HÁSKÓLI ÍSLANDS

Faculty of Social Science

MA-thesis

Anthropology

The origin of Icelandic mtDNA lineages from haplogroup C

Sigríður Sunna Ebenesersdóttir

February 2010

Leiðbeinandi: Dr. Agnar Helgason

Nemandi: Sigríður Sunna Ebenesersdóttir

Kennitala: 221182-3759

ABSTRACT

Icelanders are one of the most studied populations in human genetics. The deCODE Genetics genealogical database allows unusually detailed conclusions to be drawn about the population history of the Icelanders. In this study, genetic and genealogical data were combined to make inferences about a rare mtDNA haplotype found in Iceland that belongs to haplogroup C1. While most of the Icelandic mtDNA pool originated in the British Isles or Scandinavia, this haplotype is likely to have a more distant origin, either from East Asians or Native Americans.

The goal of the research described in this thesis was twofold. First, we wanted to know its time of entry onto the Icelandic mtDNA pool. In all, 744 individuals were sequenced for their mtDNA control region in this study in addition to the 1538 mtDNA control region sequences available beforehand. Of these 2,266 individuals, seven were haplogroup C1 carriers yielding a population frequency of about 0.3%. The seven Icelandic haplogroup C1 carriers could be traced back to four matrilineal ancestors, born between 1710 and 1740 AD from a limited geographic region. We conclude that a single founding female ancestor gave rise to the Icelandic haplogroup C lineage, who must have arrived sometime between the starting date of the settlement of Iceland (874 AD) and 1690 AD.

The second goal was to trace the geographical origin of the Icelandic C1 haplotype outside Iceland. To this end, the complete mtDNA genome was sequenced from eleven Icelandic C1 carriers, allowing it to be placed within the phylogeny of haplogroup C1. Phylogenetic trees were constructed from all 61 complete haplogroup C1 sequences published to date, along with the Icelandic C1 sequence. Interestingly, this analysis revealed that the Icelandic sequence belongs to a new a very distinct sub-clade in C1, named C1e. This leaves us with an even greater mystery than before. The Icelandic C1e haplotype could either be of Native American, European or Asian origin. Further research is necessary to determine which of these regions gave rise to C1e, but at present the strongest evidence points to a Native American origin.

PREFACE

This thesis outlines the results of a research-based Master's project in biological anthropology conducted at the University of Iceland and deCODE Genetics in the years 2007-2009. The study was supervised by Dr. Agnar Helgason (deCODE Genetics and the University of Iceland). The project and the thesis accounted for 90 ECTS of 120 ECTS units required to obtain an MA degree in Anthropology form the University of Iceland.

ACKNOWLEDGMENTS

Firstly, I would like to thank my supervisor, Dr. Agnar Helgason, for his guidance, support and patience during the past two and a half years. I feel truly privileged to have worked with him. Many thanks go to Hjalti Sigurðarson for helping me with Photoshop work. Special thanks to Ásgeir Sigurðsson for his help in cloning PCR amplifications, and the advice he has given me during my lab work. And finally, my greatest thanks go to my love Tóti and my daughter Ísold, whom I could not have done this project without.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABELS

1 INTRODUCTION

Historical sources indicate that Iceland was settled by Vikings just before 870 AD, prior to that Iceland had no human inhabitants. Most Icelandic mitochondrial DNA haplotypes originated from the British Isles or Scandinavia, but at least one haplotype, belonging to haplogroup C, appears to have a more distant origin (Helgason *et al.*, 2000a). MtDNA sequences from haplogroup C has been found among populations in East Asia and among individuals with Native American ancestry.

In all, five Icelandic haplogroup C carriers were known prior to this study. According to the deCODE Genetics genealogy database, the matrilineal ancestry of these Icelandic haplogroup C carriers can be traced back at least eight generations within Iceland, to at least the year 1710. This shows that the Icelandic haplogroup C lineage is not the result of recent gene flow from Native American or Asian populations. It is intriguing to reflect upon what kind of interaction between Icelanders and the external world led to a woman, from a distant geographical region, to settle in Iceland prior to 1710.

If we take into account factors like population size, genetic drift and low gene flow into the Icelandic gene pool, it can be assumed that most mtDNA lineages observed in contemporary Icelanders are descended from the original set of mtDNA lineages present in female settlers 1100 years ago. Around the same time as Iceland was settled, Norse settlers founded a colony in Greenland. For a few centuries their colony thrived. Existing research revealed no evidence of admixture between Greenland Inuits and Icelanders (Helgason *et al.*, 2001; Helgason *et al.*, 2006). However, haplogroup C is absent among Eskimo-Aleut speakers (Aleut Islanders, Eskimos and Inuits) and therefore this does not account for the Icelandic haplogroup C sequences (Schurr, 2004). A preliminary phylogeographical analysis by Agnar Helgason (results not published) indicated that the Icelandic C haplotype is most closely related to Native American

haplogroup C sequences. In light of this, combined with that fact that there is firm archaeological evidence for a Viking settlement in the New World around the 10th century, it was suspected the Icelandic C haplotype could be of Native American origin. Hence, this study could provide insights into the history of pre-Colombian contact between Europeans and Americans. Thus, if it can be verified that the Icelandic haplogroup C sequence was present in the Icelandic mtDNA gene pool before 1500 AD and that it has a Native American origin, then that would indicate contact between Icelanders and Native Americans prior to Columbus's re-discovery of the New World in 1492 AD.

The aim of this study is to use genealogical and genetic data to trace the geographical origin and phylogenetic context of the Icelandic haplogroup C lineage. More specifically the aim is to determine both the geographical origin of the Icelandic haplogroup C lineage and its time of entry into the Icelandic mtDNA pool. The deCODE Genetics genealogical database was used to find additional carriers and to identify the oldest verifiable matrilineal ancestor that carried this rare haplotype in Iceland. Furthermore, the complete mtDNA genome was sequenced in a subset of Icelandic C carriers in order to clarify its phylogenetic relationship to haplogroup C sequences from other populations and to determine whether any mutational diversity exists within the Icelandic lineage. It is important to bear in mind that although inferences are only made about the genetic history of a few Icelanders in this study the results could provide information about important events in the history of Icelanders as a whole.

The main body of this thesis is organized into five chapters of which this introduction is Chapter 1. Chapter 2 introduces some of the basic genetic concepts necessary for the reading of this thesis. It presents background of state of knowledge about human mtDNA, phylogeography and evolution. Chapter 3 gives the reader a background to the main historical events relevant to this study, for example, the colonization of America, the settlement of Iceland and the Viking discovery of America. Furthermore, it discusses the Icelandic and Native American populations as subjects of population genetic research, with a focus on mtDNA. Chapter 4 explains the methods applied in the analysis of genealogical and genetic data. Also it gives information about additional data collected and used in the analysis performed in this study. Chapter 5 presents the results of this study. In this chapter, the two main points of interest are addressed, the founding age of the

Icelandic haplogroup C1 lineage and the phylogenetic context of the Icelandic C1 sequences. Finally, in Chapter 6 the main conclusions are summarized and interpretation is provided in relation to the main question of this study i.e. the phylogeographic origin of the Icelandic mtDNA lineage from haplogroup C.

2 BACKGROUND

Human genetic research can be used to shed light on our past. In particular, it reveals the dispersals of the modern *Homo sapiens* throughout the world. An effective way to do this is to study the non-recombining regions of the genome such as the mitochondrial DNA and the Y-chromosome. The Y-chromosome is paternally inherited and mtDNA is maternally inherited. Thus, the Y-chromosome carries information about the evolutionary past of males, whereas mitochondrial DNA carries information about the evolutionary history of women. By analysing the sequence variation in these two genomes, genetic lineages that are present within populations can be identified. Analyses of these regions have yielded new insights into the evolutionary history, migration and the historical relationship of human populations. That is why the Y-chromosome and mitochondrial genome have become standard tools for genetic analysis of human populations.

2.1 General structure and inheritance of mitochondrial DNA

2.1.1 The human genome

DNA is a molecule that carries the genetic code of all living things. The genetic code specifies the sequences of amino acids that are the building blocks of proteins, which participate in virtually every process within cells. It is therefore said that DNA holds the key to the structure and function of the whole body. The basic structure of DNA consists of a pair of long nucleotide chains wound around each other, forming a double helix. Each nucleotide consists of a phosphate group, a deoxyribose sugar molecule and one of the four bases: Adenine, Guanine, Cytosine or Thymine. These bases are often represented by the four letters A, G, C and T, it can be said that they form the genetic "alphabet" of DNA. It is the order of these letters in genes that specifies the genetic instructions carried out in each cell. The nucleotide chains or DNA sequences are divided into long strands called

chromosomes that are found within the nucleus of cells, with the exception of a small circular DNA molecule, found in cellular organelles called mitochondria (Jobling *et al.*, 2004; Relethford, 2005; Strachan & Read, 2004).

The human genome consists of 23 chromosome pairs. Twenty-two of these are autosomal chromosome pairs, while the remaining pair is sex-determining. For this pair, women have two X chromosomes, while men have one X and one Y-chromosome. Individuals get half of their genetic material from their mother and half from their father when the ovum and sperm germ cells merge to form a zygote (fertilized egg). During the formation of gametes recombination takes place between autosomal chromosome pairs. Each germ cell produced by an individual contains 23 chromosomes (one from each pair), each being a mosaic of the chromosomes that were inherited from the individual's parents. Generation after generation autosomes undergo this kind of reshuffling through recombination. Recombination therefore plays an important role in the preservation and production of genetic diversity from one generation to the next. After many generations of recombination the chromosome of an individual can be thought of as an intricate mosaic of small DNA segments he has inherited from his different ancestors through time. There are, however, two special parts of our genome that are inherited from one parent only. These are the mitochondrial DNA and most of the Y chromosome. The sperm does not contribute mitochondria to the zygote and the ovum never contains a Y-chromosome. Consequently, the mitochondrial genome and Y chromosome do not undergo recombination of paternal and maternal segments during the formation of sex cells like the nuclear genome does (Jobling et al., 2004; Strachan & Read, 2004).

Mutations are another important mechanism that affects genetic diversity. A mutation is defined as a change in DNA sequence, most often as a result of DNA damage or copying errors that are introduced when DNA is replicated. Mutations include a broad spectrum, ranging from insertions, duplications or inversions of one or few bases to changes in numbers of chromosomes. Among the most common kind of mutation is the substitution of a single base. These mutations can be categorized into groups, transitions, where a purine is replaced by a purine or a pyramidine is replaced by a pyramidine (A↔G or C↔T) and transversions, where a pyramidine is replaced by a purine or a purine by a pyramidine (C or T ↔ A or G). Transitions are more common than transversions. Since mutations are random,

any base in the DNA can change, coding or non-coding the same (Griffiths, 2008; Jobling et al., 2004).

2.1.2 The mitochondrial genome

Mitochondrial DNA is a circular double-stranded molecule located outside the nucleus in the mitochondria of cells. The entire sequence of human mitochondrial DNA, i.e. its genome, was established 28 years ago (Anderson *et al.*, 1981). The number of mitochondria in a cell varies with cell type, from a few hundred to thousands. Each mitochondria contains one or more identical molecule of DNA, typically 16,569 base pairs long and containing 37 genes (Jobling et al., 2004). In comparison, the nuclear genome consists of approximately 3 billion base pairs and estimated 20,000-25,000 genes. Thus, mitochondrial DNA or mtDNA, as it is often called, is only a tiny fraction of our total genome. The mitochondrial genome, unlike the much larger nuclear genome, is transmitted via the mitochondria which are in the oocyte cytoplasm. It is maternally inherited, meaning that both males and females inherit their mtDNA from their mother, but males do not transmit their mitochondria to next generation. This kind of inheritance is called "maternal". Nuclear DNA recombines after being inherited from both parents so it is difficult to predict its origins, especially after many generations. Mitochondrial DNA is on the other hand transmitted as a non-recombining unit through the mother and therefore it is possible to assess the relationship of two or more individuals through the direct female line (Jobling et al., 2004; Strachan & Read, 2004).

Unlike the nuclear genome, mtDNA is extremely compact as more than 90% of the DNA sequence codes for genes. The remaining mtDNA sequence is called the control region or D-loop. It is the only major non-coding area in human mtDNA, meaning that it does not contain instructions for making proteins or RNA molecules as the coding region does. The control region contains the two most polymorphic regions, known as hypervariable segments 1 and 2 (HVS1 and HVS2) (see Figure 1). The mutation rate of the coding region is ten times greater than that of the nuclear genome and it is even higher in the control region or a hundred times that of the nuclear genome (Jobling et al., 2004).

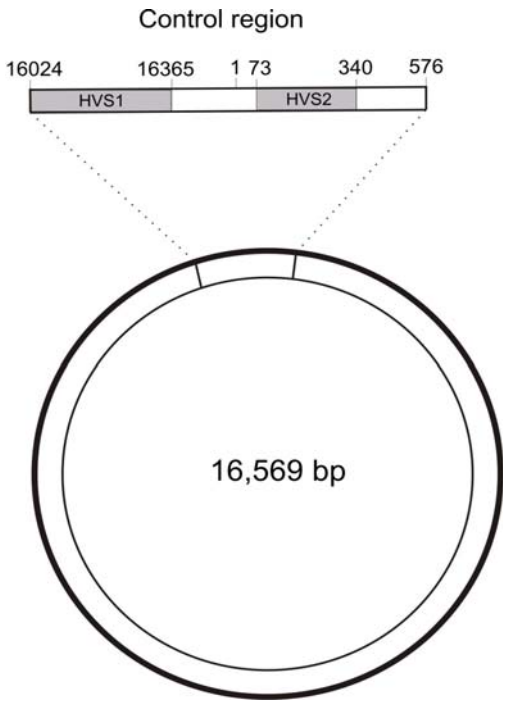

Figure 1. Mitochondrial DNA.

The outer circle represents the heavy strand which is relatively rich in G bases, and the inner circle represents the light strand which is rich in Cs. The control region contains two hypervariable segments (HVS1 and HVS2).

Most studies of human evolution based on mtDNA sequences have focused on the first and second hypervariable segments (HVS1 and HVS2) of the control region, which comprises less than 7% of the mitochondrial genome. In some studies, restriction site analysis of the whole genome is performed as well. To date, sequencing of the first segment hypervariable (HVS1) has proven the most popular method for assessing mtDNA. The sequence length of HVS1 varies in different studies. For example, some have defined the HVS1 range as 16024-16365, some as 16099-16383, while most adhere to the range of 16024-16383. The sites of the mitochondrial DNA are numbered from 1 to 16,569, each number representing a nucleotide. Because mtDNA is a circular molecule sites 16,569 and 1 are side by side. The numeration of the mtDNA circle starts just before HVS2 starts and ends just after HVS1 ends (see Figure 1) (Strachan & Read, 2004).

Within the field of mtDNA studies, mitochondrial sequences are often presented in the form of motifs used to denote sequences based on nucleotide variability at sites where they differ from the so-called revised Cambridge reference sequence (rCRS) (Andrews *et al.*, 1999). In such motifs, difference from rCRS is represented by the site and the nucleotide which is different. For example,

if a sequenced mtDNA genome has a T base at site 16223, but the rCRS has a C base at that site, and that is the only difference between the two of them, then the motif for the sequenced mtDNA is 16223T (Ingman & Gyllensten, 2001; Jobling et al., 2004). Most mutations happen at different sites. However there are some sites that have been hit multiple times and some more often than others. These sites are so-called mutational "hotspots". The drawback of relying solely on control region sequences is that it experiences a relatively high level of such recurrent mutation. The variation found by sequencing just HVS1 is often insufficient to tell whether two mutations are identical by descent or just identical by state. Two nucleotides with the same state and same site are identical by descent if they arose from the same mutation. In contrast, they are merely identical by state if they arose from two different mutation events. Recent advances in sequencing technique have made it possible to study complete mitochondrial genomes, redressing this problem. When the whole mtDNA genome is sequenced the data is usually sufficient to allow researchers to distinguish between identity by descent and identity by state for most mutations. Complete sequencing has therefore greatly improved the power of population genetic mtDNA studies as it gives more precision for estimating the timing and direction of human dispersals (Jobling et al., 2004; Torroni *et al.*, 2006). To date, there are more than 5100 complete mtDNA sequences published, so the structure of mtDNA variation has vastly improved over the last decade (Pereira et al., 2009, see also Mitomap, a human mitochondrial genome database at http://www.mitomap.org/ and http://phylotree.org/).

2.2 The genealogy and geographical dispersal of mtDNA

2.2.1 The mtDNA phylogeny

The fact that human mtDNA is maternally inherited, lacks recombination, has a high mutation rate and a high copy-number per cell makes it a tremendously important tool in evolutionary genetics. The mtDNA sequences can be used to assess the genealogical relationships of two or more individuals through the direct female line and the time to the most recent common ancestor (TMRCA) in the direct female line. This is done by tracing a line from every individual to his/her mother, then continue to trace the lineages of each of these mothers to their

mothers and so on until we reach the most recent common ancestor (MRCA). All human mtDNA sequences are related in this way and their relationships to one another are best understood as a genealogical tree. Genealogical trees representing the relationship between mtDNA sequences are called phylogenetic trees. By using specific algorithms, different kinds of phylogenetic trees can be generated from mtDNA sequences. The principle of maximum parsimony is important in the construction of phylogenetic trees. It states that the most probable tree is the one based on the fewest nucleotide changes. The same principle applies for phylogenetic networks, which are merely a different means of depicting the relationships between mtDNA sequences. A phylogenetic tree appears as a series of points connected by lines that together form branches. The tree is structured from a set of individuals that we want to study, they are represented as points where the lines start. Other points in the tree represent ancestral mtDNA sequences and the lines represent the mutation that separates the ancestral sequences from each other. It is said that individuals share the same haplotype if their mitochondrial DNA sequence being examined is identical. Haplotypes that share a specific set of polymorphism (genetic markers) reflecting their relatedness form clades of lineages that are sometimes referred to as haplogroups. Haplogroups form 'branches' or 'clades' on the mtDNA phylogenetic tree. All haplogroups form clades, however not all clades are haplogroups. The definition of haplogroups is subjective. Branches considered to be "interesting" are named and defined as haplogroups (Jobling et al., 2004; Schurr, 2004; Stein & Rowe, 2003).

Haplogroups are typically defined by specific combinations of mutations. Through time mutations accumulate along less and less related lineages or haplotypes. It is possible to estimate the divergence between those lineages based on the number of mutational differences in the phylogeny. Because these mutations happen at a probabilistically regular rate they can be used like a "clock" to estimate the TMRCA and thereby to estimate the times of human population splits. As we trace back the ancestry of a set of haplotypes, by moving backwards in time, generation after generation, we will at some point encounter the MRCA of the set. All modern humans descend from ancestral individuals, but not all people who lived in the past had descendants. Therefore, different lineages can be connected to different MRCAs. For example the members of haplogroups M*, C and D are descended from a single female ancestor, who is the MRCA of macro-haplogroup

M (see Figure 2, marked as a green circle). This kind of merging of lineages as we go backwards in time is called “coalescence”. As we go further back in time the matrilineal lineages of all humans coalesce to a woman who lived in Africa some 190 thousand years ago, so-called “mitochondrial Eve”. She is the MRCA of all people living today via the mtDNA pathway (Barton, 2007; Jobling et al., 2004; Stanyon *et al.*, 2009; Strachan & Read, 2004). This woman had a particular mtDNA type, which is the single most ancestral point in the mtDNA phylogeny (see Figure 2). This does not mean that the population 190 thousand years ago comprised of only one woman. This merely means that the mitochondria types of all other women who lived at the same time are now extinct. The time of 190 thousand years thus encompasses the history of mutations leading to the mtDNA types found in present human populations. After tens of thousands of years of mutations, many different mtDNA types have originated from the ancestral type (Cavalli-Sforza *et al.*, 1994).

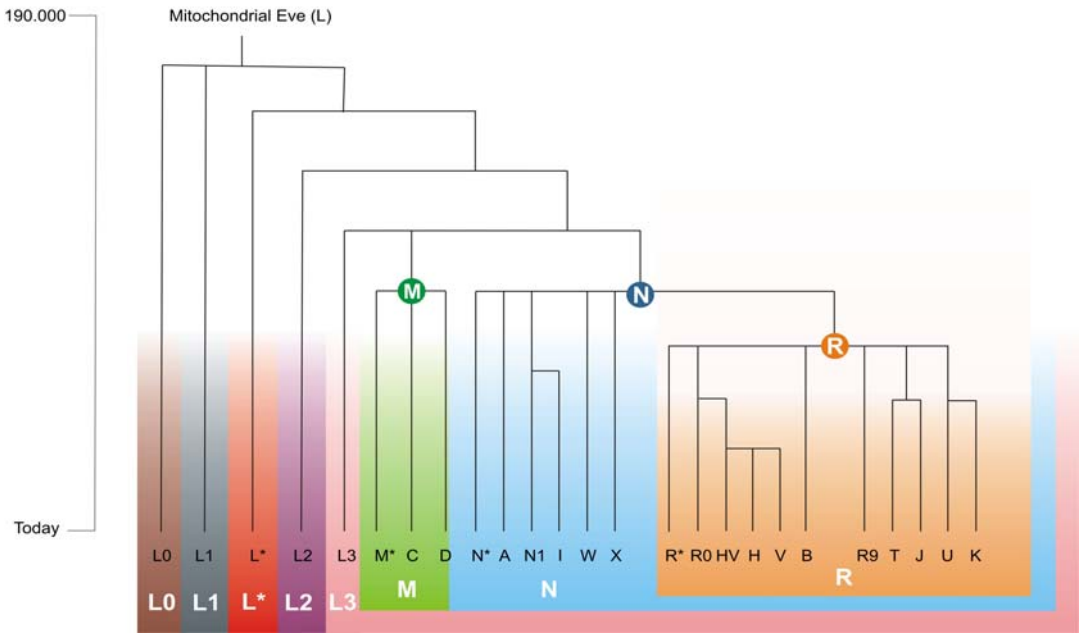

Figure 2. The phylogeny of mtDNA haplogroups

This tree shows the phylogeny of mtDNA haplogroups (modified from Behar *et al.*, 2007). Note that the mtDNA haplogroups are not related in a alphabetic order. The Native American haplogroups A, B, C and D were the first to be named, however that does not mean they are all closely related. The circles indicate the macro-haplogroups. The haplogroups highlighted by a *suffix represent female lines that are paraphyletic. That means, for example, that members of M* represent the collection of all female lineages that descent from super haplogroup M except those leading to haplogroups C and D (Jobling et al., 2004).

