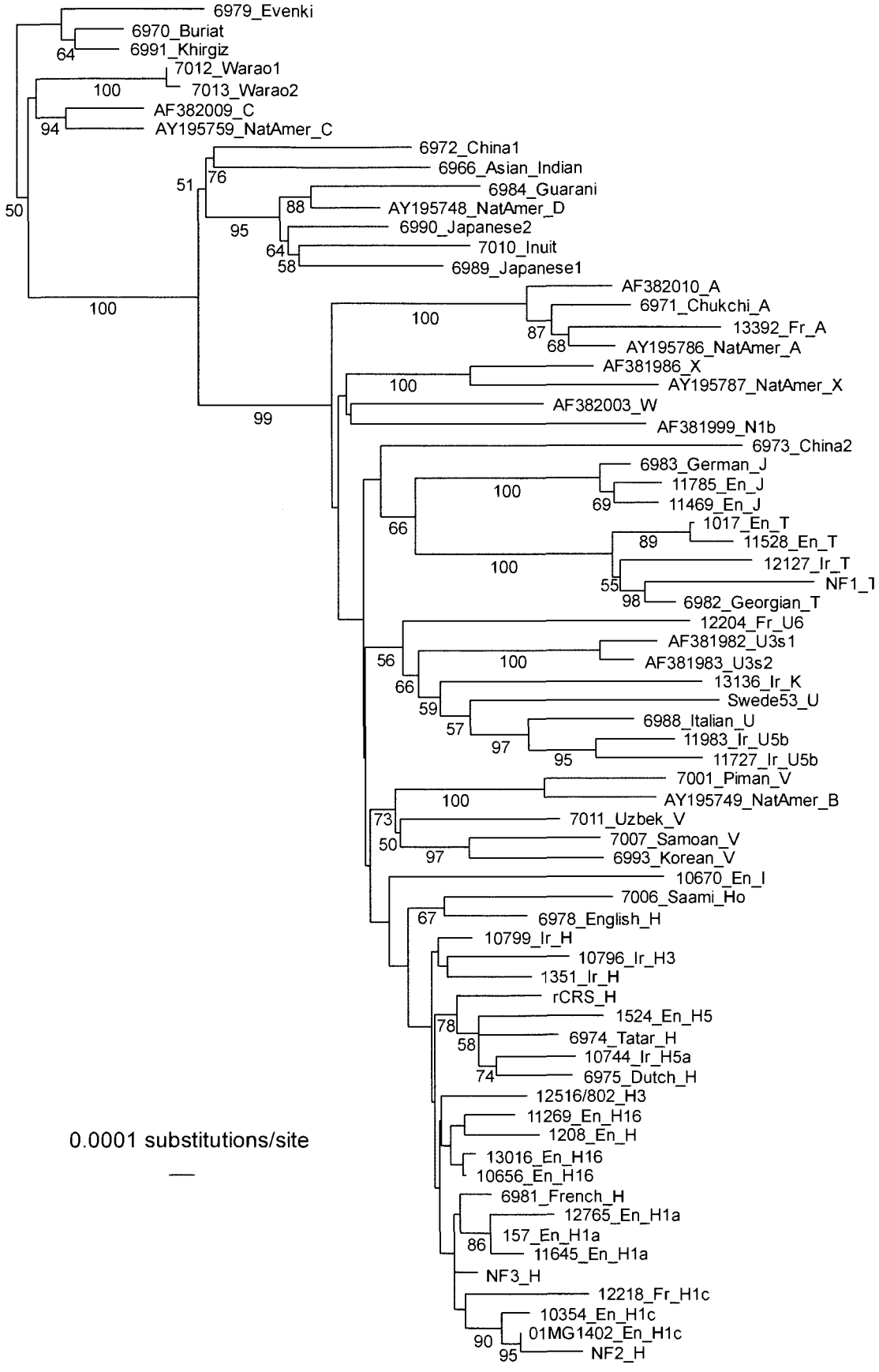
6979_Evenki
6970_Buriat
6991_Khirgiz
7012_Warao1
7013_Warao2
AF382009_C
AY195759_NatAmer_C
6972_China1
6966_Asian_Indian
6984_Guarani
AY195748_NatAmer_D
6990_Japanese2
7010_Inuit
6989_Japanese1
AF382010_A
6971_Chukchi_A
13392_Fr_A
AY195786_NatAmer_A
AF381986_X
AY195787_NatAmer_X
AF382003_W
AF381999_N1b
6973_China2
6983_German_J
11785_En_J
11469_En_J
1017_En_T
11528_En_T
12127_Ir_T
NF1_1
6982_Georgian_T
12204_Fr_U6
AF381982_U3s1
AF381983_U3s2
13136_Ir_K
Swede53_U
6988_Italian_U
11983_Ir_U5b
11727_Ir_U5b
7001_Piman_V
AY195749_NatAmer_B
7011_Uzbek_V
7007_Samoan_V
6993_Korean_V
10670_En_I
7006_Saami_Ho
6978_English_H
10799_Ir_H
10796_Ir_H3
1351_Ir_H
rCRS_H
1524_En_H5
6974_Tatar_H
10744_Ir_H5a
6975_Dutch_H
12516/802_H3
11269_En_H16
1208_En_H
13016_En_H16
10656_En_H16
6981_French_H
12765_En_H1a
157_En_H1a
11645_En_H1a
NF3_H
12218_Fr_H1c
10354_En_H1c
01MG1402_En_H1c
NF2_H
0.0001 substitutions/site

Table 8. A comparison of identified single nucleotide polymorphisms using two different sequencing methodologies (manual (dd) vs. microarray (chip)) in two Newfoundland individuals.

12204				**13392**			
nt position	Ref seq	dd seq	chip seq	nt position	Ref seq	dd seq	chip seq
750	A	G	G	750	A	G	G
794	T	A	A	1438	A	G	G
1193	T	C	C	1736	A	G	G
1438	A	G	G	1809	A	C	C
1692	A	T	T	2638	A	C	C
2706	A	G	G	2706	T	G	G
3347	A	G	G	3315	T	A	A
4768	A	G	G	4247	A	C	C
5119	A	G	G	4768	G	G	G
5470	G	A	A	4823	T	G	G
7027	C	T	T	7027	A	T	T
7804	G	A	A	**7896**	**G**	**G**	**A**
8472	**T**	**T**	**C**	8026	G	A	A
8587	**T**	**A**	**T**	**8636**	**C**	**T**	**C**
8636	**C**	**T**	**C**	8790	C	T	T
8859	A	G	G	8859	A	G	G
11466	A	G	G	11718	G	A	A
11718	G	A	A	12006	G	A	A
11928	T	C	C	12091	C	A	A
12307	A	G	G	12704	C	T	T
12371	G	A	A	**12713**	**T**	**T**	**C**
14178	A	G	G	12939	G	A	A
14271	**C**	**C**	**G**	14765	C	T	T
14765	C	T	T	14910	C	T	T
15042	G	A	A	15325	A	G	G
15325	A	G	G	15669	T	C	C
15529	T	C	C	15849	T	C	C
15631	C	T	T				

Figure 7 An example the mtDNA re-sequencing microarray for individual 13392 (Carr *et al.*, 2007). The region shown tiles a reference sequence of 15 452 bases (it excludes the Control Region) in a 160 row x 488 column array. Both the sense and anti-sense strands are tiled onto the array for a total of >31 Kb. Each nucleotide position is represented in a vertical block of 4 cells in 5 rows (A, C, G, T and a blank). In each block, the cell with the highest intensity of DNA binding identifies the base present at that position. In the magnified view, the sequence of bases is easily read as the left-to-right order of successive brightest pseudo-colour squares.