Our understanding of the mtDNA phylogeny has vastly improved over the last few years because of the large numbers of samples that have been tested for mtDNA markers. It can be said that mtDNA is the most extensively studied and understood genetic marker in the genome (Salas *et al.*, 2007). The nomenclature used to describe mtDNA phylogeny is based on the ancestral relationship of haplogroups. Single capital letters are used to identify the major monophyletic haplogroups, i.e. that share a common ancestor to the exclusion of others (e.g., haplogroup C), sub-haplogroups within these clades are given numbers (e.g., haplogroup C1), and the branches within the sub-haplogroups are identified by lower case letters (e.g., C1a) and so on, depending on the need for finer scale classifications (Jobling et al., 2004).

Combinations of particular mutations represented as motifs are used to define different haplogroups of the mtDNA phylogenetic tree. For example, if the HVS1 of a particular mtDNA sequence has a T at site 16223, C at site 16298 and T at site 16327, the sequence is classed in haplogroup C. If a transition has also occurred at site 16325, replacing T with C, the sequence can then be placed within the sub-haplogroup C1. Furthermore, if the sequence has a C at site 16356, it can be placed at a specific branch C1a within the sub-haplogroup. The motif for this sequence is: 16223T-16298C-16325C-16327T-16356C.

2.2.2 Human dispersal and mtDNA phylogeography

People have been migrating from one place to another since the first humans emerged. Thus, migration has always shaped the genetic variation of humans. When we think about the dispersal of human populations throughout the world, it is useful to be clear on some definitions. In human evolutionary genetics, *migration* is defined as the movement of people from one inhabited area to another. Migration may result in gene flow, the process by which genes from one population are introduced into another population. By contrast, *colonization* is the process of movements of groups to previously uninhabited territories. In the colonization of new territory there is likely to be a specific form of genetic drift called founder effect. Genetic drift affects genetic diversity in populations, changing the frequency of sequence types through random variation in the reproductive success of individuals. This leads to a reduction in diversity through random extinction of

sequences. A founder effect refers to the reduced genetic diversity that occurs when a new population is established by a small number of individuals from a larger population. Looking back in time it is genetic drift that primarily affects the rate of coalescence of lineages. In the case of mtDNA, genetic drift is caused by random variation in the number of daughters among women (Barton et al. , 2007; Jobling et al., 2004; Strachan & Read, 2004).

When human dispersal is studied by analysis of mtDNA, the focus is on the genetic history of women. The view of migration in human history has always been male-dominated. It involves the image of the strong and masculine hero conquering distant lands. However, studies have shown that when it comes to the exchanging of genes between populations, women's movements play a much bigger role. This is partly due to the widespread practice of patrilocality, i.e. a practice whereby women tend to reside with the family or tribe of the husband, but men tend to remain close to their birthplace upon marriage. The majority of societies are patrilocal, meaning that most mtDNAs are moving between villages each generation. Cultural differences therefore clearly affect genetic diversity along with variables such as population size, genetic drift, gene flow and the presence of selection (Jobling et al., 2004; Stanyon et al., 2009; Stoneking, 1998).

The "Out of Africa" hypothesis argues that every living human being is descended from a small group in Africa that emerged 200 thousand years ago and spread from there to the rest of the world, replacing other hominin species (see Figure 3). The hypothesis is supported by evidence from mtDNA. The phylogeny of human mtDNA revealed a series of population bottlenecks, i.e. a brief, but considerable reduction in population size, leading to a burst of random genetic drift. This would have caused loss of mtDNA diversity moving away from East Africa (Stanyon et al., 2009). As small groups of humans colonized different geographic regions, the new populations became isolated from their source populations. Thus, founder effects from the source populations occurred, i.e., only a few women contributed a limited number of mtDNA to next generations. Also, studies indicate that the discovery of a new habitat increased the survival of offsprings. This would have led to the expansion of a few founder mtDNA types. Consequently, present populations still largely reflect the smaller population size of our past. Human population history can to some effect be represented by a branching tree of isolated populations. However, it must not be forgotten that the

movement of just few individuals greatly reduces neutral divergence between populations (Barton et al., 2007; Forster, 2004; Jobling et al., 2004).

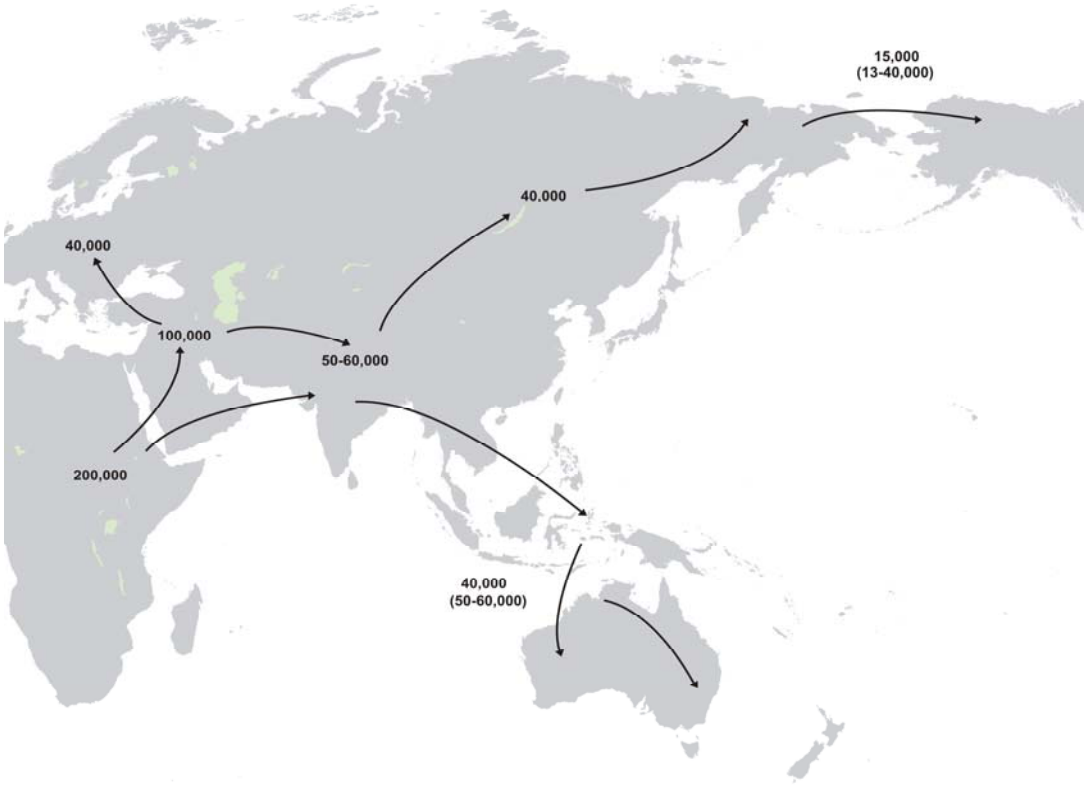

Figure 3. The migration of modern *Homo sapiens*. The scheme outlined above begins with a radiation out of East Africa about 100,000 years ago, followed by an expansion to Asia between 60,000 to 40,000 years ago. Oceania, Europe and America were colonized from Asia in that order (Cavalli-Sforza & Feldman, 2003; Jobling et al., 2004; Schurr, 2004).

The genetic diversity found in modern populations can be used to reconstruct the migration routes and geographic origin of past populations. Phylogeographic studies do precisely that. With the help of mtDNA phylogenetic trees, which are constructed based on the diversity between different haplotypes, it is possible to determine the geographical distributions of different clades. Phylogenetic trees contain clusters of lineages, each of which originated at a particular time and place, following a specific mutation event. Some lineages are geographically restricted, whereas others are extremely widespread. Thus, mtDNA lineages are distributed all around the world with distinguishable differences among geographic regions (Salas et al., 2007; Stanyon et al., 2009). A detailed mtDNA phylogeny is available, based on published mtDNA genome sequences and control region sequences combined with typing of selected coding region SNPs *(single nucleotide polymorphism)* (Jobling et al., 2004). Human mtDNA is characterized by specific distributions

along continents. The earliest mtDNA haplogroups, which make up macro-haplogroup L (L0, L1, L2, L* and L3) (see Figure 2) are largely restricted to Africa. The human dispersal from Africa to Asia is defined by the separation of haplogroup L3 into southern and northern branches, macro-haplogroups M and N, giving rise to almost every mtDNA lineage that exists outside of Africa today. The southern Asia expansion includes haplogroups M*, C and D. As a result of human dispersal these haplogroups spread to Mongolia and Siberia. Haplogroup D is also quite common in Southeast Asia and Japan. The southern Asia expansion also includes haplogroups N*, A, N1, I, W and X. Haplogroup I is thought to have spread to Europe as a result of the first major expansion to the continent. Haplogroups W and X are also found among Europeans. Macrohaplogroup N diverged into the R branch including haplogroups R*, RO, H, V, HV, B, R9, T, J, U and K, which are found in Europe, India and Pakistan. Some Eurasian lineages spread north to America, haplogroups A, B, C, D and X, where they comprise the majority of the Native American gene pool. Note that each haplogroup is divided into series of sub-haplogroups that have spread to further phylogeographic areas (Cavalli-Sforza et al., 1994; Salas et al., 2007; Sun *et al.*, 2007).

3 HISTORICAL BACKGROUND

3.1 Human dispersal to the Americas

3.1.1 Peopling of the Americas

It is thought that modern humans dispersed from Africa about 100,000 years ago and appeared in central Asia about 40,000 years ago. The Americas were one of the last continents to be colonized by humans (Goebel *et al.*, 2008; Stanyon *et al.*, 2009). During the Last Glacial Maximum (LGM) 24,000-14,000 years before present, when the ice sheets were at their maximum extent, Asia and America were connected by the ice-free part of the now submerged Beringian land bridge (Kitchen *et al.*, 2008). It is generally agreed that the ancestors of Native Americans came from northeast Asia via Beringia. However, the timing of the dispersal from Beringia and the size of the founder population remain debated in the literature (Schurr, 2004). Different migration dates have been proposed ranging from 13,000 to over 30,000-40,000 years BP and the size of the founder population has been postulated to be as high as 5,000 to as low as 80 individuals (Bonatto & Salzano, 1997; Goebel et al., 2008; Hey, 2005; Kitchen et al., 2008; Schurr, 2004; Torroni *et al.*, 1993).

The earliest settlers in the Americas are called Paleo-Indian. Until recently, the dominant explanation for the colonization of the New World was the Clovis First model, which maintains that the first Paleo-Indians were people who made a special type of blade tools, first found near the town Clovis in New Mexico (Stein & Rowe, 2003). It has been estimated that the beginning of this material culture dates to between 13,250 and 12,000 years BP (Waters & Stafford, 2007). According to this model, the first human populations entered the continent after the LGM via the Beringian landmass, passing through an interior ice-free corridor in Western Canada. However, a series of new archaeological sites found both in North and South America attest to a human presence in the continent by 14,600 years ago,

suggesting Pre-Clovis occupation (Goebel et al., 2008; Schurr, 2004; Waters & Stafford, 2007).

Several other models of human migration to the New World have also been proposed. One of the most heavily debated is the so-called "Solutrean hypothesis" that suggests an alternative origin for the Clovis culture. It proposes the colonization of North America by Europeans through an ice-edge corridor situated on the edge of the sea-ice that filled the Atlantic Ocean during the LGM. According to this scheme, European hunter-gatherers supposedly migrated across the Atlantic Ocean using boats to sail along the edge of the ice that filled the North Atlantic. However, there is no direct evidence that such sea-voyages actually took place. The hypothesis is primarily based on the apparent similarity of Clovis tools and those associated with the so-called Solutrean culture, dominant in France and Spain, 25,000-18,000 years BP, and the inference that this similarity indicate a common origin and therefore interaction between the two cultures (Westley & Dix, 2008).

The general consensus is that the ancestors of Native Americans entered the Americas via Beringia. Some studies support the so-called Direct colonization model, which purports that the ancestors of Native Americans reached Beringia just before 15,000 years BP and then moved rapidly on into the Americas (Tamm *et al.*, 2007). However, other studies lean more towards the so-called "Three stage model", wherein the ancestors of Native American are thought to have occupied Beringia before the LGM, remaining isolated there until entering the Americas around 15,000 years ago. The first stage of this model represents a single migration from Asia into Beringia, the second stage isolation in Beringia and the third stage a rapid expansion out of Beringia into Americas. This is the most widely accepted hypothesis by human geneticists today (Achilli *et al.*, 2008; Fagundes *et al.*, 2008b; Kitchen et al., 2008; Tamm et al., 2007). In a variation on this theme, an analysis formed by Perego *et al.*, (2009) indicates a dual origin for the first Americans, suggesting that over a short period of time several isolated Beringian source populations migrated into the Americas.

The combination of genetic and archaeological evidence indicates that humans colonized the Americas around 15,000 years ago, after the LGM and that the dispersal occurred along the deglaciated Pacific coastline (Achilli et al., 2008; Goebel et al., 2008). However, some studies suggest another route, through the ice-free corridor between the Laurentide and Cordilleran ice sheets that opened around

14,000 years BP (see Figure 4) (Fagundes et al., 2008b; Perego et al., 2009; Wang *et al.*, 2007). Some studies propose that the population expansion to the Americas started earlier, towards the end of the LGM around 18,000-19,000 years ago (Achilli et al., 2008; Fagundes et al., 2008a; Fagundes et al., 2008b). This would have been possible as the coastal route was largely ice free by 19,000 BP (Kelly, 2002).

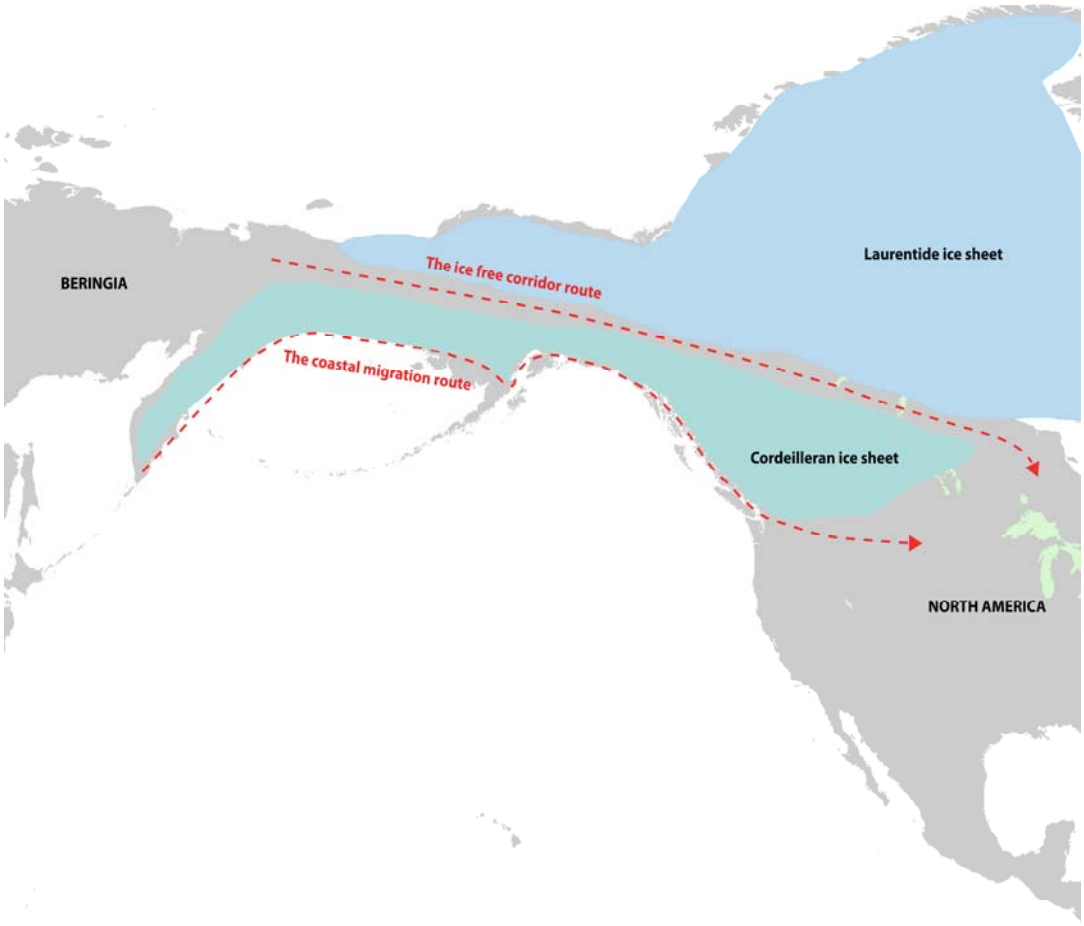

Figure 4. Two suggested migration routes into the Americas.
According the costal migration hypothesis the dispersal from Siberia to America occurred through the Pacific coastline. Some studies suggest another route, through the ice-free corridor between the Laurentide and Cordilleran ice sheets (modified from Westley & Dix, 2008)

3.1.2 Native American haplogroups

Nine Native American mtDNA clades deriving from supposed founding haplotypes have been identified. The first four letters of the alphabet have been used to designate four of the "pan-American" haplogroups: A, B, C and D. Later they were distinguished from their Siberian progenitors by numbers indicating Native American specific sub-clades A2, B2, C1 and D1 (Schurr, 2004; Tamm et al.,

2007). Five less frequent haplogroups, X2a, C4c, D2a, D3, D4h3, have also been identified. These nine haplogroups define the core makeup of the Siberian ancestors of Paleo-Indians that spread to the Americas (Brown, 1998; Perego et al., 2009; Tamm et al., 2007). The most frequent haplogroups, A2, B2, C1 and D1, are found throughout North and South America, suggesting that the initial migration to the continent was likely a swift process rather than a gradual diffusion. Furthermore, their coalescence time is similar suggesting they diverged simultaneously in the continent (Fagundes et al., 2008a; Fagundes et al., 2008b; Perego et al., 2009; Starikovskaya *et al.*, 2005; Tamm et al., 2007; Wang et al., 2007).

The inferred ancient Siberian gene pool, that is thought to be represented by contemporary Siberian populations, is closely related to the Native American haplogroups. Phylogenetic analyses of individuals from Siberia suggest that the founding haplotypes of the Native American mtDNA haplogroups originated in different parts of Siberia (Starikovskaya et al., 2005). The ancestors of Beringian migrants that gave rise to the Native American haplogroup A2 are thought to have originated from the Ket of the Lower Ob and the Mansi of the Lower Yenisei (see Figure 5). The Native American haplogroup B2 has a common origin with the Tubalar and Tuvan, indigenous populations inhabiting the Ob River in the west to the Upper Yenisei region in the east. Haplogroups similar to the Native American C1 and D1 have been found among the Ulchi from the Lower Amur/Sea of Okhotsk region in Siberia, they have been associated with the Selemdzha culture of blade industries, which originated approximately 25,000 years ago in that area (Ingman & Gyllensten, 2007; Starikovskaya et al., 2005; Volodko *et al.*, 2008). Haplogroup C1 has also been detected in Nanai who are located in the same area as the Ulchi, the Buryat population that are Mongolic-speakers who live near Lake Baikal in Southeast Russia and in Japan (Derenko *et al.*, 2007; Ingman & Gyllensten, 2007; Starikovskaya et al., 2005; Tanaka *et al.*, 2004). It is hard to be certain of the precise geographic origin of the "pan-American" haplogroups. The considerable information at hand merely provides us with an idea of possible areas.

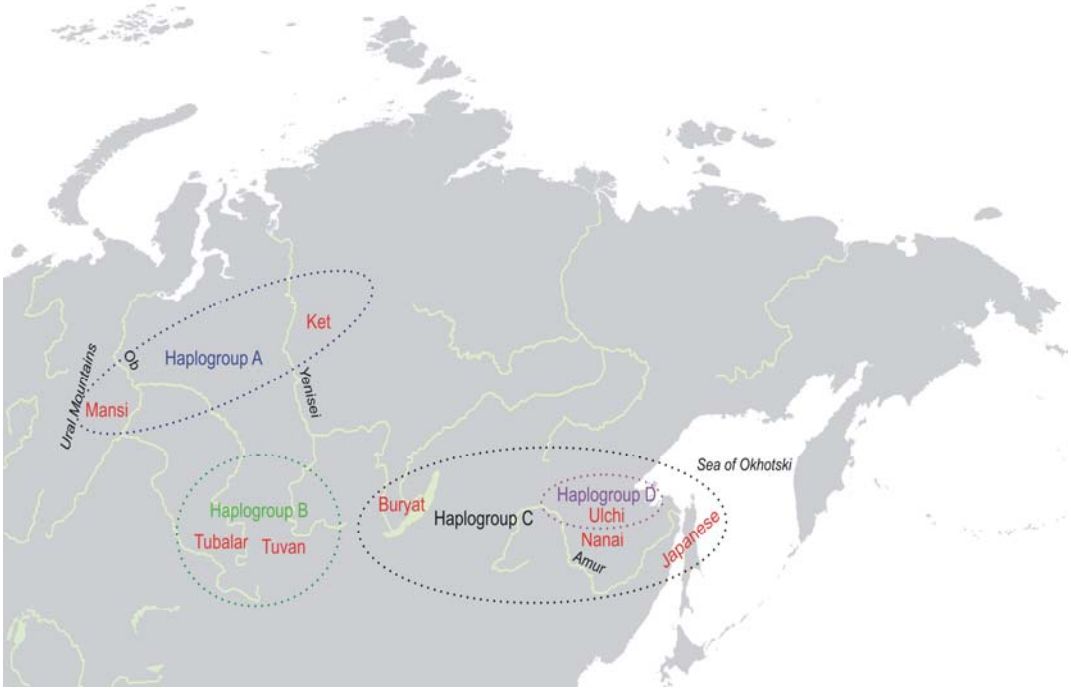

Figure 5. Origin of the four major "pan-American" haplogroups.