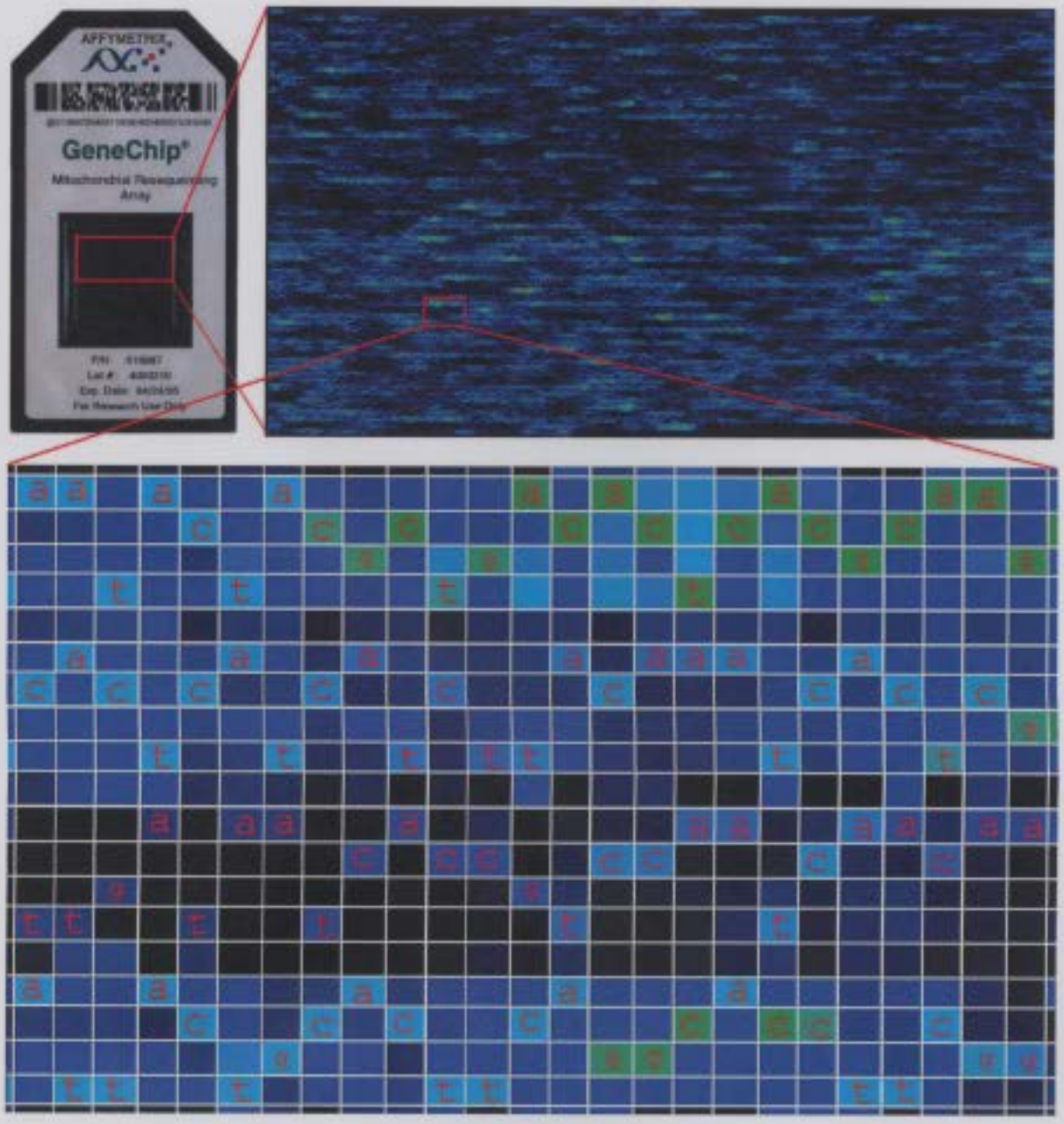
AFFYMETRIX
GeneChip®
Mitochondrial Resequencing
Array

Table 9. Efficiency, accuracy, and error of the microarray data for NF individual 13392 (dS/N = 0.20)

	Correct	Incorrect
Microarray: SNP	**24** + *3* = 27 **0.16** + *0.02* = 0.18%	**0** + *0* = 0 0%
Microarray: no SNP	**14237** + *1186* = 15423 **92.14** + *7.67* = 99.81%	**0** + *2* = 2 **0** + *0.01* = 0.01%
	14261 + *1189* = 15450 (99.99%)	**0**(0%) + 2(0.01%)*N*

Table 10. Efficiency, accuracy, and error of the microarray data for NF individual 12204 (dS/N = 0.20)

	Correct	Incorrect
Microarray: SNP	**25** + *0* = 25 **0.16** + *0.00* = 0.16%	**0** + *0* = 0 0%
Microarray: no SNP	**15079** + *344* = 15422 **97.59** + *2.23* = 99.82%	**0** + *4* = 4 **0** + *0.03* = 0.03%
	15104 + *344* = 15448(99.97%)	**0**(0%) + 4(0.03%)*N*

Table 11. A list of sites that demonstrated poor hybridization across all ten samples that were sequenced via the GeneChip method (1208, 10354, 10670, 10796, 11269, 11528, 12127, 13016, 12204, and 13392).

Nucleotide Position	Reference Sequence	Nucleotide Position	Reference Sequence
231	C	4867	C
232	C	4870	T
233	C	4932	A
385	C	5325	C
386	C	5600	C
387	C	5742	C
388	C	6275	C
390	C	6827	C
391	C	6828	C
1111	C	6829	C
2535	C	6830	C
2598	C	6831	A
2914	C	6896	C
2956	C	7248	C
2957	C	7458	C
2997	C	7702	C
2998	C	7711	C
3000	T	7988	C
3001	C	8369	C
3002	C	8370	C
3003	C	8685	C
3004	C	8956	C
3005	A	8957	C
3006	T	8984	C
3015	C	8985	C
3323	C	8986	C
3566	C	9622	C
3567	C	9623	C
3568	C	9624	C
3678	C	9708	C
3679	C	10365	C
4191	C	10366	C
4192	A	10367	C
4195	A	10941	C
4199	C	10942	C
4225	C	10943	C
4308	C	11102	C
4309	C	11103	C
4310	C	11297	C
4641	C	11298	C
4661	C	11299	C
4662	C	11513	C
4663	C	11514	C
4732	C	11515	C

Table 11 Cont'd

Nucleotide Position	Reference Sequence	Nucleotide Position	Reference Sequence
4733	C	11666	C
4866	C	11667	C
12397	C	13675	C
12398	C	13768	C
12482	C	13769	C
12483	C	13847	C
12484	C	13919	C
12486	C	13926	T
12487	C	13960	C
12488	C	14200	C
12557	C	14205	A
12558	C	14206	T
13076	C	14237	C
13077	C	14238	C
13107	C	14239	C
13111	C	14240	C
13112	C	14242	C
13113	T	14243	C
13114	A	14755	C
13184	C	14971	C
13185	C	14972	C
13585	C		

4.0 Discussion

4.1 Comparison of Sequencing Methods

Sequence data were available from both the dideoxy method and the microarray technology for two individuals: a comparison of the results indicates the accuracy of each method.

For individual 12204, the dideoxy method identified 26 SNPs. The microarray data confirmed 24 of these, and identified a previously undetected SNP. The dideoxy sequence and the microarray sequence were discordant at four positions out of 15452. Two of these (nt 8587 and 8636) were found to be inaccuracies in the dideoxy sequencing methodology. At nt 8587, the dideoxy sequence shows an A peak, and the chip sequence shows a T, as in the reference sequence. The chip sequence was deemed to be correct, as the absolute intensity for the T residue is 43.0% better than that for the A residue, and has a quality score of 148.7. At nt 8636, both the chip and the reference sequence show C for this position, but the dideoxy sequence shows a T. Again, the chip sequence is taken to be correct, with the C residue 58.8% more intense than that for the T, and a quality score of 222.5. At these positions the dideoxy sequence is not of high quality. With these

changes, positions 8587 and 8636 are no longer variable sites. Comparison of the two sequencing methods identified an editing error at nt 8472, where the reference sequence shows a T and the microarray sequence shows a C. Re-examination of the dideoxy sequence data shows a strong C peak. The microarray sequence also identified a previously undetected SNP at nt position 14271. Both the reference sequence and the dideoxy sequence show C for this position but the chip sequence determines a G at this position. While the G residue does have a relatively low quality score (49.1), it is taken to be the correct base for this position as it is 31.5% more intense than that of the C residue. As before, the dideoxy sequence is not of high quality for this position.

For individual 13392, the dideoxy method identified 25 SNPs. The microarray sequence data confirms 24 of these, as well as identifying a previously undetected SNP, and a single editing error. The microarray data identifies a single dideoxy sequencing inaccuracy at nt position 8636, at which the dideoxy method shows a T, but the chip and reference sequences are both C. C was determined to be the correct base at this position as the chip sequence showed the absolute intensity of C to be 46.8% more intense than T with a quality score of 190.2. The manual editing error occurred at nt 7896, which the

microarray identified as an A. Re-examination of the data showed an unambiguous A, which had been edited to a G. The microarray sequence identified a previously undetected SNP at nt 12713, where both the reference sequence and the dideoxy sequence are T at this position but the chip shows a clear C. Again the microarray sequence was taken to be correct as the absolute intensity for the C residue was 38.4% more intense than that of the T residue, with a quality score of 120.8.

Thus, employment of the microarray technology not only allowed the identification of previously unidentified SNPs, but was also able to clarify previous ambiguities. The efficiency and accuracy of re-sequencing can be compared with that previously obtained for the same two individuals by conventional dideoxy nucleotide sequencing (Flynn and Carr, personal communication). For individual 13392, use of the dS/N quality-control algorithm with a 20% threshold rule called 15212 of 15452 bases correctly (98.45% efficiency), including all 25 known SNPs (100.00% accuracy). Of the remaining 240 positions with dS/N < 0.20 and therefore initially called as "N", in 232 cases the quartet cell with the highest absolute signal strength corresponded to the correct base as identified by dideoxy sequencing. In the remaining eight cases, the quartet cell with the highest absolute signal was one

other than that identified by dideoxy sequencing and therefore remain as N's. In order to determine the overall efficiency, high-confidence calls (ds/N > 20%) are combined with the low-confidence calls (ds/N < 20%); excluding the low confidence N's gives (14261 + 1189) / 15452 = 99.99% overall efficiency (Table 9). For individual 12204, use of the dS/N quality-control algorithm with a 20% threshold rule called 15258 of 15452 bases correctly (98.74% efficiency), including all 24 known SNPs (100.00% accuracy). Of the remaining 194 positions with dS/N < 0.20 and therefore initially called as "N", in 170 cases the quartet cell with the highest absolute signal strength corresponded to the correct base as identified by dideoxy sequencing. In the remaining 24 cases, the quartet cell with the highest absolute signal was one other than that identified by dideoxy sequencing. Combining the high-confidence calls with the low-confidence calls and excluding the low confidence N's gives (15104 + 344) / 15452 = 99.97% overall efficiency (Table 10).

4.2 SNP Diversity Among Individuals Within Newfoundland Ethnic Groups

In the current investigation, the ratio of transitions to transversions was found to be approximately 11:1. A high transition to transversion rate is indicative of a population of recent origin, such as the European population of humans. This pattern is consistent with findings in previous mitochondrial DNA studies (e.g. Carr *et al.*, 1995; Marshall and Baker, 1998).

The substantial genetic diversity within each Newfoundland ethnic group is documented in the pairwise data matrix (Table 3). The greatest pairwise nucleotide difference between two English individuals was 43 between 11528 & 10670, whereas individuals 10656 and 13016 differed by only a single SNP. The greatest pairwise nucleotide difference between two Irish individuals was 50 (11727 & 12127), and the smallest number of differences was 8 (10799 & 1351). The largest number of pairwise nucleotide difference between two French individuals was 36 (12204 and 12218), however, if we still include individual 13392 as of French origin (see 4. 4) that number increases to 48 (13392 & 12204). The French individuals as a whole appear to

be part of a much more variable population when compared to the English and Irish, despite the fact that only four French individuals were available for analysis.

4.3 Relationships Among Individuals Indicated by Phylogenetic Analysis

The genome phylogeny of the 27 Newfoundlanders shows that, although the majority fall into one of the haplogroups defined previously by analysis of CR sequence data, other individuals are not readily assigned into these predetermined groups. Haplogroups are determined by the presence or absence of particular restriction sites; individuals who display the same signature of these restriction sites are grouped into the same Haplogroup (Torroni *et al.*, 1994, 1996; Richards *et al.*, 1998; Macaulay *et al.,* 1999; Richards *et al.*, 2000; Maca-Meyer *et al.*, 2001). The most numerous haplogroup was H (Figures 3, 4 and 5), which has been further subdivided into more than a dozen subgroups in recent publications (Loogväli *et al.,* 2004; Brandstätter *et al.,* 2006; Roostalu *et al.,* 2007). A total of 16 Newfoundlanders were assignable to haplogroup H. Of these, three

were assignable to a monophyletic lineage equivalent to subgroup H1a, two to one equivalent to H5, both with high statistical support (94% and 95% respectively). Three were assignable to H1c, but with less support (58%). Those assignable to H3 and H16 are problematic. A single SNP (T6776C) defines Haplogroup H3, yet Figures 4 and 5 demonstrate that the two individuals designated H3 are not each others closest relative. This supports the idea that subhaplogroup H3 represents a multifurcation node (Torroni *et al.,* 2006). Haplogroup H16 appears to be a monophyletic group; interestingly, it includes an individual that does not carry the defining C10394T SNP. This finding illustrates the greater accuracy obtained with complete mitochondrial DNA sequences to determine true phylogenetic relationships.

Another nine Newfoundlanders were assignable to haplogroups J, K, T and U, each as parts of monophyletic lineages, with the U and K lineages grouped as predicted. Individuals belonging to haplogroup T and J are supported by a bootstrap of 100%, while the branch separating these two groups received strong support with a value of 72. Individuals belonging to haplogroups U and K are grouped together with a bootstrap value of 65; within this cluster, individuals belonging to subhaplogroup U5b grouped together with 100% bootstrap support.