The approximate location of the populations carrying each haplogroup is given on the map. Origins of the haplogroups are thought to be somewhere within the circles.

Haplogroup A1 is the most common one in Native Americans, with a highest frequency in Canada, eastern United States and Mexico. Haplogroup B1 is particularly common in Native Americans in the West and Midwest North America, but is virtually absent in the northern parts of North America. Haplogroup D1 is common among native South American tribes, but found in lower frequency in North America. The mtDNA haplogroup diversity of the Na-Dene speaking Native-Americans (Apache, Navajo, Haida and Tlingit) and Eskimo-Aleut speakers (Aleuts, Eskimos and Inuit) differs from that of other Native Americans, in that they carry primarily A and D, but lack haplogroups B and C (Malhi *et al.*, 2002; Schurr, 2004). Haplogroup D3 has been found in Greenland and Canadian Inuit populations (Helgason *et al.*, 2006), whereas D2a is found in Eskimos, Na-Dene and Aleuts (Tamm et al., 2007; Volodko *et al.*, 2008). Haplogroup X can be divided to two major branches, X1 and X2. Branch X1 is restricted to Africa and the Near East, whereas branch X2 is found in Europe, western and Central Asia, North America and the Near East. The X2 branch is found in low frequency in Native Americans (X2a), but almost absent in Siberia and West Eurasians. Some studies claim that haplogroup X shows less diversity and a more recent coalescence time

that haplogroups A-D, indicating more recent population expansion. Thus, it has been argued that haplogroup X represents an independent migration event to the Americas from Asia or even Europe. This has even been used to support the aforementioned "Solutrean hypothesis" (Derenko et al., 2007; Fagundes et al., 2008b; Reidla *et al.*, 2003). However, it should be stressed that it is only a hypothesis that some American founders were of European ancestry and that does not indicate that they came to America through an Atlantic Ocean ice-edge corridor, which is a rather farfetched interpretation.

3.1.3 The phylogeny of haplogroup C

Haplogroup C splits into five distinct branches, C1, C2, C3, C4 and C5 (see Figure 6). All are found in Asia, but three of these branches, C2 C3 and C5, are absent among Native Americans that have been sampled to date (Starikovskaya et al., 2005; Tamm et al., 2007; Volodko et al., 2008). The C4c sub-clade has been found in low frequency in the Americas, but haplogroup C1 is one of the four major "pan-American" haplogroups. The Native American part of haplogroup C1 is found in higher frequency in South America than in North America, with a notable decrease in frequency in Alaska (Malhi *et al.*, 2002; Schurr, 2004). Haplogroup C1 can be divided into four sub-clades, C1a, C1b, C1c and C1d. Three of them, C1b, C1c and C1d, are found in Native Americans. However, sub-clade C1a is only found in Asia. The C1a sub-clade probably diverged from the same ancestral population as the three Native American sub-clades, but before spreading to the Americas. Assuming that the ancestral population of Native American paused in Beringia for some time, during that time specific mutations will have occurred that distinguish the Native American sub-clades lineages from their Asian sister-clades (Tamm et al., 2007).

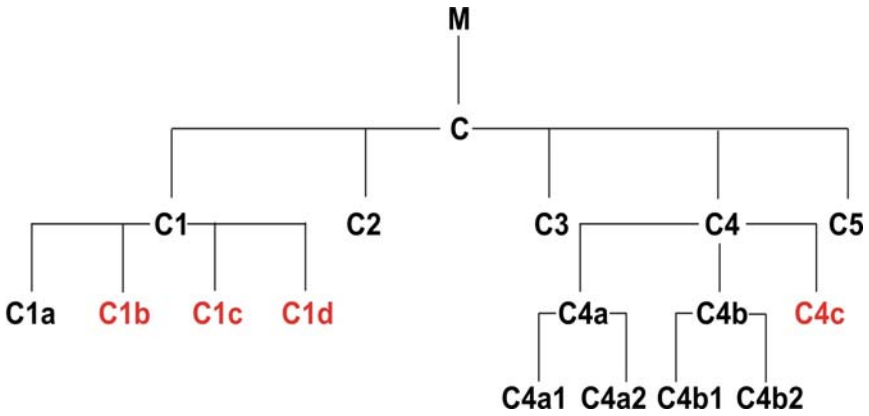

Figure 6. Schematic phylogenetic tree of haplogroup C.
The phylogeny (modified from Starikovskaya et al. 2005; Tamm et al. 2007) is rooted from macro-haplogroup M. Native American branches are defined with red colour.

3.2 The Viking exploration of the West

3.2.1 The settlement of Iceland

The most famous era in Scandinavian history, the Viking age, lasted from the late 8th century through the early 11th century. During that time Norse sailors, named Vikings, traveled extensively in their long ships, expanding east, south and west from Scandinavia. The westward spread, took them from one island to the next: to the Shetland, Orkneys, Hebrides and the Isle of Man were the first to be settled, followed by the Faroe Islands and later Iceland (Helgason *et al.*, 2001; Þorsteinsson & Jónsson, 1991). The settlement of Iceland was probably the outcome of a number of different historical factors, including improved ship-building techniques, overpopulation and crop failures in Scandinavia and the fact that the Norwegian Vikings wanted to escape the rule of Harald Fairhair who had succeeded in unifying Norway during the second half of the 9th century (Thomasson, 1980).

Iceland was one of the world's last major land masses to be colonized by humans. The Icelandic population, furthermore, is the only one in Europe whose origin has been recorded shortly after occurring (Thomasson, 1980). An unusually extensive knowledge of the early history of Iceland can be found in Icelandic medieval texts. The earliest native history that has survived is *Íslendingabók* (the Book of the Icelanders), written between 1122 and 1134 AD by a priest, Ari Þorgilsson (1068-1148). Another remarkable Icelandic historical source is *Landnámabók* (the Book of Settlements), probably composed around 1000 AD,

although it is only preserved in several late versions of the 13th century (Rafnsson, 1997). Although these texts were written some two hundred years after the events they describe, they have frequently been used as reliable accounts of the settlement, because of their descriptive richness. According to these texts the settlement of Iceland can be dated to the period 874-930 AD (Smith, 1995).

Currently available archaeological data suggest that settlement in Iceland began in the late 9th century, thus it does not refute the mediaeval texts chronology for the country's initial settlement (Rafnsson, 1997; Smith, 1995). These results are partly based on an auxiliary science called tephrochronology. A number of tephra layers can be found in the soil as a result of volcanic eruption in Iceland. All archaeological finds that are beneath a definite layer are older than the eruption which produced it and all that is found above it is younger (Karlsson, 2000). The so-called "the settlement layer" was produced by an eruption just around the time of the first settlements, dated from 871(±2 years) AD (Sveinbjarnadóttir, 2004). All remains of human habitation start just above this settlement layer. In many cases, the layer can be found in the turf of which the oldest houses were built, suggesting that the houses were built a few decades after the eruption (Karlsson, 2000). The houses and artifacts of the Vikings along with heathen graves *(kuml)* are the main archaeological evidence for the early settlement in Iceland (Sveinbjarnadóttir, 2004). Before Christianity was mandatory by law, around 1000 AD, people were buried according to heathen customs. These heathen graves are different from the Christian graves. Males were usually buried with weapons or tools, personal belongings and clothes, while women were often buried with jewellery or other personal belongings. Both men and women could also be accompanied by a horse and the corpse was occasionally buried in a boat or a ship. Over 300 heathen burials have been found in Iceland (Rafnsson, 1997; Vésteinsson, 1998).

3.2.2 Icelandic population history

Throughout the years, there has been considerable debate about the ancestry of the original settlers in Iceland. *Landnámabók* tells the story of over 400 original settlers, most of them Norwegian with a minority of Irish and Scots. Most of the settlers did not migrate directly from Norway, but came to Iceland after spending some years in the British Isles. Many brought with them Celtic wives and slaves.

Of those who came directly from Norway, most came from the southwest, from Sogn, Fjordane and Hordaland. Some also came from other parts of Norway and a few came from Denmark and southern Sweden (Karlsson, 2000; Thomasson, 1980). A number of anthropological and genetic studies of blood groups and other serological markers have attempted to estimate the composition of the founding population. However, the results have been inconsistent, ranging from almost entirely Celtic (Thomson, 1973) to primarily Norwegian (Wijsman, 1984). More recent analysis of the genetic origin of Icelanders indicate that contemporary Icelanders trace about 63% of their matrilineal ancestry to Scotland and Ireland and 37% to Scandinavia. In contrary, 75-80% of their patrilineal ancestry originated in Scandinavia, whereas the remaining 20-25% originated in Scotland and Ireland (Helgason et al., 2001; Helgason *et al.*, 2000a; Helgason *et al.*, 2000b). These results suggest that most male settlers were Scandinavian whereas most female settlers came from the British Isles. *Landnámabók* mentions that the Viking settlers brought with them women and slaves from the British Isles. But in light of the substantial maternal contribution from the British Isles, the number of settlers from the British Isles seems to have been greater than mentioned in written sources (Sveinbjarnadóttir, 2004).

In *Landnámabók* it says "…that in the space of sixty years Iceland was fully occupied, so that after that there was no further taking of land" (Thomasson, 1980). Archeological evidence indicates that in the last years of the settlement period, almost all the habitable areas of Iceland were taken (McGuire, 2006). It is likely its settlement were accomplished by many expeditions that were launched from various places in Norway as well as from the Viking settlements in the Orkneys, Shetland, the Hebrides and in Ireland (Rafnsson, 1997). Historians have estimated, based on the written sources, that over the period of settlement, around 8,000-20,000 individuals were involved (Helgason et al., 2000a). Over the next five hundred years the Icelandic population experienced several drastic declines in size due to famines and epidemics. It was not until after 1830 AD that the population began to increase, growing rapidly to the present size of 319,246 (Íslands, 2009; Thomasson, 1980). The extremely high death rates that have occurred throughout the centuries have affected the Icelandic gene pool. Providing many opportunities for genetic drift which have reduced the genetic diversity introduced by the original settlers (Helgason *et al.*, 2003b). Moreover, because of the natural barrier of the

North Atlantic that hindered post-settlement immigration to the island, the Icelandic population has not been affected by constant gene flow like many other populations. Thus, it can be assumed that the vast majority of mtDNA lineages observed in contemporary Icelanders are descended from the original set of mtDNA lineages present in the female settlers (Helgason *et al.*, 2009; Helgason et al., 2000a).

3.2.3 The Viking discovery of America

Towards the end of the 10th century, expeditions brought Vikings to Greenland and a few years later to North America. Around the year 1000 AD, five hundred years before Christopher Columbus is said to have discovered America, Vikings had in fact discovered and tried to colonize the continent. Remains of buildings from this period, with evidence of temporary occupation by Vikings, have been found at L'Anse aux Meadows in Newfoundland (Sawyer, 1997). The three Viking buildings found there seem to have been used at the same time and together they could accommodate 70-90 people. They are distinctly Icelandic in style, the same style as used by the initial Norse settlers in Greenland. Radiocarbon dating of the site indicates that it was occupied sometime between 980 and 1020 AD (Fergusson, 2001; Rafnsson, 1997). It is tempting to interpret this site as base camp for further exploration. However, there is no reliable evidence for that (Sawyer, 1997).

The Viking discovery of America is described in two Icelandic sagas, *Grænlendinga saga* (the Saga of the Greenlanders) and *Eiríks saga rauða* (Eric the Red´s Saga). The two sagas, often referred to as the Vínland Sagas, belong to a category of medieval Icelandic literature known as *Íslendingasögur* (the Sagas of the Icelanders). *Grænlendinga saga* is the older of the two dating from the year 1200 AD, whereas *Eiríks saga rauða* was probably composed between 1263 to 1300 AD (Rafnsson, 1997; Williamsen, 2005; Þorláksson, 2001). The Icelandic sagas are not contemporaneous with the events they describe and not historical documents as such. However, many scholars think that the historical reliability of the texts is high, although they should not be used purely as historical sources (Halldórsson, 2001; Þorláksson, 2001). The Vínland sagas are not the only documents which mention the Viking discovery of America, it is also briefly depicted in the aforementioned *Íslendingabók, Heimskringla* (The Chronicle of the

Kings of Norway), written ca. 1230 AD, and in the writings of Adam of Bremen, a German medieval chronicler, born ca. 1070 AD (Karlsson, 2000).

According to the sagas, a Norwegian born Icelandic settler, named Eiríkur Rauði *(Eric the Red)* found an ice covered land around the year 985 AD, which he named Greenland. Archeological finds show that the Norse settlement in Greenland comprised nearly 4000 people (Rosenblad & Sigurðardóttir-Rosenblad, 1993). The Viking settlements were on the west coast of Greenland. The richest settlement was in the south, known as the Eastern Settlement, and further north, near the modern capital of Greenland, Nuuk, was another settlement, known as the Western Settlement (see Figure 7) (Rafnsson, 1997). *Grænlendinga saga* tells the story of the Icelander Bjarni Herjólfsson. He blew off course sailing from Iceland to Greenland, in 985 or 986 AD, and accidentally discovered a land to the west of Greenland. Bjarni and his ship members discussed among themselves what land it could be and Bjarni was sure it could not be Greenland because it was flat and covered with woods whereas Greenland was covered by ice. The land Bjarni saw was most likely the land south of Hamilton Inlet, likely near what is now known as Sandwich Bay in Labrador (see Figure 7). Then Bjarni sailed for two days before seeing land again, which also was flat and wooded. Bjarni was, once again, convinced that this was not Greenland and decided not to go ashore. They sailed for one and a half day until they turned once again towards land. This land was described as high and mountainous with ice covered mountain tops. They sailed along the shore and saw that it was an island. This may have been Resolution Island, which is south of Baffin Island. Eventually, Bjarni held out to sea reaching Greenland in a few days time (Bergþórsson, 2000; Magnusson & Pálsson, 1965).

This discovery by Bjarni paved the way for Leif Eríksson, son of Eric the Red, who heard the story from Bjarni and according to both *Grænlendinga saga* and *Eiríks saga rauða* sailed back to those areas, around the year 1000. Leif first reached a rocky shore where glaciers covered the highlands to which he gave the name *Helluland* (Stone-slab land), most likely what is now known as Baffin Island (see Figure 7). The next anchorage was off a land that was flat and heavily wooded with long white-sanded beaches. He named this country *Markland* (Forest Land). There are white sands south of Cape Porcupine and also in a place reached after passing through the Strait of Belle Isle between Labrador and Newfoundland. Thus, it is likely that Markland occupies the land somewhere in that area. Continuing his

journey, Leif sailed along the shore of Labrador into the Gulf of St. Lawrence. Allegedly, he and his companions went ashore somewhere around Québec, naming the land *Vínland* (Wine-land), because of the grapes they found there (Bergþórsson, 2000). Scholars do not agree on the exact location of *Vínland*, but it is thought to be somewhere between Maryland to Labrador (Williamsen, 2005). Leif and his companions decided to spend winter in *Vínland*, building houses before returning to Greenland when spring came (Bergþórsson, 2000).

According to the saga material, other adventures followed after Leif and begun colonization of this new found land. One of them was Þorfinnur Karlsefni who sailed there with three ships carrying 160 people (Karlsson, 2000). The Sagas tell stories of relations between the settlers and the natives, whom the Vikings called *Skrælingjar*. The Vikings, most likely, encountered the Micmac or the Beothuk Indians, depending on the location in question (Williamsen, 2005) According to the sagas the Vikings and *Skrælingjar* sometimes traded peacefully, but sometimes they fought. As well as being numerically superior, the natives could make new weapons locally, making them the stronger party. One theory as to why the Vikings failed to colonize the continent, is that they retreated because of continuous attacks by the natives (Karlsson, 2000; Rafnsson, 1997; Rosenblad & Sigurðardóttir-Rosenblad, 1993). Although a short-lived venture, these trips may have led to a female being transported to Iceland and giving rise to the haplogroup C mtDNA lineage in Iceland.

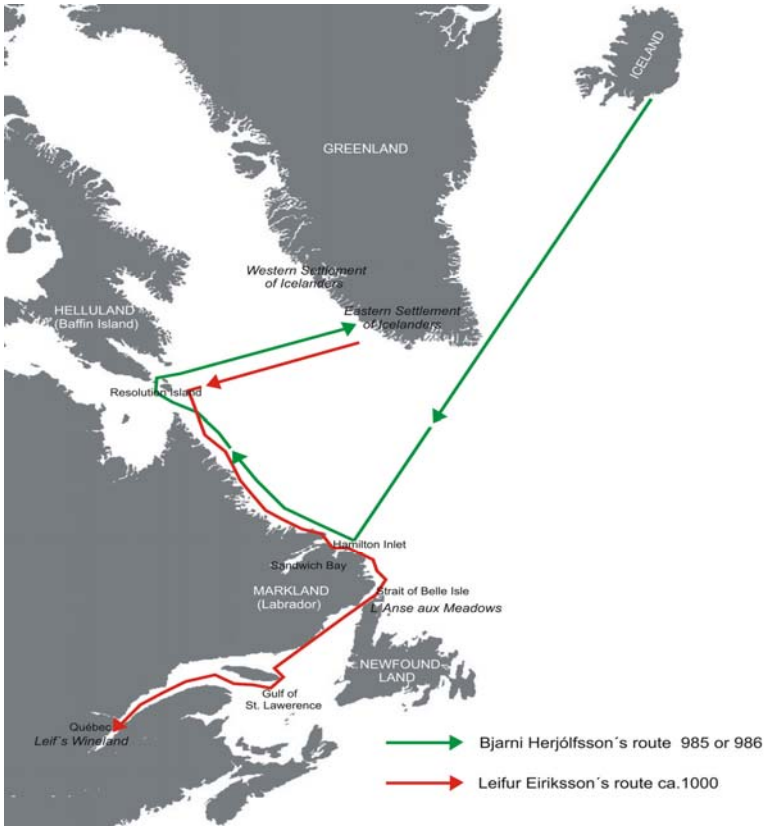

Figure 7. A map of the Viking voyages.

This map shows the Viking voyages main locations and Bjarni Herjólfsson´s and Leifs Eiriksson´s sailing routes.

4 METHODS

4.1 The genealogical database

Matrilineal genealogies were extracted from an encrypted version of the deCODE Genetics genealogical database, which contains information about 720,000 Icelanders, including almost all ancestors of the contemporary population from 1650 AD. Information for each individual, which has been assigned a unique encryption code, is recorded in the database, including gender, codes for the father and mother of the individual, date of birth and death. All dates are rounded to the nearest multiple of 5 (for example, the birth year 1710 stands for the range 1708-1712, and 1715 stands for the range 1713-1717). Furthermore, information about geographic location of individuals is recorded, obtained from censuses and parish registers (Helgason *et al.*, 2005). The encrypted version of the genealogical database often has information about the county and usually also the parish for most individuals. To work with the data I primarily used the Microsoft Access (database software). In this study two primary data sets were used. The first contained 1,538 Icelanders that had previously been sequenced for the mtDNA control region. The second contained 744 Icelanders sequenced in this study for the mtDNA control region (sites 15,811 to 775).

4.2 DNA sequencing

The DNA samples used in this study were previously obtained from Icelandic volunteers for genetic studies of disease by deCODE Genetics, where an appropriate informed consent was obtained for each individual. In this study, mtDNA nucleotide sites were numbered as in the Cambridge Reference Sequence (CRS) (Anderson et al., 1981) and mutation motifs are expressed as differences from the revised Cambridge reference sequence (rCRS) (Andrews *et al.*, 1999). A total of 11 individuals were selected for complete mtDNA sequencing. The

complete mitochondrial genome was amplified by PCR using the primers reported by Rieder *et al*., (1998) (see Table 1). A total of 744 individuals were sequenced for the control region (sites 15,811 to 774) using primer pairs 23 and 24 (see Table 1). To simplify sequencing protocols, all forward primers were prefixed with the universal M13 forward primer (-20, 5′-GTAAAACGACGGCCAG-3′) and all reverse primers suffixed with the universal M13 reverse primer (5′-CAGGAAACAGCTATGAC-3′). Each amplification reaction was carried out using 3µl of template DNA, 2µl of each primer (forward and reverse), 4µl Betaine (5M), 0.16µl MQ water, 1.6µl 10x reaction buffer 1(with NH_2SO_2) (Fermentas), 0.16µl $MgCI_2$ (250mM), 1.84µl dNTP′s (2mM) and 0.24µl Taq (Fermentas) in a final volume of 20µl. In a thermal cycler, 35 cycles of 95°C (10 mins), 95°C (30 sec), 60°C (30 sec) and 72°C (1 min) were followed by a final 10 mins at 72°C. PCR products were resolved in a 2.5% agarose gel, stained with ethidium bromide and visualized by UV light.

Following DNA amplifications, unincorporated PCR primers and deoxynucleotide triphosphates in the samples were inactivated prior to sequencing by an enzymatic treatment. This was accomplished by adding 0.1µl exonuclease (10units/µl; Usb®) and 2µl shrimp alkaline phosphatase (1units/µl; Usb®) to each 20µl PCR product and incubating at 37°C for 30 mins followed by 75°C for 15 mins to inactivate the exonuclease and alkaline phosphatase enzymes. Cycle sequencing was performed using the BigDye® Terminator v3.1 Cycle Sequencing Kit. Each amplification reaction was carried out using 2µl cleaned PCR product, 5.8µl cycle sequencing mix and 0.8µl M13 (forward or reverse primer). The plates were placed in a thermocyler and run for 25 cycles of 96°C for 10 sec, 55°C for 5 sec and 60°C for 4 mins. After high-throughput purification, the extension products were loaded onto a Cymark Sequencer. The resulting chromatograms were aligned and manually checked using the SeqManII software (Swindell & Plasterer, 1997).