An expanded data set with all 27 Newfoundland individuals was analyzed in the context of 42 other individuals of both European and non-European descent. The Neighbor-Joining tree for the analysis of the 69 individuals is presented in Figure 6. There are two main groups shown that reach statistical significance with a bootstrap value of 99; these groups designate the separation of Europeans and closely related haplogroups, with haplogroups that have been associated with Native Americans. The addition of a further 42 individuals provided increased confidence in the previous identification of monophyletic lineages. Individuals belonging to haplogroups J and T reach significance with a bootstrap value of 100, while the branch separating these groups shows support with a value of 66. Individuals belonging to haplogroups U and K are grouped together with moderate support with a bootstrap value of 56, while the two individuals further subdivided into haplogroup U5b are supported with a value of 95. Individuals belonging to Haplogroup H are not grouped together with strong bootstrap support. This finding supports that shown in figure 4 where only a few haplogroup H subgroups were shown to be monophyletic. This is probably due to the similarity of the sequences, which causes problems with the algorithm and are therefore collapsed back towards the root.

Interestingly, an individual that previously was not placed into a haplogroup, 13392, was grouped with a Chukchi individual along with two Native American sequences with 100% bootstrap support. The Chukchi are the largest native nation on the Asian side of the North Pacific (Starikovskaya *et al.*, 1998). They populate areas of Siberia and northern North America in areas such as Alaska.

4.4 Haplogroup Analysis

The majority of Newfoundland individuals (16 of 27 or 59%) were found to belong to haplogroup H (Tables 4 and 5). This is the most prevalent haplogroup, constituting 47% of modern Europeans (Sykes, 2001). Six of these sixteen individuals (38%) belong to the H1 subhaplogroup, which comprises approximately 30% of haplogroup H and 13% of the total European mtDNA pool (Loogväli *et al.,* 2004). Three Newfoundland individuals (11%) were found to belong to haplogroup T, as compared with 9% of modern Europeans; this haplogroup is found predominately in western Britain and Ireland (Sykes, 2001). Two Newfoundland individuals (7%) belong to haplogroup J, which has two distinct branches, one residing in Spain

and Portugal, the other in western Britain, and constitutes 17% of modern Europeans. Three Newfoundland individuals (11%) belong to haplogroup U. Haplogroup U is found all over Europe with the majority in western Britain and constitutes 11% of modern Europeans (Sykes, 2001). A single individual (4%) was found to belong to haplogroup K. Haplogroup K accounts for 6% of modern Europeans and is found mostly in the Mediterranean.

Two Newfoundland individuals were unassignable to any of the pre-defined haplogroups, which illustrates a disadvantage of haplogroup analysis versus analysis of complete genomes. The reliance on the control region for signatures of restriction sites does not allow for the same resolution of relationships that whole genome sequencing provides. Furthermore, previous work has suggested that attempts to classify lineages based on mutations found in the hypervariable segments have been hindered by the frequent occurrence of mutational hot-spots or fast-evolving nucleotide sites (Richards *et al.*, 2000; Loogväli *et al.*, 2004).

Individual 10670 does not fit any of the canonical CR haplogroup signatures. This individual has characteristics of both haplogroups I and J (Torroni *et al.*, 1996), and is a closer match to the former. Individual 10670 exhibits the addition of a *Dde*I site at position 10394;

this position is a shared SNP between haplogroups I and J. This individual also has a SNP at position 16389 resulting in a *Bam*HI site, which is characteristic of haplogroup I. However, this individual does not exhibit any of the other defining SNPs that represent this haplogroup. Macaulay *et al.* (1999) as well as Maca-Meyer *et al.* (2001) lists nucleotide positions of SNPs that define each of the major haplogroups of Europeans. When referring to this list, individual 10670 exhibits several of the defining mutations for haplogroup I, but lacks two characteristic SNPS (A4529T, or 10034C). The complete mtDNA sequence was entered into a BLAST search (NCBI) to determine if any other individuals had been typed that showed similar mutations. An individual of Finnish ancestry assigned to haplogroup I was found to be similar (Genbank accession number DQ489516; Finnilä *et al.*, 2000) and when analyzed with the 27 Newfoundland individuals, was found to be most closely related to individual 10670 with bootstrap support of 81% (results not shown). Therefore, for the purposes of this investigation, individual 10670 has been designated haplogroup I.

Individual 13392 was not assigned to one of the haplogroups identified by Torroni *et al.* (1996), but they did have an individual with a similar signature and they referred to this pattern of SNPs as "Other". Recall that this individual was shown to be most closely

related to a Chukchi individual (bootstrap support of 100%; Figure 6) who had been assigned to the A haplogroup (Torroni and Wallace, 1995; Starikovskaya *et al.*, 1998; Starikovskaya *et al.*, 2005). The pattern of SNPs associated with the "Other" haplogroup differs from the A haplogroup pattern at only one site – the "Other" signature has a gain of the 1715 *Dde*I site. Haplotype analysis shows that individual 13392 belongs to haplogroup A. The A haplogroup has been exclusively associated with northeastern Eurasian natives and North American First Nations peoples (Mishmar *et al.*, 2003; Reidla *et al.*, 2003). A hypothesis is that individual 13392 is the maternal descendant of a First Nations inhabitant of Newfoundland, a daughter of a French father who was adopted into the French community (Carr *et al.*, 2007). In an attempt to substantiate this hypothesis, additional demographic information was obtained. Although this individual was recruited for participation in a Y chromosome study, the maternal ancestry was known for 3 generations - all have French surnames. However, this individual is from the west coast of the island, which is a geographical region with a strong history of Mi'kmaq propagating with French men (Story *et al.*, 1999). With this knowledge and the haplogroup analysis provided in this study, it is very likely that beyond the great-grandmother was a Mi'kmaq maternal ancestor.

Current literature supports the idea that there is one common mitochondrial "Eve" of whom we are all descendants (Cann *et al.*, 1987). More recently, all non-African humans have a common ancestor of ~37,500 years ago. It has been suggested that there are seven "Daughters of Eve", each of whom corresponds to one of the seven haplogroups of Europeans: U, X, H, V, T, K, and J (Torroni *et al.*, 1996; Sykes, 2001). Five of these seven haplogroups were found among the 27 Newfoundland individuals, and the relative proportions of Newfoundlanders assignable to these five haplogroups do not differ significantly from those expected for typical western European populations (Tables 6 and 7: Richards *et al.*, 2000; Sykes, 2001; Torroni *et al.*, 2006). No Newfoundland individual was found to belong to haplogroup X or V. However, only 6% of modern Europeans belong to haplogroup X and they are largely confined to eastern and central Europe, which is not a source region of the founding population of Newfoundland. Haplogroup V constitutes only 5% of modern Europeans; they are located in Scandinavia and Finland, which contributed little to the founding population of Newfoundland. Haplogroup I, although not included in the seven "Daughters of Eve", is principally a European haplogroup detected at low frequencies across western Eurasia with slightly greater representation in northern

and western Europe (Macaulay *et al.*, 1999). It has been shown to constitute approximately 2% of Finland (Torroni *et al.*, 1996), and individual 10670 was most closely related to a Finnish individual assigned to haplogroup I. This provides further evidence for the limited loss of haplogroup diversity in the Newfoundland population, as a haplogroup that is present in the Finnish population at a low frequency has been preserved despite the fact that it not a representative haplogroup of the majority of the founding population.

4.5 Conclusions and Future Directions

The genetics of small populations and the historical pattern of settlement and demography of Newfoundland have led to expectations of low genetic diversity. It has been stated that founder populations are ideal for studying disease genes as isolation, inbreeding, and founder effects, reduce the genetic complexity of the disorder so that it can be more easily identifiable (Sheffield *et al.*, 1998). However, in this study, this expected loss of haplogroup diversity was not observed; comparisons to the European population did not show a decrease in genetic variation or any statistically significant alteration in

the patterns of diversity. The main haplogroups that are found in Europe are found in Newfoundland at similar frequencies, which suggests that genetic variation may not be reduced as rare haplogroups are preserved. This may be because there appears to be pockets of founding populations across the island of Newfoundland, and individuals used in this study were collected from the St. John's region, not from localized areas. Support for this idea comes from a recent publication that demonstrated that a randomized collection of 200 Newfoundland individuals was comparable to the outbred European population (Service *et al.*, 2006).

In conclusion, analysis of the whole mitochondrial genome provides direct inference of mitochondrial population structure. It yields numerous sites for comparison and eliminates the effects of sampling when investigating select genes from the genome. A whole mitochondrial genome study also provides the opportunity to compare rates and patterns of mitochondrial DNA evolution.

Furthermore, the use of complete mitochondrial DNA sequence data provides accurate information regarding phylogenetic relationships. This study illustrated a limitation in using haplogroup data based on CR sequences - not all haplogroups were found to be monophyletic. The major strength in the use of complete mtDNA data

is the ability to identify monophyletic lineages with increased confidence.

Possible future directions include increased sampling in order to determine if the Newfoundland population contains homogeneous isolates. Once these isolates have been ascertained, further studies can be conducted in order to identify the SNPs associated with the particular genetic condition prevalent in that area.

5.0 Literature Cited

Adcock, G.J., Dennis, E.S., Easteal, S., Huttley, G.A., Jermiin, L.S., Peacock, W.J., & Thorne, A. (2001). Mitochondrial DNA sequences in ancient Australians: implications for modern human origins. *Proceeding of the National Academy of Sciences of the United States of America, 98(2)*: 537-542.

Alfonso-Sanchez M.A., Martinez-Bouzas C., Castro A., Pena J.A., Fernandez-Fernandez I., Herrera R.J., & de Pancorbo M.M. (2006). Sequence polymorphisms of the mtDNA control region in a human isolate: the Georgians from Swanetia. *Journal of Human Genetics*, *51*(5): 429-439. Epub 2006 Apr. 1.

Allard, M.W., Polanskey, D., Wilson, M.R., Monson, M.L., & Budlowle, B. (2006). Evaluation of Variation in Control Region Sequences for Hispanic Individuals in the SWGDAM mtDNA Data Set. *Journal of Forensic Sciences*, *51*(3): 566-573.

Andermann, E., Jacob, J.C., Andermann, F., Carpenter, S., Wolfe, L., & Berkovic, S.F. (1988). The Newfoundland aggregate of neuronal ceroid-lipofuscinosis. *American Journal of Medical Genetics. Supplement, 5*: 111-116.

Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H.L., Coulson, A.R., Diouin, J., Eperson, I.C., Nierlich, D.P., Roe, B.A., Sanger, F., Schreier, P.H.,Smith, A.J.H., Staden, R., & Young, I.G. (1981). Sequence and organization of the human mitochondrial genome. *Nature, 290(5806)*: 457-465.

Andrews, R.M., Kubacka, I., Chinnery, P.F., Lightowlers, R.N., Turnbull, D.M., & Howell, N. (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature Genetics, 23(2)*: 147.

Arnason, U., Gullberg, A., Janke, A., & Xu X. (1996). Pattern and timing of evolutionary divergences among hominoids based on analyses of complete mtDNAs. *Journal of Molecular Evolution, 43(6):* 650-651.

Avise, J.C., Arnold, J., Ball, R.M., Bermingham, E., Lamb, T., Neigel, J.E., Reeb, C.A., & Saunders, N.C. (1987). Intraspecific phylogeography: the molecular bridge between population genetics and systematics. *Annual Review of Ecology and Systematics, 18*: 489-522.

Avise, J.C., Nelson, W.S., & Sibley, C.G. (1994). Why one kilobase sequences from mitochondrial DNA fail to solve the Hoatzin phylogenetic enigma. *Molecular Phylogenetics and Evolution, 3(2)*: 175-184.

Barbujani, G. & Bertorelle, G. (2001). Genetics and the Population history of Europe. *Proceedings of the National Academy of Sciences of the United States of America, 98(1)*: 22-25.

Bear J.C., Nemec, T.F., Kennedy, J.C., Marshall, W.H., Power, A.A., Kolonel, V.M., & Burke, G.B. (1988) Inbreeding in outport Newfoundland. *American Journal of Medical Genetics, 29(3)*: 649-660.

Brandstätter, A., Salas, A., Niederstätter, H., Gassner, C., Carracedo, A., & Parson, W. (2006). Dissection of mitochondrial superhaplogroup H using coding region SNPs. *Electrophoresis, 27(13):* 2541-2550.

Cabot, E.L., & Beckenbach, E.T. (1989). Simultaneous editing of multiple nucleic acid and protein sequences with ESEE. *Computer applications in the biosciences: CABIOS, 5(3)*: 233-234.

Cann, R.L., & Wilson, A.C. (1983). Length mutations in human mitochondrial DNA. *Genetics, 104(4):* 699-711.

Cann, R.L., Stoneking, M., & Wilson, A.C. (1987). Mitochondrial DNA and human evolution. *Nature, 325(6099):* 31-36.

Carr, S.M., Snellen, A.J., Howse, K.A., & Wroblewski, J.S. (1995). Mitochondrial DNA sequence variation and genetic stock structure of Atlantic Cod (*Gadus morhua*) from bay and offshore locations on the Newfoundland continental shelf. *Molecular Ecology, 4*: 79-88.

Carr, S.M., Marshall, H.D., Duggan, S.M., Flynn, S.M.C., Johnstone, K.A., Pope, A.M., & Wilkerson, C.D. (2007). Phylogeographic Genomics of mitochondrial DNA: Highly-resolved patterns of intraspecific evolution and a multi-species, microarray-based DNA sequencing strategy for biodiversity studies. *Comparative Biochemistry and Physiology, Part D: Genomics and Proteomics* (in press)

Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X.C., Stern, D., Winkler, J., Lockhart, D.J., Morris, M.S., & Fodor, S.P. (1996). Accessing genetic information with high-density DNA arrays. *Science, 274(5287): 610-614.*

Finnilä, S., Hassinen, I.E., Ala-Kokko, L., & Majamaa, K. (2000). Phylogenetic network of the mtDNA haplogroup U in Northern Finland based on sequence analysis of the complete coding region by conformation-sensitive gel electrophoresis. *American Journal of Human Genetics, 66(3):* 1017-1026.

Flynn, S.M.C., & Carr, S.M. (2007). (personal communication)

Griffiths, A.J.F., Miller, J.H., Suzuki, D.T., Lewontin, R.C., & Gelbart, W.M. (2000). *An Introduction to Genetic Analysis – 7th ed*. W.H. Freeman: New York.

Hacia, J.G., Sun, B., Hunt, N., Edgemon, K., Mosbrook, D., Robbins, C., Fodor, S.P., Tagle, D.A., & Collins, F.S. (1998). Strategies for mutational analysis of the large multiexon ATM gene using high-density oligonucleotide arrays. *Genome Research; 8(12)*: 1245-58

Hacia, J.G. (1999). Resequencing and mutational analysis using oligonucleotide microarrays. *Nature Genetics*, 21(1 Suppl): 42-47.