Sequencing of the control region resulted in high quality sequences between sites 15,811 and 775 for all 744 individuals. Sequencing of complete mtDNA genomes resulted in high quality sequences for all eleven individuals, except for a 327 nucleotide long fragment in the coding region (sites 5895-6222). One coding region fragment, sites 11948-12775, was amplified for two additional haplogroup C carriers, using primer pair 18 (see Table 1), and another segment of the coding

segment, sites 13338-14268, was amplified for six additional haplogroup C carriers, using primer pair 20 (see Table 1).

As chromatograms were unclear between sites 5895 and 6222, due to an apparent insertion of multiple Cs, PCR amplifications for primer pairs 8 and 9 (sites 5255-6642) (see Table 1) were cloned using TOPO® cloning kit. Each cloning reaction was carried out using 2µl of PCR product, 0.5µl salt solution, 2.5µl dilute salt solution and 1µl TOPO® vector in a final volume of 6µl. Chemical transformation was carried out using one Shot® *E.coli* that was thawed on ice, 2µl TOPO cleaning reaction was added to a vial of One Shot® *E.coli* and mixed, incubated in ice for 30 minutes and then the cells were heat-shocked for 30 seconds at 42°C. Recovery and plating was carried out using 250µl of room temperature SOC medium added to cells, then the tubes were shaken at 37°C for 1 hour, 20µl from each transformation was spread into LB plates containing X-Gal and 50µg/ml kanamycin. The plates were then incubated overnight at 37°C, after which the plasmid DNA was isolated and analyzed by sequencing. A total of eleven samples were cloned, but only seven gave results. These seven samples resulted in a total of 41 clones.

Table 1. PCR primers used for mtDNA amplification.

The universal M13 forward and reverse primers are indicated in red. The amplification length and overlap shown in the table is based on the primers alone (i.e. not including the M13 primers). The overlap number given on the right side of each primer pair refers to the overlap with the subsequent primers pairs.

Primer Name	Primer Sequence 5´- 3´ with M13 F or R primers	Start Site	Length	Overlap
1F	GTAAAACGACGGCCAGCTCCTCAAAGCAATACACTG	611		
1R	CAGGAAACAGCTATGACTGCTAAATCCACCTTCGACC	1411	840	202
2F	GTAAAACGACGGCCAGCGATCAACCTCACCACCTCT	1245		
2R	CAGGAAACAGCTATGACTGGACAACCAGCTATCACCA	2007	802	204
3F	GTAAAACGACGGCCAGGGACTAACCCCTATACCTTCTGC	1854		
3R	CAGGAAACAGCTATGACGGCAGGTCAATTTCACTGGT	2669	860	196
4F	GTAAAACGACGGCCAGAAATCTTACCCCGCCTGTTT	2499		
4R	CAGGAAACAGCTATGACAGGAATGCCATTGCGATTAG	3346	887	208
5F	GTAAAACGACGGCCAGTACTTCACAAAGCGCCTTCC	3169		
5R	CAGGAAACAGCTATGACATGAAGAATAGGGCGAAGGG	3961	832	215
6F	GTAAAACGACGGCCAGTGGCTCCTTTAACCTCTCCA	3796		
6R	CAGGAAACAGCTATGACAAGGATTATGGATGCGGTTG	4654	898	203
7F	GTAAAACGACGGCCAGACTAATTAATCCCCTGGCCC	4485		
7R	CAGGAAACAGCTATGACCCTGGGGTGGGTTTTGTATG	5420	975	207
8F	GTAAAACGACGGCCAGCTAACCGGCTTTTTGCCC	5255		
8R	CAGGAAACAGCTATGACACCTAGAAGGTTGCCTGGCT	6031	814	201
9F	GTAAAACGACGGCCAGGAGGCCTAACCCCTGTCTTT	5855		
9R	CAGGAAACAGCTATGACATTCCGAAGCCTGGTAGGAT	6642	827	214
10F	GTAAAACGACGGCCAGCTCTTCGTCTGATCCGTCCT	6469		
10R	CAGGAAACAGCTATGACAGCGAAGGCTTCTCAAA TCA	7315	886	211
11F	GTAAAACGACGGCCAGACGCCAAAATCCATTTCACT	7148		
11R	CAGGAAACAGCTATGACCGGGAATTGCATCTGTTTTT	8095	987	205
12F	GTAAAACGACGGCCAGACGAGTACACCGACTACGGC	7937		
12R	CAGGAAACAGCTATGACTGGGTGGTTGGTGTAAATGA	8797	900	196
13F	GTAAAACGACGGCCAGTTTCCCCCTCTATTGATCCC	8621		
13R	CAGGAAACAGCTATGACGTGGCCTTGGTATGTGCTTT	9397	816	214
14F	GTAAAACGACGGCCAGCCCACCAATCACATGCCTAT	9230		
14R	CAGGAAACAGCTATGACTGTAGCCGTTGAGTTGTGGT	10130	940	205
15F	GTAAAACGACGGCCAGTCTCCATCTATTGATGAGGGTCT	9989		
15R	CAGGAAACAGCTATGACAATTAGGCTGTGGGTGGTTG	10837	891	182
16F	GTAAAACGACGGCCAGGCCATACTAGTCTTTGCCGC	10672		
16R	CAGGAAACAGCTATGACTTGAGAATGAGTGTGAGGCG	11472	840	203
17F	GTAAAACGACGGCCAGTCACTCTCACTGCCCAAGAA	11314		
17R	CAGGAAACAGCTATGACGGAGAATGGGGGATAGGTGT	12076	802	196
18F	GTAAAACGACGGCCAGTATCACTCTCCTACTTACAG	11948		
18R	CAGGAAACAGCTATGACAGAAGGTTATAATTCCTACG	12772	866	166
19F	GTAAAACGACGGCCAGAAACAACCCAGCTCTCCCTAA	12571		
19R	CAGGAAACAGCTATGACTCGATGATGTGGTCTTTGGA	13507	977	242
20F	GTAAAACGACGGCCAGACATCTGTACCCACGCCTTC	13338		
20R	CAGGAAACAGCTATGACAGAGGGGTCAGGGTTCATTC	14268	970	207
21F	GTAAAACGACGGCCAGGCATAATTAAACTTTACTTC	14000		
21R	CAGGAAACAGCTATGACAGAATATTGAGGCGCCATTG	14998	938	206
22F	GTAAAACGACGGCCAGTGAAACTTCGGCTCACTCCT	14856		
22R	CAGGAAACAGCTATGACAGCTTTGGGTGCTAATGGTG	15978	1162	180
23F	GTAAAACGACGGCCAGTCATTGGACAAGTAGCATCC	15811		
23R	CAGGAAACAGCTATGACGAGTGGTTAATAGGGTGATAG	5	765	205
24F	GTAAAACGACGGCCAGCACCATTCTCCGTGAAATCA	16420		
24R	CAGGAAACAGCTATGACAGGCTAAGCGTTTTGAGCTG	775	954	203

4.3 Phylogenetic analysis

Median-joining networks (Bandelt *et al.*, 1995) were generated using the program PhyloNet (written by Dr. Agnar Helgason). Polymorphic sites were weighted to correct for mutational hotspots. Sites in the control region networks were weighted according to the following scheme, five sites were given the low weight 1: 16093, 16183i, 16189, 16311, 152, and two sites were given the low weight 2: 522-523d and 16519. In a median joining network generated from HVS1 (sites 16024-16383) and HVS2 (sites 1-534) sequences, three sites that are characteristic of C1 sub-clades were given high weight (8): 16051, 493 along with 249d which is defining for CZ, this was needed because the revised Cambridge reference sequence (rCRS) was included. Furthermore, the dinucleotide (CA) repeat situated between sites 522 and 523 was ignored in this network along with any length variation or mutations in the C-stretch between nucleotides at sites 302-315 because they interfered with the structure of the network. In a median-joining network generated from HVS1 (sites 16040-16400) sequences three sites were given high weight (8): 16223, 16298 and 16327. This was necessary so that the revised Cambridge reference sequence (rCRS) would not get lost in the structure of the network. The remaining sites were assigned the weight of 4.

4.4 Sequence data from the literature

The sample-set used in this study includes a total of 61 complete and 13 coding region C1 sequences published in GenBank by the end of May 2009. Haplogroup C1 membership was inferred by the control region motif: 16223T-16298-16325C-16327T for HVS1 and 73G-249d-236G-290d-291d. The sequence data along with population information is listed in Table 2. In addition to the 61 complete sequences listed in Table 2, a total of 213 control region C1 sequences, HVS1 (sites 16040-16400), were collected from the literature and used to construct a median joining-network. Of these 213 sequences, 92 were deposited in GenBank by Monson K.L., (2002) (FBI database), 84 by Behar et al., (2007), 28 by Vona *et al.*, (2005) two by Kolman *et al.*, (1996), one by Vigilant *et al.* (1991), six sequences remain unpublished to date, but can be found in GenBank under the accession numbers, AB059881, AB059891, AB059931, AB059939, AB059963,

AB059969, AB059881 and AB059891. All 61 complete C1 sequences (Table 2) and all but one of the 92 sequences deposited in GenBank by Monson K.L., (2002) (FBI database) were used along with one additional sequence (Brandstatter *et al.*, 2007) to construct a median-joining network for control region HVS1 (sites 16024-16383) and HVS2 (sites 1-534) sequences. These sequences were downloaded in FASTA format from the GenBank database and aligned using SeqManII (Swindell & Plasterer, 1997) sequence analysis software and a sequence manipulation tool (SeqTool) written by Dr. Agnar Helgason. The automatic alignment was manually checked, in order to ensure consistency in the handling of deletions and insertions. Sequences that had ambiguous nucleotides (N) were included in the analysis if it could be inferred that the Ns were unlikely to be mutations, i.e. if no other haplogroup C sequence had a confirmed mutation on the site in question then we concluded that the N represented the consensus allele state of haplogroup C sequences. In cases where we could not make such a conclusion the sequence was dropped from the analyses.

Table 2. List of complete and coding region sequences from haplogroup C collected and used in this study. The label refers to numbers or/and names used in this study to refer to the sequences. These special labels were used to differentiate sequences. Note that in this study sequences obtained from Just et al. (2008) are referred to by the name Parsons, as was the author name provided when the sequences were originally deposited in the GenBank

GenBank ID	Reference/submitted	Geographic or ethnic origin	Haplogroup	Label	Note
AY519496	Starikovskaya et al., (2005)	Ultchi, Siberia	C1a		
AP008311	Tanaka et al., (2004)	Japanese	C1a		
EF153779	Derenko et al., (2007)	Buryat, Siberia	C1a		
EU007858	Ingman & Gyllenstein, (2007)	Nanaitci, Siberia	C1a		
DQ282447	Just et al., (2008)	Hispanic	C1b	Parsons 7	
DQ282448	Just et al., (2008)	Hispanic	C1b	Parsons 8	
DQ282449	Just et al., (2008)	Hispanic	C1b	Parsons 9	
DQ282450	Just et al., (2008)	Hispanic	C1b	Parsons 10	
DQ282451	Just et al., (2008)	Hispanic	C1b	Parsons 11	
DQ282452	Just et al., (2008)	Hispanic	C1b	Parsons 12	
DQ282453	Just et al., (2008)	Hispanic	C1b	Parsons 13	
DQ282454	Just et al., (2008)	Hispanic	C1b	Parsons 14	
DQ282455	Just et al., (2008)	Hispanic	C1b	Parsons 15	
DQ282456	Just et al., (2008)	Hispanic	C1b	Parsons 16	
DQ282457	Just et al., (2008)	Hispanic	C1b	Parsons 17	
DQ282458	Just et al., (2008)	Hispanic	C1b	Parsons 18	
DQ282461	Just et al., (2008)	Hispanic	C1b	Parsons 21	
DQ282464	Just et al., (2008)	Hispanic	C1b	Parsons 24	
DQ282469	Just et al., (2008)	Hispanic	C1b	Parsons 29	
DQ282475	Just et al., (2008)	Hispanic	C1b	Parsons 35	
DQ282476	Just et al., (2008)	Hispanic	C1b	Parsons 36	
EU431085	Achilli et al. ,2008)	USA	C1b	2	
EU597545	Hartman et al.,(2009)	Mexico, Pima	C1b		
EU597557	Hartman et al., (2009)	Mexico, Pima	C1b		
EF657584	Herrnstadt et al. (2002)	Native American	C1b	417 Herrnstadt	Coding area only
DQ112846	Kivisild et al., (2006)	Navajo, North America	C1b		Coding area only
EF657282	Herrnstadt et al., (2002)	Native American	C1b	145 Herrnstadt	Coding area only
EU095549	Tamm et al., (2007)	Wayuu, Colombia and Venezuela	C1b		
AF382009	Maca-Meyer et al., (2001)	Canary	C1b		
EU095222	Fagundes et al., (2008)	WaiWai, Brazil and Guyana	C1b		

EU095223	Fagundes et al., (2008)	Zoro, South America	C1b	2	
EU095224	Fagundes et al., (2008)	Zoro, South America	C1b	1	
EU095225	Fagundes et al., (2008)	Quechua, South America	C1b	2	
EU095226	Fagundes et al., (2008)	Quechua, South America	C1b	1	
EU095227	Fagundes et al., (2008)	Arara, Brazil	C1b		
EU095228	Fagundes et al., (2008)	Poturujara, South America	C1b		
EU095229	Fagundes et al., (2008)	Yanomamö, Brazil and Venezuela	C1b	2	
EU095230	Fagundes et al., (2008)	Yanomamö, Brazil and Venezuela	C1b	3	
EU095231	Fagundes et al., (2008)	Yanomamö, Brazil and Venezuela	C1b	1	
AY195759	Mismar et al., (2003)	North America	C1b		
DQ282462	Just et al., (2008)	Hispanic	C1c	Parsons 22	
DQ282463	Just et al., (2008)	Hispanic	C1c	Parsons 23	
DQ282465	Just et al., (2008)	Hispanic	C1c	Parsons 25	
DQ282466	Just et al., (2008)	Hispanic	C1c	Parsons 26	
DQ282467	Just et al., (2008)	Hispanic	C1c	Parsons 27	
DQ282459	Just et al., (2008)	Hispanic	C1c	Parsons 19	
DQ282460	Just et al., (2008)	Hispanic	C1c	Parsons 20	
DQ282468	Just et al., (2008)	Hispanic	C1c	Parsons 28	
DQ282470	Just et al. ,(2008)	Hispanic	C1c	Parsons 30	
DQ282471	Just et al., (2008)	Hispanic	C1c	Parsons 31	
EF079875	Achilli et al., (2008)	Dominican Republic	C1c		
EU431086	Achilli et al., (2008)	Canada	C1c		
EU431087	Achilli et al,. (2008)	USA	C1c	1	
EU617323	Family Tree DNA	Unknown	C1c	Greenspan	
EU327973	Family Tree DNA	Unknown	C1c	Greenspan	
EU327891	Family Tree DNA	Unknown	C1c	Greenspan	
EF657317	Herrnstadt et al., (2002)	Native American	C1c	177 Herrnstadt	Coding area only
EF657324	Herrnstadt et al., (2002)	Native American	C1c	183 Herrnstadt	Coding area only
EF657329	Herrnstadt et al., (2002)	Native American	C1c	188 Herrnstadt	Coding area only
EF657355	Herrnstadt et al., (2002)	Native American	C1c	210 Herrnstadt	Coding area only
EF657547	Herrnstadt et al., (2002)	Native American	C1c	384 Herrnstadt	Coding area only
EF657588	Herrnstadt et al., (2002)	Native American	C1c	420 Herrnstadt	Coding area only
DQ112888	Kivisild et al., (2006)	Maya, South America	C1c		Coding area only
EU095527	Tamm et al., (2007)	Arsario, South America	C1c		

EU095544	Tamm et al., (2007)	Kogui, Colombia	C1c		
DQ282472	Just et al., (2008)	Hispanic	C1d	Parsons 32	
DQ282473	Just et al., (2008)	Hispanic	C1d	Parsons 33	
DQ282474	Just et al., (2008)	Hispanic	C1d	Parsons 34	
AF347012	Ingman et al., (2000)	Warao, Venezuela and Guyana	C1d		
AF347013	Ingman et al., (2000)	Warao, Venezuela and Guyana	C1d		
EF657504	Herrnstadt et al., (2002)	Native American	C1d	345 Herrnstadt	Coding area only
DQ112789	Kivisild et al., (2006)	Colombia	C1d		Coding area only
EU095537	Tamm et al., (2007)	Corequaje, Colombia	C1d		
EF657314	Herrnstadt et al., (2002)	Native American	C1*	174 Herrnstadt	Coding area only

URLs for data presented herein are as follows: GenBank, http://www.ncbi.nlm.nih.gov/Genbank/

5 RESULTS

5.1 Icelandic haplogroup C carriers

In a previous study of mtDNA variation in Icelanders (Helgason et al., 2000a) two copies of a haplogroup C sequences were found in a sample of 401 control region sequences (see Figure 8). In a subsequent study, 552 new Icelandic control region sequences were published (Helgason et al., 2003b). Additionally, 585 unpublished Icelandic control region sequences were available at deCODE Genetics. In all, mtDNA control region sequences from a random sample of 1538 Icelanders were available when this study began, which revealed five individuals to be carriers of identical haplogroup C control region mtDNA sequences. No other haplogroup C lineages were found in this sample. It was suspected that one of these five individuals did not carry haplogroup C, because his mother carried a different sequence, indicating the possibility of a sample handling error. This sample was included in this study to determine whether the error occurred in the handling of his or his mother's sample.

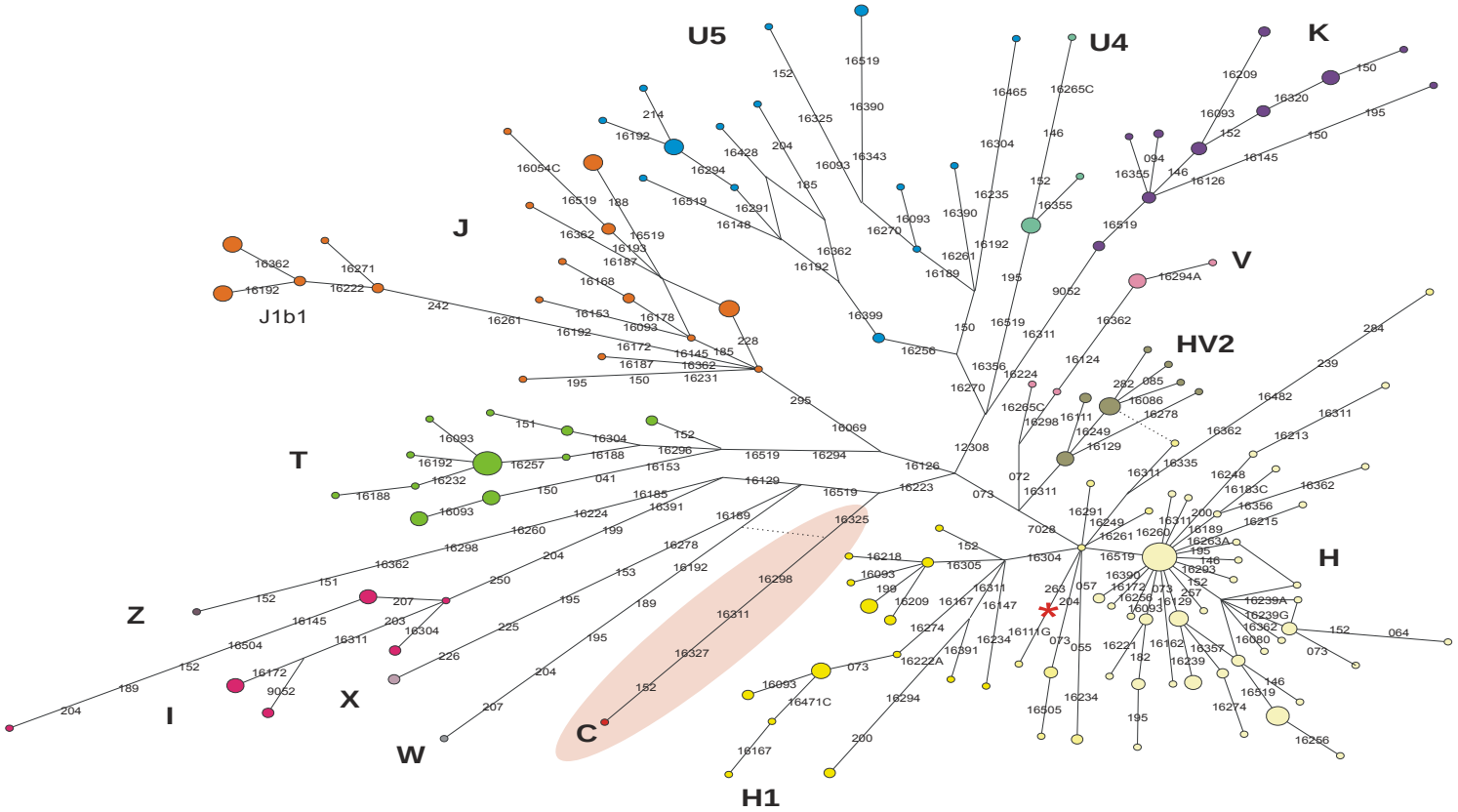

Figure 8. Network of mtDNA lineages from the Icelandic population.

This median-joining network (modified from Helgason et al. 2000a) shows HVS1 (16028-16519), HVS2 (1-297), and the coding region SNPs 7028, 9052 and 12308. The bold labels represent the haplogroups the lineages belong to. Circles are proportional to lineage frequency, the smallest representing single lineage and the largest 21 copies of the same lineage. Lines represent genetic differences. The numbers by the lines represent the site of mutations (transversions are indentified by site number and base). Insertions and deletions are not shown in the network. The Cambridge reference sequence is indicated by a red asterisk. The haplogroup C lineage is identified by an orange shaded ellipse.