Hartl, D.L., & Jones, E. W. (2005). *Genetics: Analysis of Genes and Genomes – 6th ed*. Jones and Bartlett Publishers: Sudbury, Massachusetts.

Hillis, D.M., Moritz, C., & Mable, B.K. (eds). (1996). Molecular Systematics, 2nd edition. Sinauer Associates, Inc.: Sunderland.

Ingman, M., Kaessmann, H., Pääbo, S., & Gyllensten, U. (2000). Mitochondrial genome variation and the origin of modern humans. *Nature*, *408(6813)*: 708-713.

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

Kaessmann, H. & Pääbo, S. (2002). The genetical history of humans and the great apes. *Journal of Internal Medicine, 251(1)*: 1-18.

Lee S.D., Lee Y.S., & Lee J.B. (2002). Polymorphism in the mitochondrial cytochrome B gene in Koreans. An additional marker for individual identification. *International Journal of Legal Medicine, 116*(2): 74-78.

Loogväli, E.L., Roostalu, U., Malyarchuk, B.A., Derenko, M.V., Kivisild, T., Metspalu, E., Tambets, K., Reidla, M., Tolk, H.V., Parik, J., Pennarun, E., Laos, S., Lunkina, A., Golubenko, M., Barac, L., Pericic, M., Balanovsky, O.P., Gusar, V., Khusnutdinova, E.K., Stepanov, V., Puzyrev, V., Rudan, P., Balanovska, E.V., Grechanina, E., Richard, C., Moisan, J.P., Chaventre, A., Anagnou, N.P., Pappa, K.I., Michalodimitrakis, E.N., Claustres, M., Golge, M., Mikerezi, I., Usanga, E., & Villems, R. (2004). Disuniting uniformity: a pied cladistic canvas of mtDNA haplogroup H in Eurasia. *Molecular Biology and Evolution, 21(11)*: 2012-2021.

Maca-Meyer, N., Gonález, A.M., Larruga, J.M., Flores, C., &Cabrera, V.M. (2001). Major genomic mitochondrial lineages delineate early human expansions. *BMC Genetics, 2(1)*: 13.

Macaulay, V., Richards, M., Hickey, E., Vega, E., Cruciani, F., Guida, V., Scozzari, R., Bonne-Tamir, B., Sykes, B., & Torroni, A. (1999). The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. *American Journal of Human Genetics, 64(1):* 232-249.

Maitra, A., Cohen, Y., Gillespie, S.E., Mambo, E., Fukushima, N., Hoque, M.O., Shah, N., Goggins, M., Califano, J., Sidransky, D., & Chakravarti, A. (2004). The Human MitoChip: a high-throughput sequencing microarray for mitochondrial mutation detection. *Genome Research, 14(5):* 812-819.

Mannion, J.J. (ed). (1977). *The peopling of Newfoundland: essays in historical geography*. Institute of Social and Economic Research, Memorial University of Newfoundland, St. John's, Newfoundland.

Marshall, H.D., & Baker, A.J. (1998). Rates and patterns of mitochondrial DNA sequence evolution in Fringilline Finches (*Fringilla spp.*) and the Greenfinch (*Carduelis chloris*). *Molecular Biology and Evolution, 15(6)*: 638-646.

Marshall, H.D. (2003). (personal communication)

Mishmar, D., Ruiz-Pesini, E., Golik, P., Macaulay, V., Clark, A.G., Hosseini, S., Brandon, M., Easley, K., Chen, E., Brown, M.D., Sukernik, R.I., & Olckers, A. (2003). Natural Selection shaped regional mtDNA variation in humans. *Proceedings of the National Academy of Sciences of the United States of America, 100(1)*: 171-176.

Moore, S.J., Green, J.S., Fan, Y., Bhogal, A.K., Dicks, E., Fernandez, B.A., Stefanelli, M., Murphy, C., Cramer, B.C., Dean, J.C., Beales, P.L., Katsanis, N., Bassett, A.S., Davidson, W.S., & Parfrey, P.S. (2005). Clinical and genetic epidemiology of Bardet-Biedl syndrome in Newfoundland: a 22-year prospective, population-based, cohort study. *American Journal of Medical Genetics, 132(4):* 352-360.

Nei, M. (1987). *Molecular Evolutionary Genetics*. New York: Columbia University Press.

Nei, M. & Kumar, S. (2000). *Molecular Evolution and Phylogenetics*. Oxford University Press: New York.

Parfrey, P.S., Davidson, W.S., & Green, J.S. (2002). Clinical and genetic epidemiolgy of inherited renal disease in Newfoundland. *Kidney International, 61(6)*: 1925-1934.

Pope, A.M. (2003). An investigation of the ethnic composition of the Newfoundland population based on whole mitochondrial genomes. B.Sc. (hons.) thesis, Memorial University of Newfoundland, St. John's, Newfoundland and Labrador.

Prowse, D.W. (1972). *A history of Newfoundland from the English, Colonial, and foreign records*. MacMillan and Co., LTD: London and New York.

Reidla, M., Kivisild, T., Metspalu, E., Kaldma, K., Tambets, K., Tolk,, H.V., Parik, J., Loogväli, E.L., Derenko, M., Malyarchuk, B., Bermisheva, M., Shadanov, S., Pennarun, E., Gubina, M., Golubenko, M., Damba,L., Fedorova, S., Gusar, V., Grechanina, E., Mikerezi, I., Moisan, J.P., Chaventre, A., Khusnutdinova, D., Osipova, L., Stepanov, V., Voevoda, M., Achilli, A., Rengo, C., Rickards, O., De Stefano, G.F., Papiha, S., Beckman, L., Janicijevic, B., Rudan P., Anagnou, N., Michalodimitrakis, E., Koziel, S., Usanga, E., Geberhiwot, T., Hernstadt, C., Howell, N., Torroni, A., & Villems, R. (2003). Origin and diffusion of mtDNA haplogroup X. *American Journal of Human Genetics, 73(5):* 1178-1190.

Richards, M.B., Macaulay, V.A., Bandelt, H.J, & Sykes, B.C. (1998). Phylogeography of mitochondrial DNA in western Europe. *Annals of Human Genetics, 62 (Pt 3):* 241-260.

Richards, M., Macaulay, V., Hickey, E., Vega, E., Sykes, B., Guida, V., Renga, C., Sellito, D., Cruciani, F., Kivisild, T., Villems, R., Thomas, M., Rychkov, O., Rychkov, Y., Gölge, M., Dimitrov, D., Hill, E., Bradley, D., Romano, V., Calì, F., Vona, G., Demaine, A, Papiha, S., Triantaphyllidid, C., Stefanescu, G., Hatina, J., Velledi, M., Di Rienzo, A., Novelletto, A., Oppenheim, A., Nørby, S., Al-Zaheri, N., Samntachiara-Benerecetti, S., Scozzari, R., Torroni, A., & Bandelt, H-J. (2000). Tracing European Founder lineages in the near eastern mtDNA pool. *The American Journal of Human Genetics, 67(5):* 1251-1276.

Rieder, M.J., Taylor, S.L., Tobe, V.O., & Nickerson, D.A. (1998). Automating the identification of DNA variations using quality-based fluorescence re-sequencing: analysis of the human mitochondrial genome. *Nucleic Acids Research, 26(4)*: 967-973.

Roff, D.A., & Bentzen, P. (1989). The statistical analysis of mitochondrial DNA polymorphisms: chi 2 and the problem of small samples. *Molecular Biology and Evolution, 6(5)*: 539-545.

Roostalu, U., Kutuev, I., Loogväli, E.L., Metspalu, E., Tambets, K., Reidla, M., Khusnutdinova, E., Usanga, E., Kivisild, T., & Villems, R. (2007). Origin and expansion of haplogroup h, the dominant human mitochondrial DNA lineage in west eurasia: the near eastern and caucasian perspective. *Molecular Biology and Evolution, 24(2):* 436-448.

Rowe, F.W. (1980). *A history of Newfoundland and Labrador*. McGraw-Hill Ryerson: Toronto.

Russell, P.J. (1998). Genetics, 5th edition. Benjamin Cummings: California.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution, 4(4):* 406-425.

Sanger F., Nicklen S., & Coulson A.R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America, 74(12)*: 5463-5467.

Service, S.K., Temple Lang, D.W., Freimer, N.B., & Sandkuijl, L.A. (1999). Linkage-disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. *American Journal of Human Genetics, 64(6)*: 1728-1738.

Service, S., DeYoung, J., Karayiorgou, M., Roos, J.L., Pretorious, H., Bedoya, G., Ospina, J., Ruiz-Linares, A., Macedo, A., Palha, J.A., Heutink, P., Aulchenko, Y., Oostra, B., van Duijn, C., Jarvelin, M.R., Varilo, T., Peddle, L., Rahman, P., Piras, G., Monne, M., Murray, S., Galver, L., Peltonen, L., Sabatti, C., Collins, A., & Freimer, N. (2006). Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. *Nature Genetics, 38(5)*: 556-560.

Sheffield, V.C., Stone, E.M., & Carmi, R. (1998). Use of isolated inbred human populations for identification of disease genes. *Trends in Genetics, 14(10)*: 391-396.

Sigurgardottir S., Helgason A., Gulcher J.R., Stefansson K., & Donnelly P. (2000). The mutation rate in the human mtDNA control region. *American Journal of Human Genetics, 66*(5): 1599-1609.

Silva, W.A., Bonatto, S.L., Holanda, A.J., Ribeiro-dos-Santos, A.K., Paixão, B.M., Goldman, G.H., Abe-Sandes, A., Rodriguez-Delfin, L., Barbose, M., Paçó-Larson, M.L., Petzl-Erler, M.L., Valente, V., Santos, S.E.B., & Zago, M.A. (2002). Mitochondrial genome diversity of Native Americans supports a single early entry of founder populations into America. *American Journal of Human Genetics, 71(1)*: 187-192.

Starikovskaya, Y.B., Sukernik, R.I., Schurr, T.G., Kogelnik, A.M., & Wallace, D.C. (1998). MtDNA diversity in Chukchi and Siberian Eskimos: implications for the genetic history of Ancient Beringia and the peopling of the New World. *American Journal of Human Genetics, 63(5):* 1473-1491.

Starikovskaya, E.B., Sukernik, R.I., Derbeneva, O.A.,Volodko, N.V., Ruiz-Pesini, E., Torroni, A., Brown, M.D., Lott, M.T., Hosseini, S.H., Huoponen, K, & Wallace, D.C. (2005). Mitochondrial DNA diversity in Indigenous Populations of the Southern Extent of Siberia, and the origins of Native American Haplogroups. *Annals of Human Genetics, 69 (Pt 1):* 67-89.

Story, G.M., Kirwin, W.J., & Widdowson, J.D.A. (1999). Dictionary of Newfoundland English, 2nd edition. University of Toronto Press: Toronto

Swofford D. (2002). PAUP*: phylogenetic analysis using parsimony (and other methods) v. 4.0 Beta. Florida State University.

Sykes, B.C. (2001). Daughters of Eve. W.W. Norton, New York: New York

Tamarin, R.H. (2002). Principles of Genetics, 7th edition. McGraw-Hill

Torroni, A., Lott, M.T., Cabell, M.F., Chen, Y.S.,Lavergne, L., & Wallace, D.C. (1994). mtDNA and the origin of Caucasians: identification of ancient Caucasian-specific haplogroups, one of which is prone to a recurrent somatic duplication in the D-loop region. *American Journal of Human Genetics, 55(4):* 760-776.

Torroni, A., & Wallace, D.C. (1995). MtDNA haplogroups in Native Americans. *American Journal of Human Genetics, 56(5)*: 1234-1238.

Torroni, A., Huoponen, K., Francalacci, P., Petrozzi, M., Morelli, L., Scozzari, R., Obinu, D., Savontaus, M.L., & Wallace, D.C. (1996). Classification of European mtDNAs from an analysis of three European populations. *Genetics, 144(4):* 1835-1850.

Torroni, A., Achilli, A., Macaulay, V., Richards, M., & Bandelt, H.J. (2006). Harvesting the fruit of the human mtDNA tree. *Trends in Genetics, 22(6):* 339-345.

Wallace, D.C. (1994). Mitochondrial DNA sequence variation in human evolution and disease. *Proceedings of the National Academy of Sciences of the United States of America, 91(19):* 8739-8746.

Woods, M.O., Young, T-L., Parfrey, P.S., Hefferton, D., Green, J.S., & Davidson, W.S. (1999). Genetic heterogeneity of Bardet-Biedl syndrome in a distinct Canadian population: evidence for a fifth locus. *Genomics, 55(1)*: 2-9.

Woods, M.O., Hyde, A.J., Curtis, F.K., Stuckless, S., Green, J.S., Pollett, A.F., Robb, J.D., Green, R.C., Croitoru, M.E., Careen, A., Chaulk, J.A., Jegathesan, J., McLaughlin, J.R., Gallinger, S.S., Younghusband, H.B., Bapat, B.V., & Parfrey, P.S. (2005). High frequency of hereditary colorectal cancer in Newfoundland likely involves novel susceptibility genes. *Clinical Cancer Research: an official journal of the American Association for Cancer Research, 11(19 Pt 1):* 6853-6861.

Young, T-L., Woods, M.O., Parfrey, P.S., Green, J.S., Hefferton, D., & Davidson, W.S. (1999). A founder effect in the Newfoundland population reduces the Bardet-Biedl syndrome 1 (BBS1) interval to 1 cM. *The American Society of Human Genetics, 65(6)*: 1680-1687.

Appendix A A portion of the questionnaire regarding maternal ancestry relevant to the mitochondrial sampling.

A Study to Determine The Ethnic Composition of The Newfoundland Population

Thank you for taking part in our study on the origins of the Newfoundland people.
Please complete as many the questions as you can.
If you have any questions call 777-7286 (collect).

Name: _____

The next 5 questions are about your **mother's mother**.

1. Her name before she was married:_____
2. Where did she live when she was a child? _____
3. What is/was her religion? _____
4. Is she of (tick one): 9 English, 9 Irish, 9 French or 9some other ancestry (ethnic background)?
5. Where did **her** mother come from?

   _____
6. Was any member of your mother's or grandmother's family adopted?

   _____
7. Was any member of your mother's or grandmother's family raised by someone other than their own parents?

   _____

Thank you for your help.

Please place this questionnaire along with the signed consent form in the enclosed envelope and send it back to us. The envelope has our address and a stamp on it.

Ban Younghusband
Discipline of Genetics
Memorial University of Newfoundland
St John's NF A1B 3V6
Telephone 709-777-7286

Appendix B Sequences of the 24 primer pairs used to amplify the whole mtDNA genome in overlapping regions, as described by Rieder *et al.*, (1998).