The control region motif for the Icelandic haplogroup C sequence is 16223T-16298C-16311C-16325C-16327T for HVS1 and 73G-152C-249d-263G-290d-291d for HVS2. Haplogroup C membership is inferred by the control region motif 73G-249d-263G-16223T-16298C-16327T. Haplogroup C can be divided into several sub-clades, one of which, haplogroup C1, is distinguished by a transition at site 16325 and a deletion at site 290-291 (see Table 3) (Starikovskaya *et al.*, 2005; Tamm *et al.*, 2007).

Table 3. Control region mutations used to define haplogroup C and C1 sub-clade.
Mutations that are sown in bold are haplogroup specific mutations.

Hg	HVS1	HVS2
C	**16223T 16298C 16327T**	**73G 249d 263G**
C1	16223T 16298C **16325C** 16327T	73G 249d 263G **290-291d**
C1a	16223T 16298C 16325C 16327T **16356C**	73G 249d 263G 290-291d
C1b	16223T 16298C 16325C 16327T	73G 249d 263G 290-291d
C1c	16223T 16298C 16325C 16327T	73G 249d 263G 290-291d
C1d	**16051G** 16223T 16298C 16325C 16327T	73G 249d 263G 290-291d

On the basis of this information, it could be concluded that the Icelandic haplogroup C sequence belonged to haplogroup C1. Four sub-clades of C1, named a, b, c and d have been identified to date (Starikovskaya et al., 2005; Tamm et al., 2007). Based only on the control region, there is no telling where the Icelandic sequence belongs in this classification system. It is, however, unlikely the Icelandic sequence belongs to C1a, because the Icelandic motif does not carry the 16356C mutation that characterizes C1a sequences. Haplogroup C1a sequences have only been found in Asian populations whereas the other C1 sub-haplogroups are only found in individuals with Native American ancestry. In light of this it was suspected that the Icelandic C1 sequence had a Native American origin, although it could be surmised that the Icelandic C1 sequence did not belong to haplogroup C1d, since it does not carry the 16051G transition. The Icelandic sequence has two additional control region mutations, 16311C and 152C, that are not common among existing C1 control region sequences.

5.1.1 The matrilineal genealogies of existing haplogroup C carriers

Because haplogroup C sequences are unusual in European populations we wanted to know more about the history of this lineage in the Icelandic population. Using the deCODE Genetics genealogical database, the matrilines of the five existing haplogroup C1 carriers were reconstructed. The database records genealogical links for about 720,000 individuals, including all 320,000 living Icelanders and most of their ancestors. All analyses were performed on an encrypted version of the genealogical database, where names are excluded and birth dates are rounded to the nearest multiple of 5. The matrilines of the five individuals deemed to be haplogroup C1 carriers were traced back in time to their earliest known ancestor. Three of the individuals were descended from the same ancestor, called A, born 1720 (see Figure 9, green circles). The matriline of the fourth individual was traced back to a different ancestor called B, born 1740 (see Figure 10, green circle). Finally, the fifth individual, the one that was likely a sample handling error, was traced back to yet another ancestor called C, born 1645 (see Figure 11, green circle).

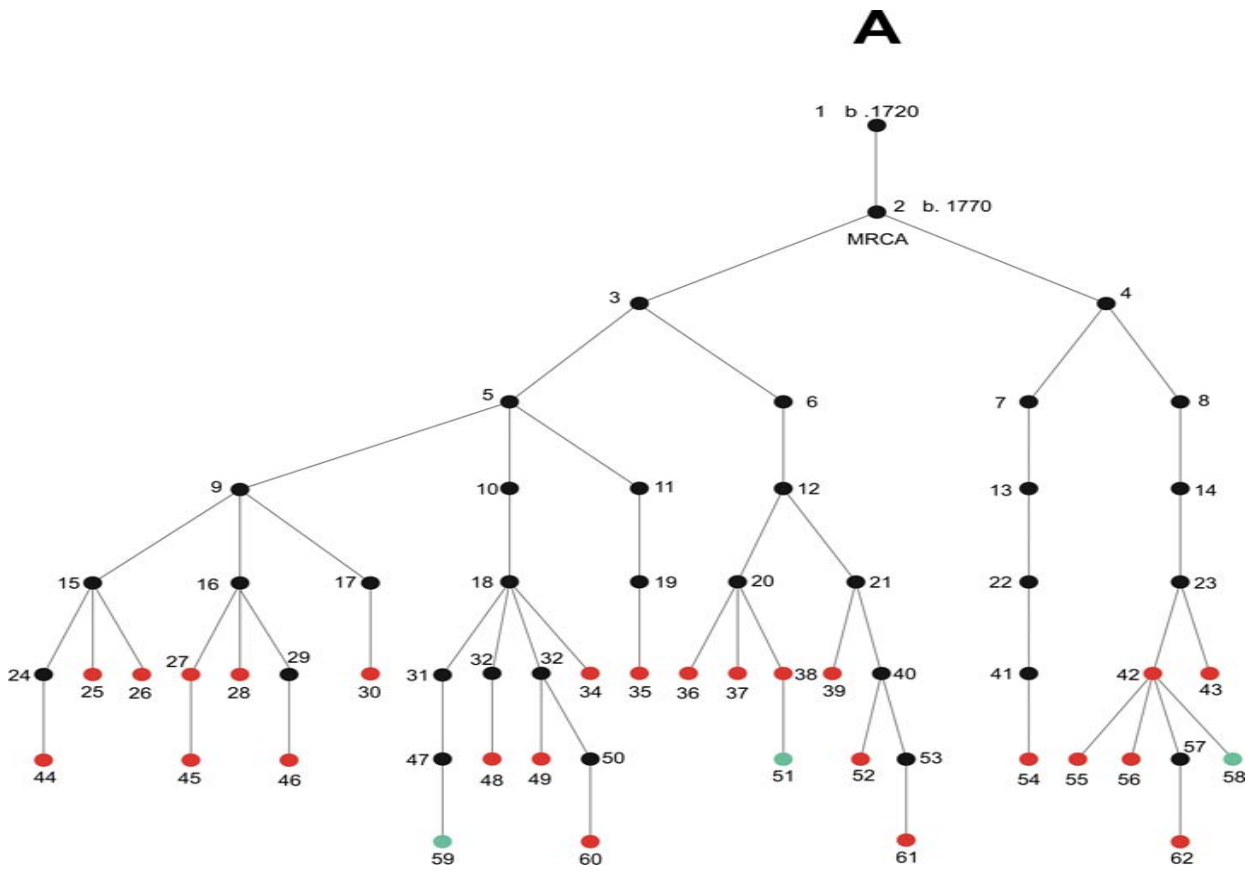

Figure 9. The matrilineal descendants of ancestor A, born 1720.

This genealogical tree shows the matrilines of descendants of ancestor A sequenced in this study (HVS1 and HVS2). The green circles represent individuals sequenced prior to this study. The red circles represent the 25 individuals sequenced for HVS1 to verify that the MRCA (A2) carried the haplogroup C1 lineage. The MRCA has two matrilines that are both are represented in this genealogical tree. Ancestor A has a total of 233 contemporary matrilineal descendants.

.

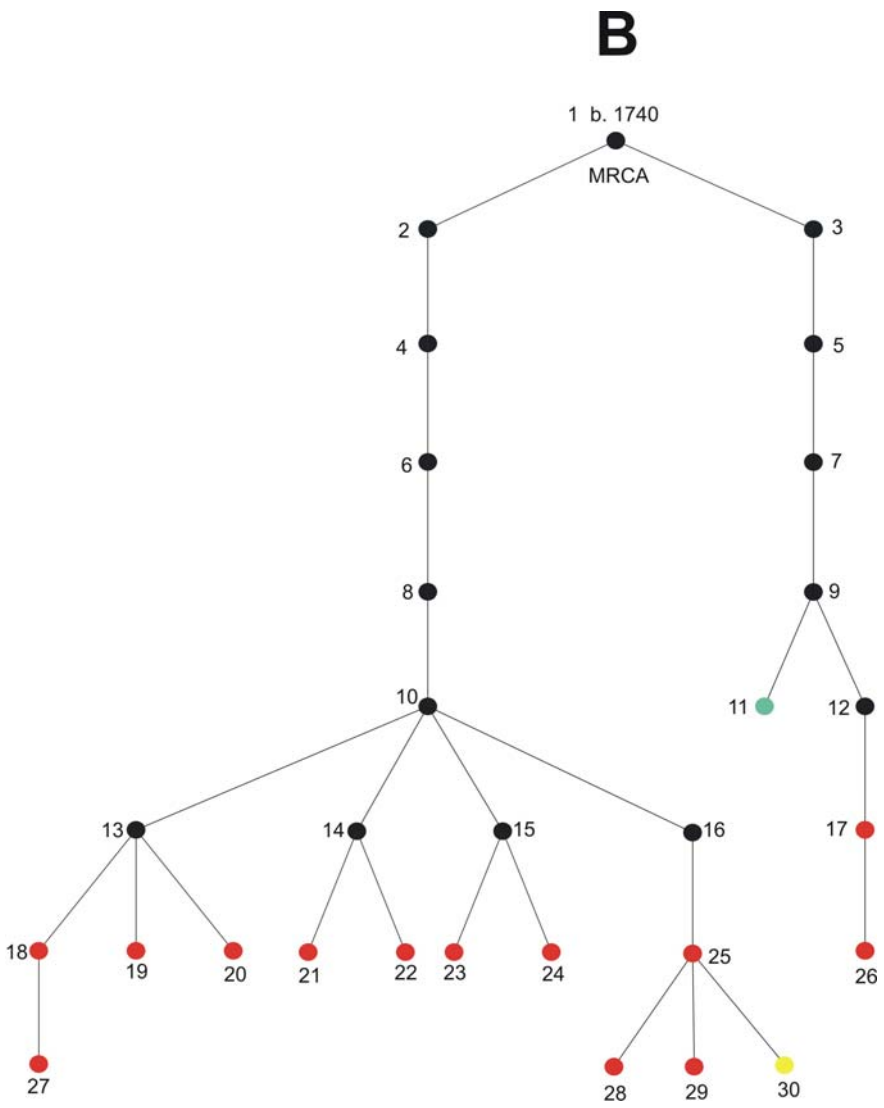

Figure 10. The matrilineal descendants of ancestor B, born 1740.

This tree shows the matrilines of descendants of ancestor B sequenced in this study (HVS1 and HVS2). The green circle represents the individual sequenced prior to this study. The yellow circle represents the individual that was sequenced in an attempt to find more haplogroup C1 carriers (see section 5.1.2). The red circles represent the 13 individuals sequenced to verify that ancestor B belongs to haplogroup C. The MRCA has two matrilines that are both are represented in this genealogical tree. Ancestor B has a total of 90 contemporary matrilineal descendants.

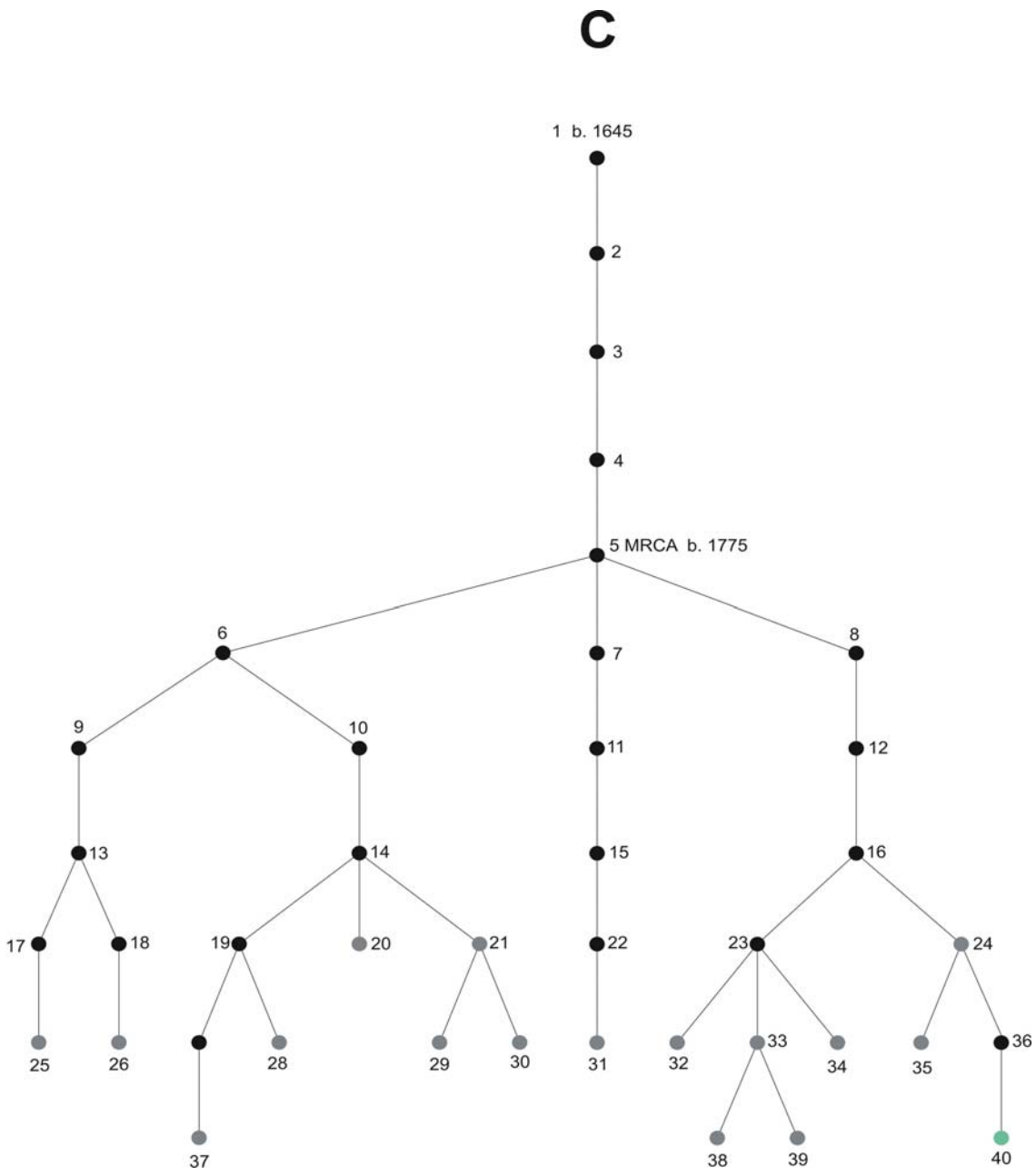

Figure 11. The matrilineal descendants of ancestor C, born 1645.

This tree shows the matrilines of descendants of ancestor A sequenced in this study (HVS1 and HVS2). The green circle represents the individual sequenced prior to this study. The grey circles represent the 16 individuals sequenced to conclude that the MRCA (C5) did not belong to haplogroup C1. The MRCA has three matrilines that are all are represented in this genealogical tree. Ancestor C has a total of 92 contemporary matrilineal descendants.

5.1.2 Detecting more individuals carrying haplogroup C

In an attempt to find more haplogroup C carriers, the original sample of 1538 control region sequences was extended. However, the additional samples were not selected randomly. The deCODE Genetics genealogical database was used to select primarily individuals that carried an unknown mtDNA sequence, i.e. individuals that could not matrilineally be linked to one of the 1538 previously sequenced individuals. The cohort used in this study consisted of contemporary Icelanders, defined as those born in the years 1895 to 2005, a total of 408,915 individuals. In the first step, the matrilines of the 1538 previously sequenced individuals were traced back in time to their earliest known matrilineal ancestors, a total of 792 ancestors. In the second step, matrilines were traced forward in time from those 792 ancestors to a total of 250,242 contemporary matrilineal descendants, born between 1895 and 2005 (see Figure 12, the light orange ellipse). Assuming that these individuals carry the same mtDNA sequence as their already sequenced matrilineal relative from the original sample of 1538 individuals, we sought to sample only from the remaining 158,673 individuals (born 1895-2005), who's mtDNA sequence was unknown. The matrilines of these 158,673 individuals were traced back in time to their earliest known matrilineal ancestors. An attempt was made to select at least one line of descent from as many independent ancestors as possible.

DNA samples were only available for a portion of these Icelanders at deCODE Genetics. As the DNA samples available for this study are stored on 96-well DNA plates (and a descision was taken not to create new plates for this study), a statistical approach was applied to select eight plates that maximized the number of matrilineal ancestors with an unknown mtDNA haplotype. The mtDNA HVS1 was sequenced for the samples of these plates, a total of 744 samples who are descendants of 536 independent ancestors. By tracing the matrilines of those 536 ancestors forward in time, 186,489 contemporary matrilineal descendants were identified (see Figure 12, the light blue ellipse).

There was some intersection between the two sets of ancestors and descendants. Nonetheless, this approach identified the mtDNA sequence of 304 additional matrilineal ancestors with previously unknown mtDNA sequences, from whom 46,466 matrilineal descendants could be traced.

Combined data sets consisted of mtDNA sequences from 2,275 individuals. Because of a slight overlapping between the two data sets, where the same individual came up in both samples, our resulting data set included 2,266 individuals which could be traced back to a total of 1,096 matrilineal ancestors. These ancestors, in turn left a total of 296,708 matrilineal descendants, which amounts to 72.56% of the contemporary Icelandic mtDNA pool.

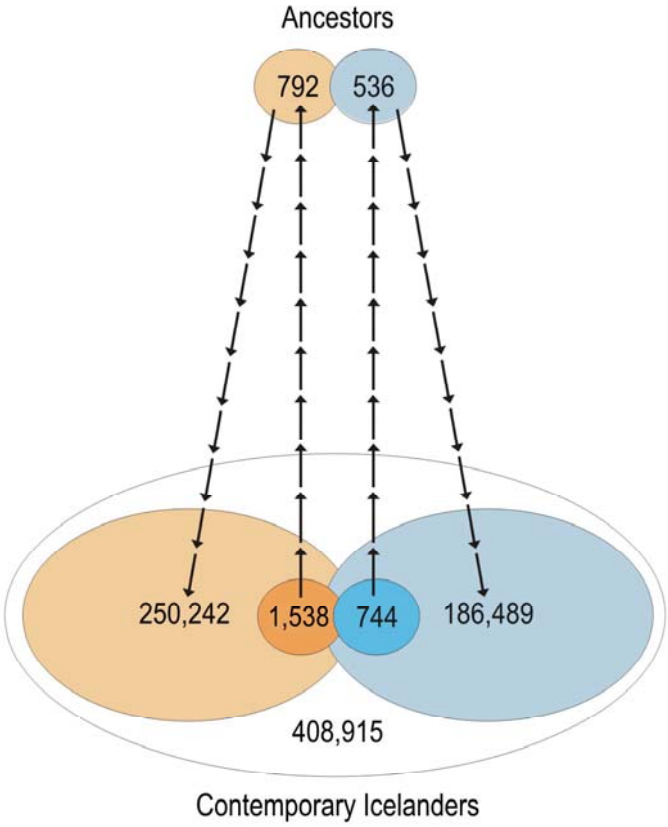

Figure 12. Matrilines traced back and forward in time

All Icelanders born in the years 1895 to 2005 are represented by the white ellipse. The dark orange ellipse represents individuals sequenced previous to this study. Based in these individuals the control region sequence of the other individuals in the light orange ellipse could be inferred using the genealogical database. The dark blue ellipse represents the individuals sequenced for this study. The light blue ellipse represents individuals whose sequence could be inferred using the genealogical database.

The new sample of 744 individuals sequenced for HVS1 (sites 16055-16410), contained three new haplogroup C1 carriers. The first individual could be traced back to a new matrilineal ancestor called D, born 1710 (see Figure 13, yellow circle). The second individual could be traced back to a new matrilineal ancestor

called E, born 1720 (see Figure 14, yellow circle). Finally, the third individual could be traced back to a previously known matrilineal ancestor, called B (see Figure 11, yellow circle). To sum it up, from the combined sample of 2,266 individuals, eight individuals were possible haplogroup C1 carriers and they could be traced back to five ancestors, called A, B, C, D and E.

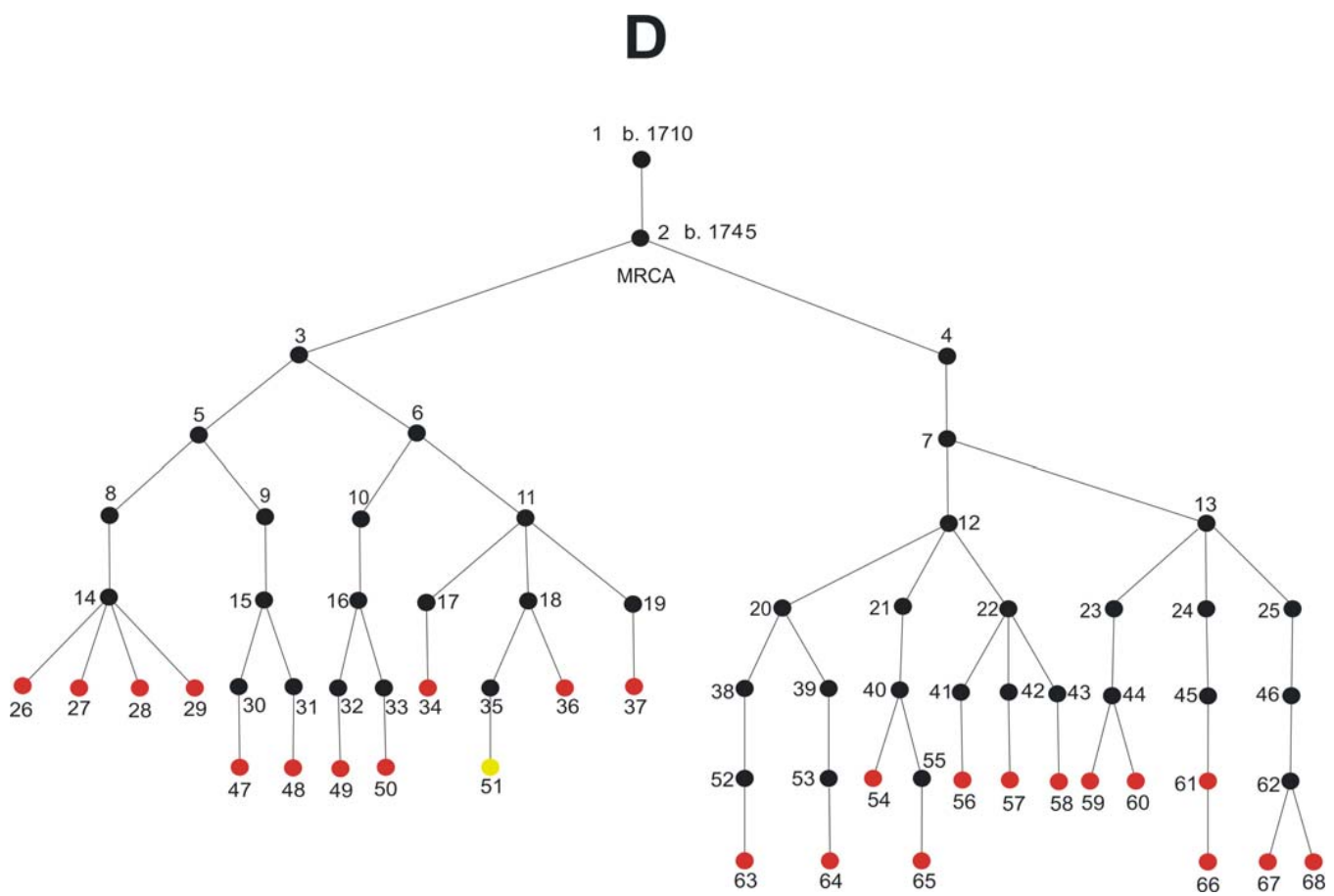

Figure 13. The matrilineal descendants of ancestor D, born 1710.

This tree shows the matrilines of descendants of ancestor D sequenced in this study (HVS1 and HVS2). The yellow circle represents the individual that was sequenced in an attempt to find more haplogroup C1 carriers. The red circles represent the 24 individuals sequenced to verify that the MRCA (D2) belongs to haplogroup C. The MRCA has two matrilines that are both are represented in this genealogical tree. Ancestor D has a total of 269 contemporary matrilineal descendants

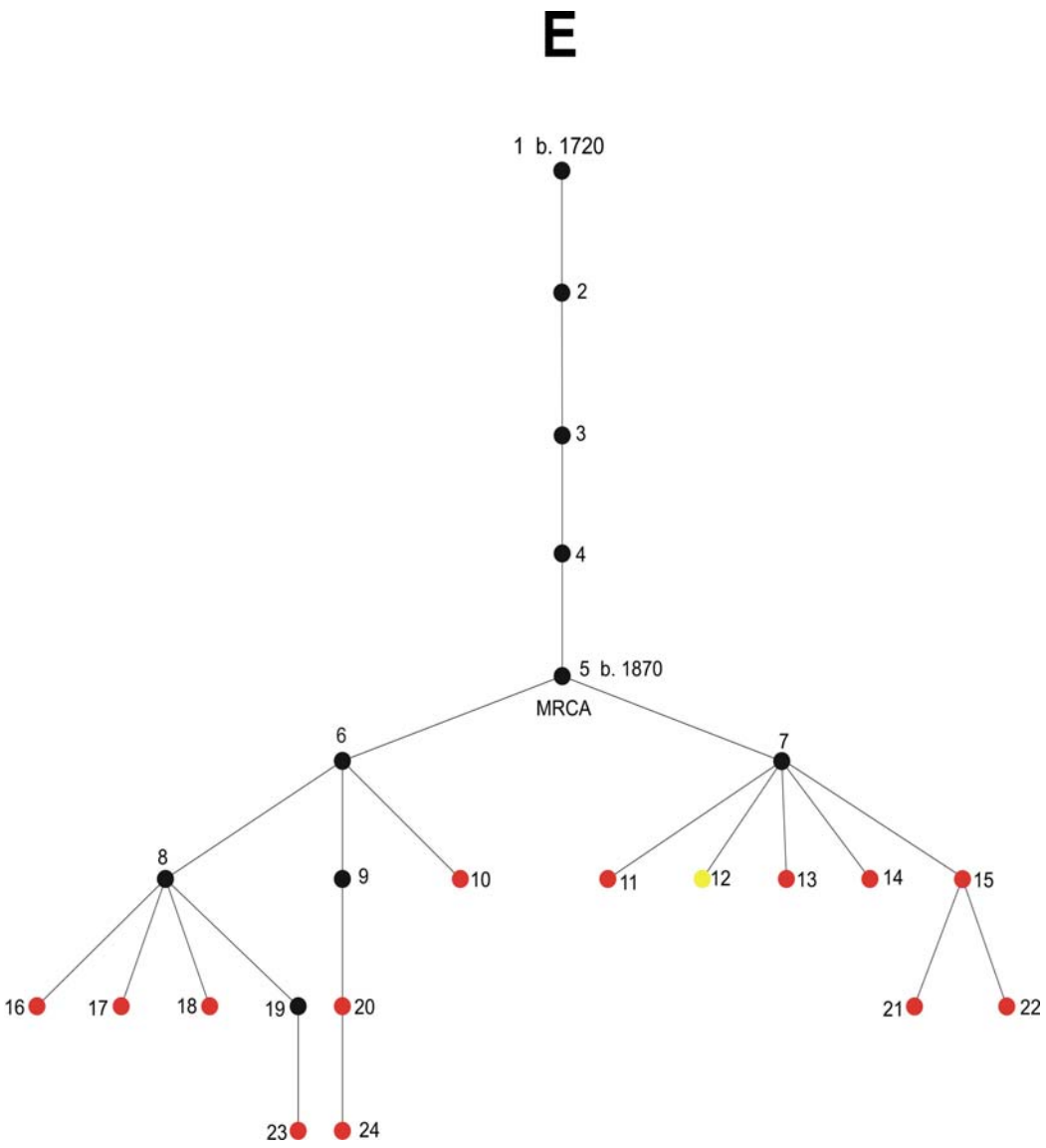

Figure 14. The matrilineal descendants of ancestor E, born 1720.

This tree shows the matrilines of descendants of ancestor E sequenced in this study (HVS1 and HVS2). The yellow circle represents the individual that was sequenced in an attempt to find more haplogroup C carriers. The red circles represent the 13 individuals sequenced to verify that the MRCA (E5) belongs to haplogroup C1. The MRCA has two matrilines that are both are represented in this genealogical tree. Ancestor E has a total of 79 contemporary matrilineal descendants.

5.1.3 Verification of haplogroup C carriers

Three of ancestor A´s and two of ancestor B´s descendants, had been sequenced at this point and confirmed to be haplogroup C1 carriers. Because more than one descendant had been identified that carried haplogroup C1 for both these ancestors it could be assumed that the ancestors also carried haplogroup C1. However, in the case of ancestors, C, D and E, the assumption of them being carriers of haplogroup C1 was based on only one descendant. To confirm that this assumption was not the result of an error either in the genealogical database or in sample handling, more descendants were sequenced. More descendants for ancestors A and B were also sequenced to verify that independent matrilines found within both lineages were in fact descendants of those ancestors and thus haplogroup C1 carriers. If all tested descendants of a particular ancestor (the MRCA) carry the haplogroup C1 sequence, then it is possible to infer that the ancestor carries it.

The matrilines of all five ancestors, A, B, C, D, and E were traced forward in time, to a total of 763 contemporary matrilineal descendants. The mtDNA control region (sites 15811-775) was sequenced for a subset of descendants, who represented multiple independent lines of descent from each ancestor. At least one individual from each line of descent from the MRCA was included in the sample. From a sample of 91 descendants, 75 individuals were confirmed to be haplogroup C1 carriers. They were all descendants of ancestors A, B, D and E, represented with red circles in Figures 10, 11, 13 and 14. As initially suspected, 16 descendants of ancestor C carried another mtDNA sequence, the same one as originally determined for the mother of the doubted carrier of haplogroup C1. These individuals are represented with gray circles in Figure 12. From these results, it was concluded that four ancestors. A, B, D and E were haplogroup C1 carriers, but not ancestor C. It can be concluded that the MRCAs of the four lineages carried haplogroup C1, thus we can say with certainty that the Icelandic haplogroup C1 lineages had a common ancestor prior to 1740 AD, which is the birth date of the oldest MRCA that is known. All the descendants that were tested and thought to be haplogroup C1 carriers based on information from the genealogical database turned out to be so, suggesting that the genealogical database is accurate. Furthermore, a study performed by Sigurðardóttir *et al.,* (2000) on the reliability of maternal links over

the period of time sampled in the database suggested that the information given in the database were 99,3% accurate (95% confidence interval [CI] 98.4%-99.8%). Thus, it is likely that virtually all maternal links obtained from the genealogical database are reliable, which would mean that the earliest known Icelandic haplogroup C1 carrier was born 1710 AD (the age of the earliest known Icelandic haplogroup C1 ancestor).

5.1.4 Geographical analysis of haplogroup C1 ancestors

Information about the geographical location of individuals is recorded in the deCODE Genetics genealogical database, constructed from censuses and parish registers. We thought that the observed geographical range of ancestors might tell us something about the age of the C1 lineage in Iceland or the number of founding copies. The results of a study performed by Helgason *et al*. (2005) of the geographic ancestry of Icelanders indicated that individuals tended to live in the same region as most of their ancestors five generations previously. This population structure was particularly evident in earlier times. Thus, it can be surmised that identity by descent of alleles had a strong geographical component in the past. Because of the geographical stratification of the Icelandic gene pool, combined with the fact that C1 is a rare lineage, we expected the ancestors of the C1 haplotype to be concentrated in a limited geographical range. The earliest known ancestors and the MRCAs of lineages A and E were born in Rangárvallarsýsla and the same applies for the MRCA of lineage D (the location of the earliest ancestor ,D1, is not known). The earliest known ancestor of lineage B (who is also the MRCA) was born in an adjacent county called V-Skaftafellssýsla.

The limited geographic spread of the four lineages supports the notion of a single founding ancestor, born in Iceland prior to 1710 AD (the age of the earliest known Icelandic haplogroup C1 ancestor) and located in Rangárvallarsýsla. It cannot be ruled out that all four ancestors, A, B, D and E, were daughters of a single a woman born sometimes around the end of the 17th century. However, as the MRCA of the four lineages are not all located in the same county and given the migration rate of the period, it is likely that a few generations have passed since the lineages shared a common ancestor. Furthermore, the result of a study performed

by Helgason *et al.*, (2003a) on the recent evolutionary history of mtDNA in the Icelandic gene pool revealed a high rate of genetic drift among matrilines. Since only a small portion of all possible matrilineal ancestors have contributed their mtDNA to the contemporary Icelandic population it is unlikely that four sisters born almost 300 years ago would all be among them. The more likely scenario is that the MRCA of the four lineages was born at least a several generations prior to 1710 AD contributing her mtDNA to a number of lineages of which only four were transmitted successfully to present day.

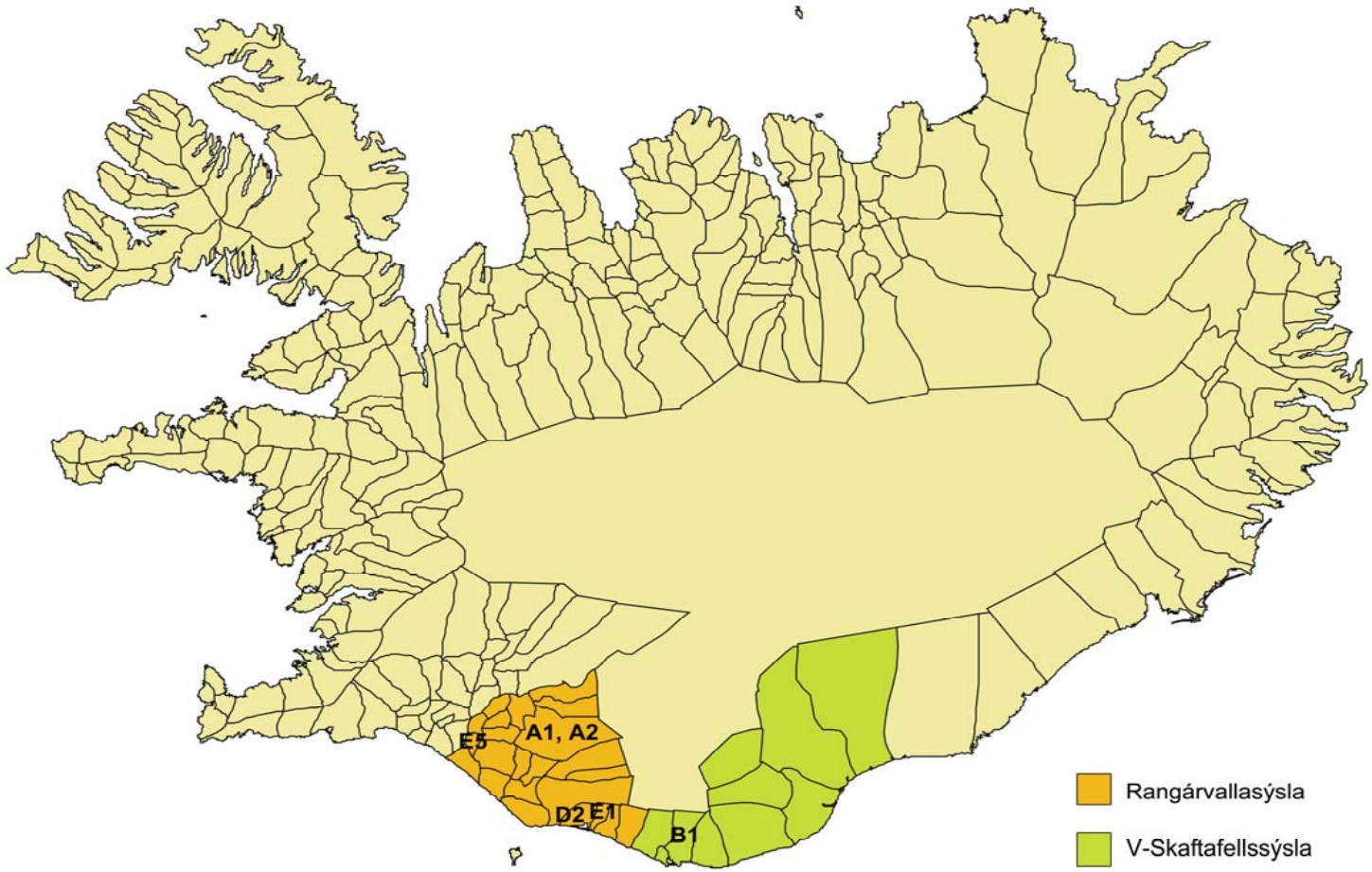

Figure 15. Geographical distribution of the earliest known ancestors and MRCAs of Icelandic haplogroup C1 carriers, showing both county and parish location.

Iceland is divided into 21 counties and 295 parishes (the lines on the map identify each parish). The earliest known ancestors and MRCAs of haplogroup C1 carriers belonged to two adjacent counties Rangárvallarsýsla (identified by orange shading), and V-Skaftafellssýsla (identified by green shading). Capital letters refer to the lineages, A, B, D, and E and the numbers refer to the number given to the earliest known ancestors and MRCAs in Figures 10, 11, 13 and 14. We did not have information about the location of the earliest known ancestor of lineage D. Note that B1 is both the earliest known ancestor and the MRCA of lineage B.

5.2 Analysis of complete mtDNA sequences for haplogroup C1

5.2.1 The complete mtDNA sequence of Icelandic haplogroup C1 carriers

In the hope of locating the Icelandic sequence in the phylogeny of haplogroup C1 sequences at the highest level of molecular resolution the complete mtDNA for several Icelandic haplogroup C1 carriers were sequenced. Complete mtDNA sequences were generated for eleven Icelandic haplogroup C1 carriers. An attempt was made to select at least one independent line of descent from the common the MRCA for each of the verified Icelandic haplogroup C1 ancestors, A, B, D and E. This was done in the hope of finding mutations that had occurred within the Icelandic C1 haplogroup. This could potentially give us more information about the founding age of the Icelandic sequence.

The complete Icelandic C1 sequence has a total of 49 mutations that distinguish it from the rCRS (Andrews et al., 1999) (see Table 4). The complete sequences of the 11 Icelandic haplogroup C1 carriers were almost all identical. They shared eleven specific mutations that distinguished them from other known C1 mtDNA sequences, that is 152C, 534T, 3395G, 3507T, 5895-5899i(5C-12C), 7331T, 13651G, 13966G, 14324C, 15613G and 16311C.

Table 4. List of sites and nucleotide changes in the Icelandic haplogroup C1 sequence compared to the rCRS.

Site 1-5000	Nucleotide change	Site 5001-13900	Nucleotide change	Site 13901-16569	Nucleotide change
73	T→C	5895-5899	5-12C insertion	13966	A→G
152	T→C	7028	C→T	14318	T→C
249	A deletion	7196	T→A	14324	T→C
263	A→G	7331	C→T	14766	C→T
290	A deletion	8584	G→A	14783	T→C
291	A deletion	8701	A→G	15043	G→A
315	C insertion	8860	A→G	15301	G→A
489	T→C	9540	T→C	15326	A→G
534	C→T	9545	A→G	15487	A→T
750	A→G	10398	A→G	15613	A→G
1438	A→G	10400	C→T	16223	C→T
2706	A→G	10873	T→C	16298	T→C
3395	A→G	11719	G→A	16311	T→C
3507	C→T	11914	G→A	16325	T→C
3552	T→A	12705	C→T	16327	C→T
4715	A→G	13263	A→G		
4769	A→G	13651	A→G		

The complete mtDNA genome was sequenced for three matrilineal descendants of ancestors A, A60, A61 and A62, and three descendants of ancestor B descendants, B26, B27 and B29 (see Figures 16 and 17). They all shared the same eleven mutations, but individual B26, descendant of ancestor B, carried an additional transition at site 12561. In order to verify this mutation, two individuals, B11 and B17, belonging to the same line of descent as individual B26, were sequenced for a fragment in the coding region, sites 11948-12772 (see Figure 17). They also carried the 12561A mutation and thus it could be inferred that individuals B9 and B12 must also have carried the mutation. However, it can not be established whether the mutation arose in individual B3, B5, B7 or B9. The two other descendants of ancestor B descendants, B27 and B29, did not carry that mutation, therefore it can be assumed that the MRCA (B1) did not carry it. The complete mtDNA sequence was also sequenced for three descendants of ancestor D (D48, D60 and 65) and two descendants of ancestor E (E21 and E24) (see Figures 18 and 19). They all shared the same eleven mutations as the descendants of ancestor A and B. In addition, the descendants of ancestor E carried an extra transition (A→G) at site 13537. To confirm this mutation, six more descendants of ancestor E (E11, E16, E18, E20, E22 and E23) were sequenced for the fragment in the coding region containing this mutation (sites 13338-14268). The result showed they all carried the 13567G mutation. From this it can be inferred that the MRCA of lineage E (E5) also carried it. However, it can not be established whether it occurred in individual E4, E3, E2, E1 or some earlier matrilineal ancestor (see Figure 19).

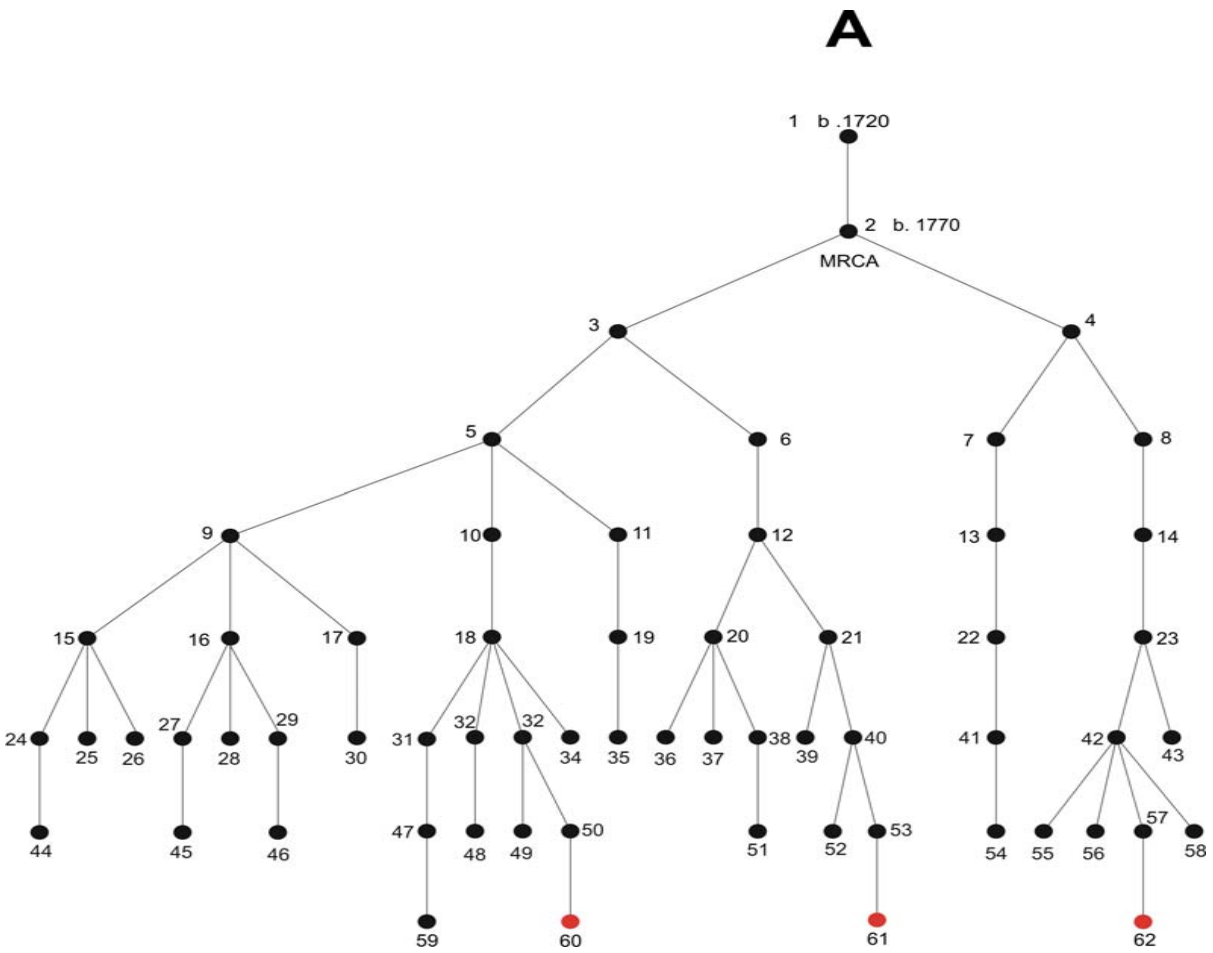

Figure 16. The matrilineal descendants of ancestor A, born 1720. The complete mtDNA was sequenced for three descendants of ancestor A, represented with red circles.