Appendix B: Sequences of the 24 primer pairs used to amplify the whole human mtDNA genome in overlapping regions, as described by Rieder *et al.* (1998).

Primer region	Primer sequence (5′-3′)	Length of PCR product (bp)	Overlap with preceding region (bp)
h01	F: CTCCTCAAAGCAATACACTG	840	202
	R: TGCTAAATCCACCTTCGACC		
h02	F: CGATCAACCTCACCACCTCT	802	204
	R: TGGACAACCAGCTATCACCA		
h03	F: GACTAACCCCTATACCTTCTGC	860	196
	R: GGCAGGTCAATTTCACTGGT		
h04	F: AAATCTTACCCCGCCTGTTT	887	208
	R: AGGAATGCCATTGCGATTAG		
h05	F: TACTTCACAAAGCGCCTTCC	832	215
	R: ATGAAGAATAGGGCGAGGG		
h06	F: TGGCTCCTTTAACCTCTCCA	898	203
	R: AAGGATTATGGATGCGGTTG		
h07	F: ACTAATTAATCCCCTGGCCC	975	207
	R: CCTGGGGTGGGTTTTGTATG		
h08	F: CTAACCGGCTTTTTGCCC	814	201
	R: ACCTAGAAGGTTGCCTGGCT		
h09	F: GAGGCCTAACCCCTGTCTTT	827	214
	R: ATTCCGAAGCCTGGTAGGAT		
h10	F: CTCTTCGTCTGATCCGTCCT	886	211
	R: AGCGAAGGCTTCTCAAATCA		
h11	F: ACGCCAAAATCCATTTCACT	987	205
	R: CGGGAATTGCATCTGTTTTT		
h12	F: ACGAGTACACCGACTACGGC	900	196
	R: TGGGTGGTTGGTGTAAATGA		

Appendix B (continued)

h13	F: TTTCCCCCTCTATTGATCCC	816	214
	R: GTGGCCTTGGTATGTCCTTT		
h14	F: CCCACCAATCACATGCCTAT	940	205
	R: TGTAGCCGTTGAGTTGTGGT		
h15	F: TCTCCATCTATTGATGAGGGTCT	891	182
	R: AATTAGGCTGTGGGTGGTTG		
h16	F: GCCATACTAGTCTTTGCCGC	840	203
	R: TTGAGAATGAGTGTGAGGCG		
h17	F: TCACTCTCACTGCCCAAGAA	802	196
	R: GGAGAATGGGGGATAGGTGT		
h18	F: TATCACTCTCCTACTTACAG	866	166
	R: AGAAGGTTATAATTCCTACG		
h19	F: AAACAACCCAGCTCTCCCTAA	977	242
	R: TCGATGATGTGGTCTTTGGA		
h20	F: ACATCTGTACCCACGCCTTC	970	207
	R: AAGGGGTCAGGGTTCATTC		
h21	F: GCATAATTAAACTTTACTTC	938	206
	R: AGAATATTGAGGCGCCATTG		
h22	F: TGAAACTTCGGCTCACTCCT	1162	180
	R: AGCTTTGGGTGCTAATGGTG		
h23	F: TCATTGGACAAGTAGCATCC	765	205
	R: GAGTGGTTAATAGGGTGATAG		
h24	F: CACCATTCTCCGTGAAATCA	954	203
	R: AGGCTAAGCGTTTTGAGCTG		

Appendix C Sequences of the 14 primer pairs used to amplify the whole mtDNA genome in overlapping regions, as modified from Rieder *et al.*, (1998).

Appendix C: Sequences of the 14 primer pairs used to amplify the complete mitochondrial genome in overlapping regions, as modified from Rieder *et al.*, (1998).

Primer region	Primer sequence (5′-3′)	Length of PCR product (bp)	Overlap with preceding region (bp)
h01-h02	F: CTCCTCAAAGCAATACACTG R: TGGACAACCAGCTATCACCA	1642	202
h03-h04	F: GACTAACCCCTATACCTTCTGC R: AGGAATGCCATTGCGATTAG	1747	200
h05-h06	F: TACTTCACAAAGCGCCTTCC R: AAGGATTATGGATGCGGTTG	1730	212
h07	F: ACTAATTAATCCCCTGGCCC R: CCTGGGGTGGGTTTTGTATG	975	207
h08	F: CTAACCGGCTTTTTGCCC R: ACCTAGAAGGTTGCCTGGCT	814	201
h09-h10	F: GAGGCCTAACCCCTGTCTTT R: AGCGAAGGCTTCTCAAATCA	1713	207
h11-h12	F: ACGCCAAAATCCATTTCACT R: TGGGTGGTTGGTGTAAATGA	1887	208
h13-h14	F: TTTCCCCCTCTATTGATCCC R: TGTAGCCGTTGAGTTGTGGT	1756	205
h15-h16	F: TCTCCATCTATTGATGAGGGTCT R: TTGAGAATGAGTGTGAGGCG	1730	193
h17-h18	F: TCACTCTCACTGCCCAAGAA R: AGAAGGTTATAATTCCTACG	1688	200
h19	F: AAACAACCCAGCTCTCCCTAA R: TCGATGATGTGGTCTTTGGA	977	242
h20	F: ACATCTGTACCCACGCCTTC R: AAGGGGTCAGGGTTCATTC	970	207

Appendix C (continued)

h21-h22	F: GCATAATTAAACTTTACTTC R: AGCTTTGGGTGCTAATGGTG	2100	206
h23-h24	F: TCATTGGACAAGTAGCATCC R: AGGCTAAGCGTTTTGAGCTG	1719	192

Appendix D A discussion of the patterns of molecular evolution evident in this study.

Appendix D

The rate of nucleotide substitution was much higher at the third codon position as compared to the first and second positions. This is expected, as substitutions at the third position typically result in silent mutations. There are, however, some substitutions at the third position that result in amino acid changes, as well as, substitutions at the first position in Leucine codons that result in silent mutations. It is therefore necessary to determine the rate of synonymous and nonsynonymous substitutions separately. According to a strict neutral theory of molecular evolution, the rates of synonymous and nonsynonymous substitution should be equal, but this has not generally been observed. Rather, the rate of synonymous substitution has been shown to be equal to that of neutral nucleotide substitution and to be similar amongst genes (Nei & Kumar, 2000). By comparison, the rate of nonsynonymous substitution is generally much lower and varies substantially among genes (Nei & Kumar, 2000). Kimura (1983) demonstrated that the observed variation was due to selection (cited in Nei & Kumar, 2000). There are, however, genes that demonstrate higher rates of nonsynonymous substitutions (COXI, COXII, ATP6, and COXIII in the present study). In this case, previous investigators have

suggested that the mutations are advantageous and are therefore maintained in the population by natural selection.

It is known that transitions occur more frequently than transversions. Transitions occur when alternative purines or alternative pyrimidines are substituted into the sequence. A transversion refers to the replacement of a purine by a pyrimidine or vice versa. Among closely related individuals or species, observed transitions routinely outnumber transversions by at least 2:1; this is termed the transition to transversion ratio. In the current investigation, the ratio was found to be approximately 11:1. A high transition to transversion rate indicates a population of recent origin, such as the Western European human population. This pattern is consistent with findings in previous mitochondrial DNA studies (e.g. Carr *et al.*, 1995; Marshall and Baker, 1998).