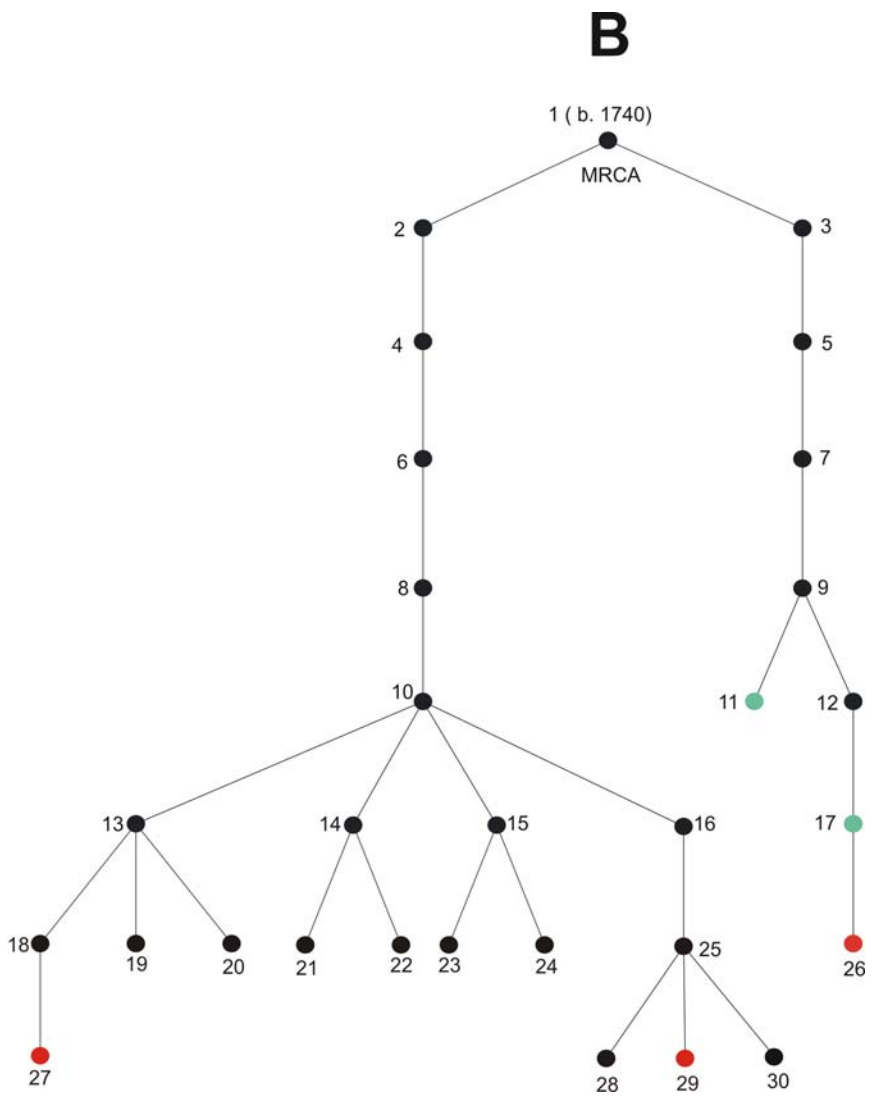

Figure 17. The matrilineal descendants of ancestor B, born 1740.

The complete mtDNA was sequenced for two descendants of ancestor B, represented by red circles. To confirm the 12561A mutation that individual B26 carried, a fragment of the mitochondrial genome, that spans sites 11,948-12,772 of the coding region, was sequenced, for two individuals, represented with green circles.

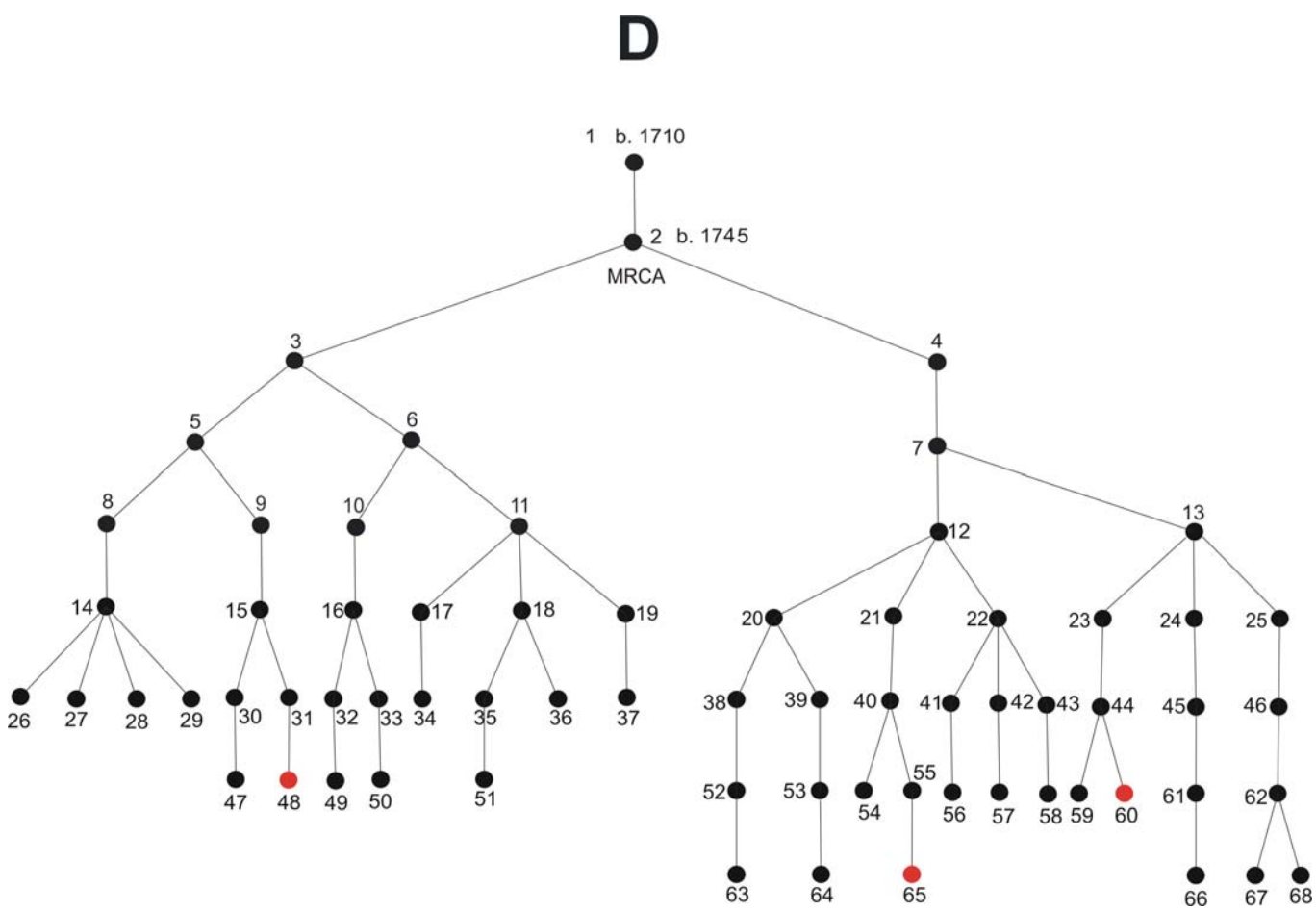

Figure 18. The matrilineal descendants of ancestor D, born 1710. The complete mtDNA was sequenced for three descendants of ancestor D, represented with red circles.

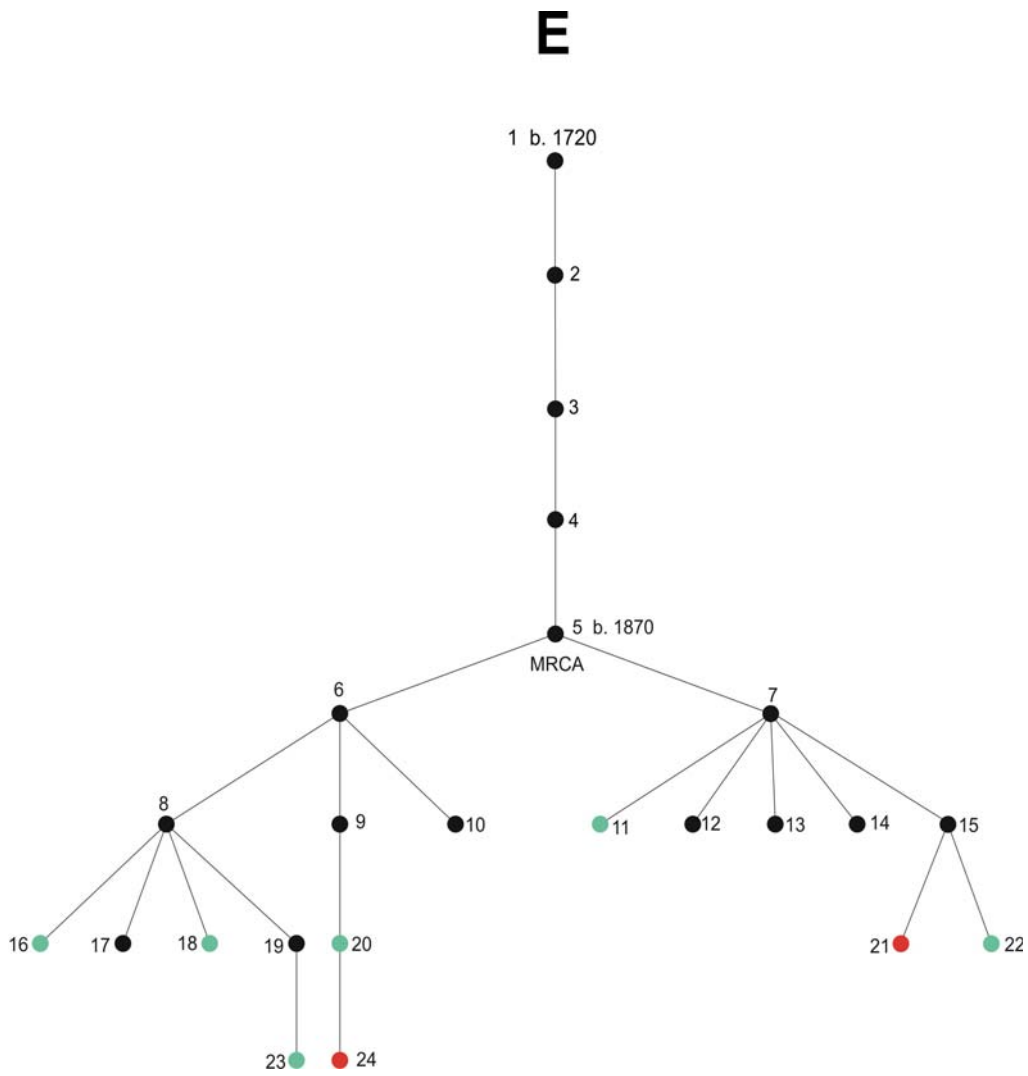

Figure 19. The matrilineal descendants of ancestor E, born 1720.

The complete mtDNA was sequenced for two descendants of ancestor E, represented with red circles. To confirm the 13537G mutation that individuals E21and E24 carried. A fragment of the mitochondrial genome, that spans sites 11,948-12,772 of the coding region, was sequenced for six individuals, represented with green circles.

5.2.2 The founding age of Icelandic haplogroup C1

We know that the mutation, 12561A, found in some descendants of ancestor B, arose sometime between the year 1770 (descendant B3) and 1870 (B9). Thus, it cannot be used to help infer the time of entry for the C1 haplotype into the Icelandic mtDNA pool. In contrast, the transition (A→G) at site 13567 found in all matrilineal descendants of ancestor E can, al least in principle, help estimate the age of the C1 lineage in the Icelandic mtDNA pool. We do know that it arose sometime before the year 1870, which is the date of birth of the MRCA, E5, or in the MRCA herself. It has been estimated that mutations occur in the coding region at the rate of one every 3533 years (Soares *et al.*, 2009). According to this mutation rate, we would expect one mutation every 883 years in the matrilines leading to the four haplogroup C1 MRCAs. The average birth date of the four MRCAs is 1781 AD. To simplify calculations we assumed that only one coalescence event had occurred for all the four lineages i.e. that the matrilines are independent and that they all coalesce at the same time to a single ancestor. This maximizes the number of generations between the MRCAs for the four lineages and their MRCA. According to this assumption our best estimate for the birth date of the MRCA of all four Icelandic C1 lineages is 898 AD (see Figure 20). Interestingly, this falls exactly within the period of the settlement of Iceland, dated 874-930 AD (Smith, 1995). In contrast, if we assume that the ancestors of the three lineages that do not carry the mutation were sisters, then the total number of years that have passed since the birth of the MRCA for the four lineages is minimized. When the number of generation is minimized based on these assumption the probability of observing one coding region mutation is highest if the MRCA was born 38 AD (see Figure 20). Another possibility is that the ancestors of the four lineages were all the daughters of the same woman, but then the 13567G mutation would have had to occur in the meiosis leading to ancestor E. However, it is important to realize that there is very little power to reliably estimate the founding age of the C1 lineage based on a single mutation. This is merely our best guess based on the information that is available. The limited power of this approach is highlighted by the 12561A mutation in lineage which we know is not more than 240 years old. However, if we

did not have information about ancestor B1, we would overestimate the age of the mutation by almost 1000 years. Nonetheless, this is our best estimate.

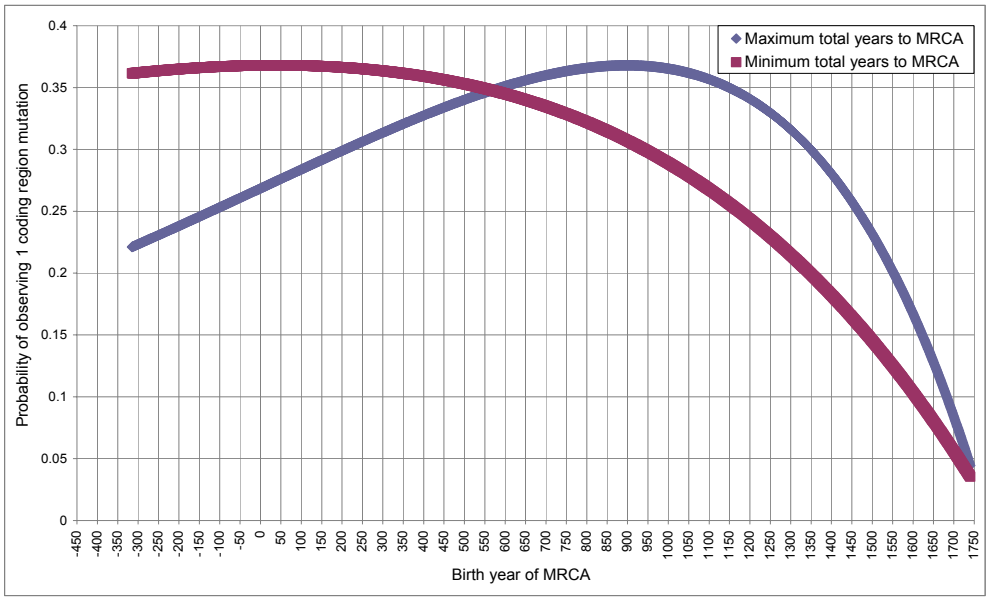

Figure 20. This graph shows the probability of observing 1 coding region mutation relative to possible birth year of the MRCA for the four Icelandic C1 lineages. The year 1750 is entered as the first possible birth date of the MRCA. This date is based on the earliest MRCA for the four lineages.

5.2.3 Heteroplasmy within the Icelandic C1 lineage

Almost all mtDNA sequences include a homopolymeric tract of five cytosine (C) bases at sites 5895-5899. The Icelandic C1 lineage has an insertion of Cs at this site that interfered with the direct PCR sequencing reaction after this C tract (regardless of which strand was sequenced). This problem is due to a heteroplasmic mixture of mtDNAs, observed in direct sequencing, each containing C tracts of different lengths (see Figure 21). To obtain better sequence data for the region beyond the C tract and furthermore to determine how many Cs the Icelandic haplogroup C1 sequence has and to help determine the number of Cs inserted in the original mutation event, the segment was amplified by PCR and then cloned and sequenced. Sequencing of the clones revealed variation in the length of the C tract between and within individuals. Since somatic cells contain hundreds or thousands of copies of mtDNA, but mutations occur in single copies of mtDNA, individuals may be heteroplasmic for particular mtDNA mutations and this variation within individuals can be transmitted for multiple generations. In such cases, the mtDNAs that carry the original alleles at the mutating site are said to carry the wild type (Jobling *et al.*, 2004; Strachan & Read, 2004), which in the case of the 5895-5899 C tract is 5Cs.

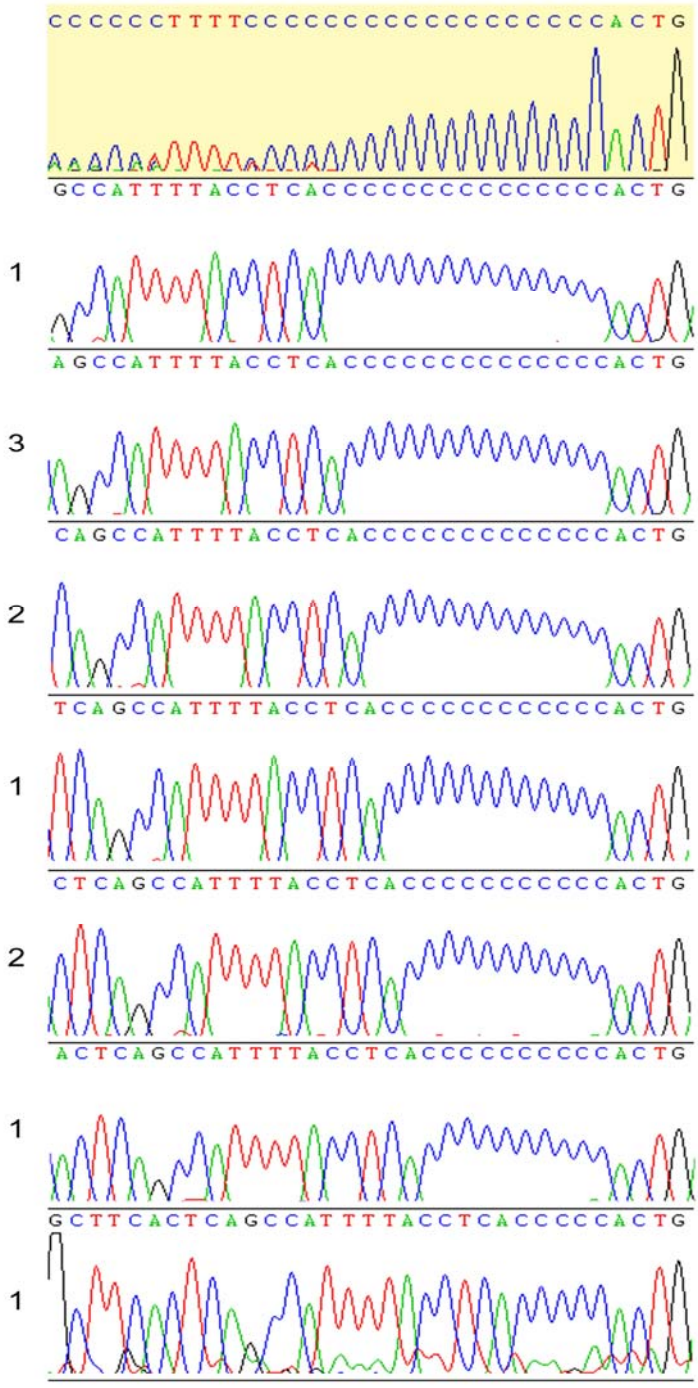

Figure 21. Length variation at sites 5895-5899 in a single individual.

The length variation shown was observed within individual D48 (see Figure 18). The first chromatogram, indicated with yellow background, was obtained through direct sequencing and thus the sequence that comes after the C tract is no good (it is shown in reverse and complemented and should thus be read from right to the left). The other chromatograms show cloned sequences from the same individual. The polymorphic C tract ranges in length from 15 to 5 nucleotides. The numbers shown to the left of the chromatograms indicate the number of clones with that particular C tract length from individual D48.

A total of 41 clones were sequenced from seven individuals that carry the Icelandic C1 haplotype. Three samples were from descendants of ancestor A, one sample from a descendant of ancestor B, two samples from descendants of ancestor D and one sample from a descendant of ancestor E. When the length variation within the heteroplasmic samples was analyzed it emerged that the homopolymeric tracts were predominantly 11, 12, 13, or 14 nucleotides long, but cloning also revealed lengths as short as 5 nucleotides (the wild type) and as long as 17 nucleotides in some cases (see Table 5). Unfortunately, the number of clones sequenced is too small to provide conclusive results about the likely length of the original insertion event, which must have occurred within a matrilineal ancestor of ancestor A, B, D and E. However, Table 5 shows that the modal homopolymeric tract length across the seven samples is 12 and this is also the length present in the greatest number of individuals. The tract homopolymeric length 12 is therefore represents our best guess for the length of the insertion mutation. Replication slippage in the polycytosine tracts occurs more readily when a specific number of cytosine residues are present. Studies indicate that of the number of cytosine goes above 8Cs for the HVS2 and above 7Cs for HVS2 slippage occurs frequently (Irwin *et al.*, 2009). The slippage is rare when the string of Cs is below this threshold. This is consistent with our result.

Table 5 The C tract length polymorphism at site 5895-5899.

The length of the homopolymeric tracts of cytosine identified are shown as Cx, were x is the length of the homopolymeric (C) tract.

Individual	C5	C10	C11	C12	C13	C14	C15	C17	*Total*
A60				2			1		*3*
A61			2		3				*5*
A62			1	1	1	2	1	1	*7*
B26		1		2		1	1		*5*
D48	1	1	2	1	2	3	1		*11*
D65		1	1	2		1	1	1	*7*
E21	1			1	1				*3*
Total	*2*	*3*	*6*	*9*	*6*	*7*	*5*	*2*	

The so-called mtDNA bottleneck hypothesis was formulated on the basis of a study performed by Hauswirth & Laipis (1982), according to which the number of mtDNA molecules is reduced to a few at some stages of oogenesis. Every

generation the population of mitochondrial genomes will pass through a bottleneck when the oocyte segregates a few of its mother's genomes. The segregation rate of mtDNA mutations is influenced by the size of the bottleneck. If the bottleneck is large the segregation will be slow. The narrower the bottleneck the faster the segregation occurs i.e. new mutation is expected to become fixed after a few generations (Lutz *et al.*, 1999). Most reports of the rate at which a lineage becomes fixed for a mutation give evidence for rapid segregation events as the result of a relatively small mitochondrial bottleneck size. According to Bendall et al., (1997) the most likely size of the mtDNA bottleneck is 1-27 segregating units and Marchindon *et al.,* (1997) have estimated that the bottleneck size is likely to be one to five segregating units. Simulations have been used to determine the number of generations for which heteroplasmy is likely to be retained. According to a study performed by Bendall *et al.*, (1997) the average number of generations it takes for a mutation to become fixed with bottleneck size of 3 segregation units is 1.4 generations, but when the when the bottleneck is 20 segregation units it takes about 12.7 generations for a mutation to become fixed in a lineage.

Mutations in mitochondrial DNA arise in a single molecule, and in time they either become fixed in the lineages of matrilineal descendants or they are lost. It is of interest to note that the wild-type length of 5C is observed in two individuals who also carry the longer variants, descendants of ancestors D and E. This suggests that the C insertion at site 5895-5899 may be relatively recent and therefore the MRCA of the C1 lineages in Iceland. We know that 7C insertion occurred before the year 1700, because that is the first possible birth date for the matrilineal ancestor of ancestors A, B, D and E. Thus, it can be concluded that the 7C insertion is at least 300 years old. Furthermore, since the wild type length is still observed in at least two of the four Icelandic C lineages it can be concluded that the wild type and mutational variants have maintained for at least 7 generations (see Figure 18). This is substantially longer that indicated by previous research on heteroplasmy (Bendall *et al.*, 1997; Howell *et al.*, 2003).

These results raise an interesting question that we did not anticipate at the beginning of this study. The fact that we have access to genealogical information for the past ten centuries in Iceland gives us a certain advance in studying the

mechanism of heteroplasmy. Thus, we are going to follow up on this finding by doing more cloning on the individual sequences in this study and by cloning additional haplogroup C1 carriers. Perhaps that will throw light on the rate at which population becomes fixed for a new allele and the size of the meiotic bottleneck affecting that process.

5.2.4 The phylogenetic context of the Icelandic C1 sequence.

A detailed analysis of previously published data was performed to infer the location of the Icelandic C1 sequence in the phylogeny of haplogroup C1 at the highest level of molecular resolution – that of complete mtDNA sequences. By analyzing previous information scattered throughout the literature a total of 61 complete haplogroup C1 sequences were assembled, 57 Native American and four Asian. These sequences were manually constructed into a phylogenetic tree along with three Icelandic C1 sequences (see Figure 22).

The four Asian complete mtDNA C1 sequences belong to sub-clade C1a, an Asian-specific branch of C1. The Japanese (Tanaka *et al.*, 2004) sequence along with the sequences of Buryat (Derenko *et al.*, 2007), Ulchi (Starikovskaya *et al.*, 2005), and Nanai (Ingman & Gyllensten, 2007) form a separate cluster, defined by transitions at sites 3826 and 7598 in the coding region and 16356 in the control region (see Figure 22).

All 57 Native American complete mtDNA C1 sequences fall into one of the three previously identified Native American sub-clades C1b, C1c and C1d (see Figure 22) (Achilli et al., 2008; Fagundes et al., 2008b; Maca-Meyer *et al.*, 2001; Mishmar *et al.*, 2003; Tamm et al., 2007). The Native-American sub-clade C1b is defined by transitions at sites 493 and CA deletion at site 522-523 in the control region, which is prone to recurrence. The Native-American sub-clade C1c is distinguished by two coding region mutations a sites 1888 and 15930. Sequences that belong to the Native American sub-clade C1d are defined by a transition at site 16051 and CA deletion at site 522-523 in the control region, and a mutation at site 7697 in the coding region. Sub-clades C1c and C1b are distributed throughout the Americas, but if sequences with unknown Hispanic origin are put aside, sub-clade C1d is restricted to South-American populations. However, the amount of data is

limited, so it does not provide conclusive information about the geographic distribution of the Native American sub-clades.

Interestingly, the Icelandic sequence has none of the mutations that characterize known C1 sub-clades. In fact, the phylogenetic tree, shown in Figure 22, reveals that the Icelandic sequences belong to a distinct branch or sub-clade of haplogroup C1. In all, eleven mutations differ between the Icelandic C1 sequence and the others sequences from haplogroup C1. These mutations are 152C, 534T, 3395G, 3507T, 5895i(5C-17C), 7331T, 13651G, 13966G, 14324C, 15613G and 16311C, in addition to a reversion at the fast evolving site 16519. In light of this, it seems appropriate to assign the Icelandic sequence to a distinct sub-clade, named C1e, following existing mtDNA haplogroup nomenclature (see Figure 22) (Kivisild *et al.*, 2006; Torroni *et al.*, 2006). The number of complete mtDNA sequences from haplogroup C1 is still limited, thus it is likely that further sampling will reveal haplotypes that will fall into the same branch as the Icelandic sequence.

In addition to the 61 complete mtDNA sequences belonging to haplogroup C1, 13 Native American coding region sequences belonging to C1 have been published. Another phylogenetic tree was constructed that included all 74 sequences available, both complete and coding region sequences, to see if that would give us additional information about the phylogeny of the Icelandic haplogroup C1 sequence (see Figure 23). The thirteen Native American coding region sequences were assigned to sub-clades based in their coding region motifs. Nine of the coding region sequences are 15,446 bases long, starting at site 577 and ending at site 16023 (Herrnstadt et al., 2002) and three are 15,585 bases long, starting at site 436 and ending at site 16021 (Kivisild *et al.*, 2006). Twelve of theses coding region sequences could be assigned to one of the Native American sub-clades C1b, C1c or C1d, based on their coding region motifs (Herrnstadt et al., 2002; Just et al., 2008). Note that according to this phylogenetic tree (Figure 23) not all sequences that belong to the C1d sub-clade carry the CA deletion at site 522-523, thus, it can be concluded that sub-clade C1d is in fact not defined by that deletion as we would have concluded if we would not have included the coding region sequences in the phylogenetic tree, as seen in Figure 22. One of the thirteen coding region sequences, 174 Herrnstadt (Herrnstadt et al., 2002), does not have

any branch-specific mutations or a mutation that occur within one of the sub-clades that can be used to locate it within known C1 sub-clades. It does not belong to sub-clades C1a or C1c and it does not have any of the "Icelandic-specific" mutations. The 174 Herrnstadt sequence could either be a distinct lineage or it could belong to sub-clade C1b. Sub-clade C1b is only distinguished with mutations in the control region, and therefore we can not say whether the 174 Herrnstadt sequence belongs in that sub-clade based solely on the coding region motif. Although, it seems likely based on the fact that all other known Native American sequences fall into one of three Native American C1 branches. None of the sequences that were included in this phylogenetic analysis had any of the mutations that separate the Icelandic C1e sub-clade from the others C1 sub-clades. This further supports the conclusions that the Icelandic lineage belongs to a distinct sub-clade.

A third phylogenetic analysis was performed using only the coding region (sites 577-16021) of all 74 previously published sequences plus the three Icelandic sequences (see Figure 24). The coding region tree shows nineteen distinct lineages that branch off from the C1 root. The coding data alone provides sufficient information to define three of the four previously indentified branches of haplogroup C1, i.e. C1a, C1c, C1d, along with the C1e branch, because these branches are defined by mutations that occur within the coding region. On the other hand, the C1b branch is defined by control region mutations and therefore it can not be distinguished on the bases of coding region mutations. It can be assumed that all sequences that do not fall into one of the branches that are defined by coding region mutations belong to C1b. However, that would not be possible of we did not have the phylogenetic trees from complete sequences to compare it to. Until more sequences are found that belong to the C1e branch we must assume that all twelve mutations that the Icelandic sequences have in common and distinguish them from other C1 sequences are characteristic of the C1e branch as shown in Figure 24.

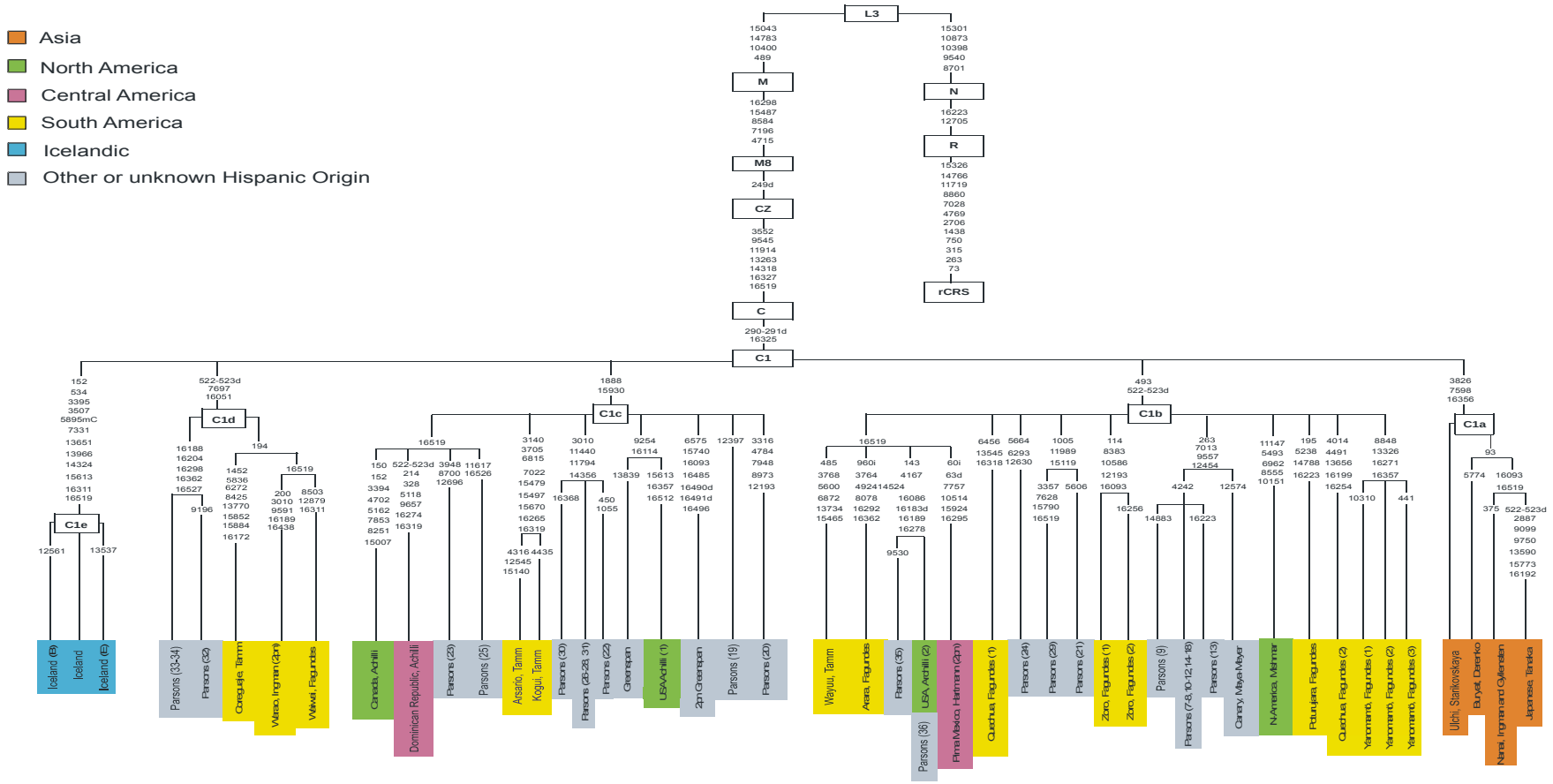

Figure 22. The phylogeny of haplogroup C1, complete sequences.

This schematic phylogenetic tree shows a total of 64 complete sequences (47 haplotypes). The phylogeny is rotted from the rCRS (Anderson *et al.*, 1981). Sub-haplogroup labels follow existing classifications. Deletions are indicated by a "d" and insertions are indicated by an "i". In all, 57 complete sequences fall into the Native American sub-clades C1d, C1c and C1b origin. The four Asian complete C1 sequences belong to sub-clade C1a, an Asian specific branch of C1. The three Icelandic sequences belong to a sub-clade named C1e. Note that mC corresponds to multiple Cs (the length variation at site 5895-5899) (see Section 5.2.3).

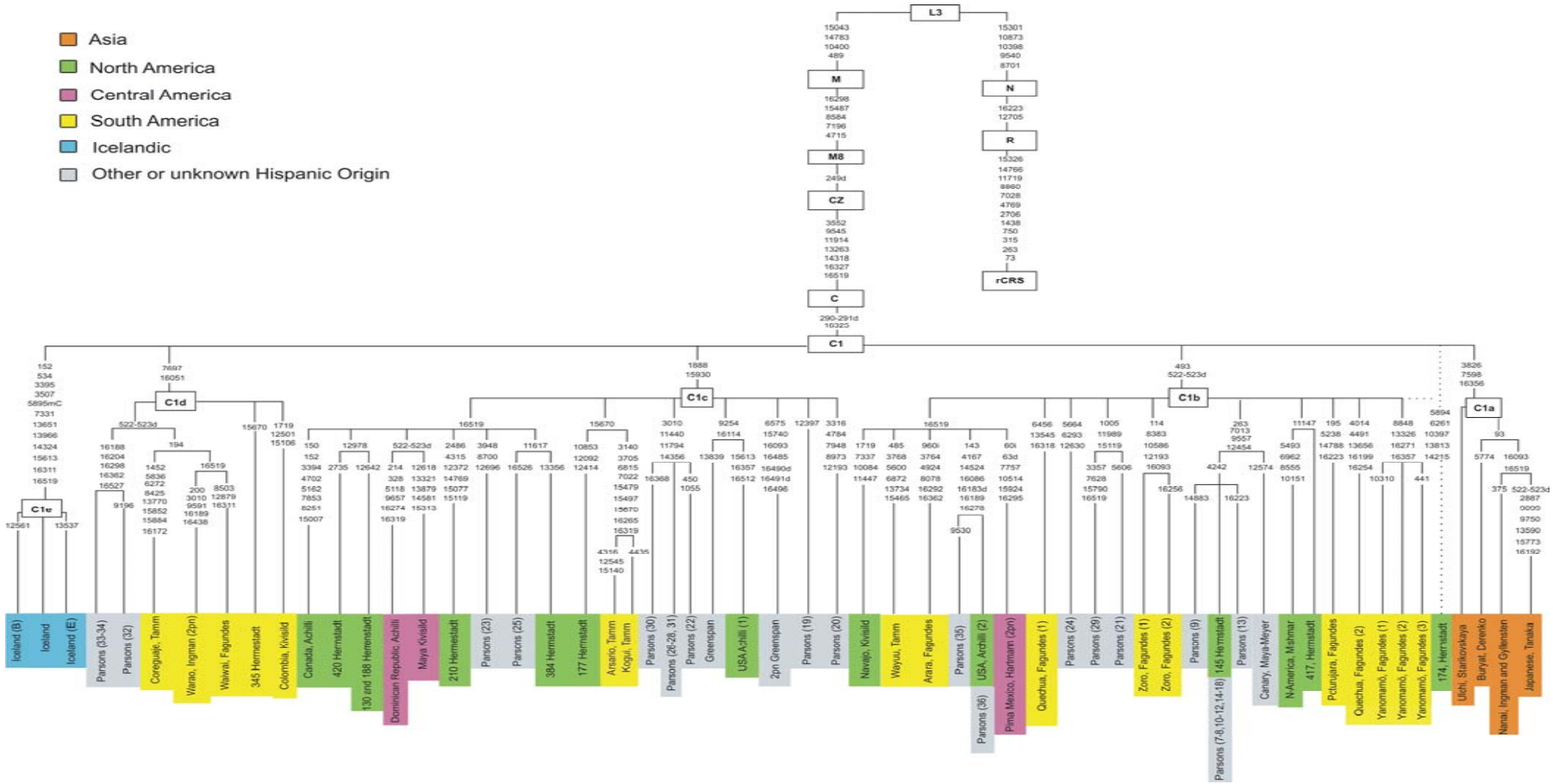

Figure 23. The phylogeny of haplogroup C1, complete sequences and coding region sequences.

The phylogeny of all 77 sequences (58 haplotypes) analyzed in this study is shown here, including complete sequences (64), shown in Figure 14, and coding region sequences (13). All coding region sequences except fore one sequence, 174 from Herrnstadt et al., (2002), could be assigned to one of the three Native American sub-clades. Note that mC corresponds to multiple Cs (the length variation at site 5895-5899) (see Section 5.2.3).

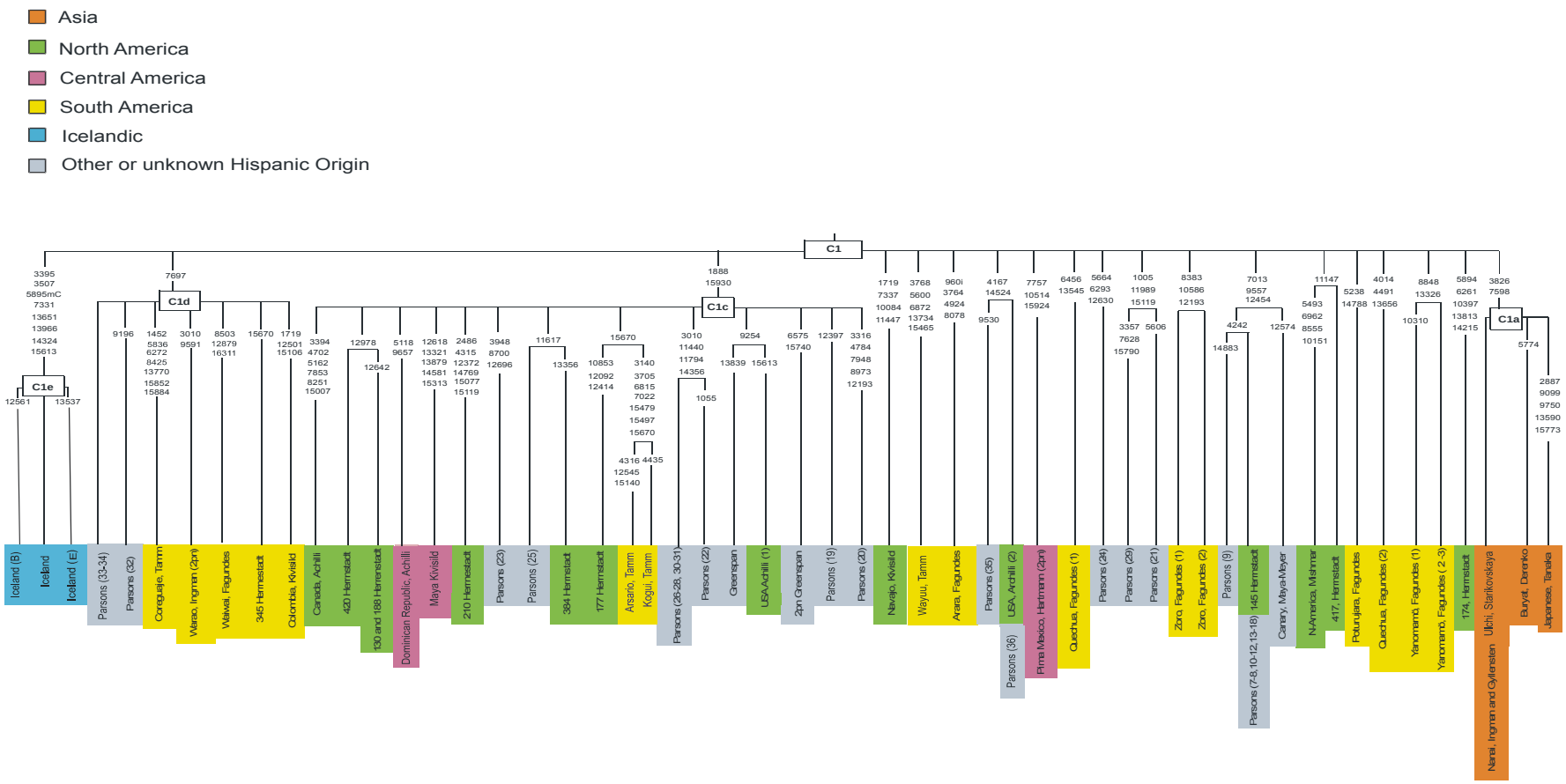

Figure 24. The phylogeny of haplogroup C1, coding region.

The phylogeny tree shows all 77 sequences (58 haplotypes) analyzed in study. It is identical to the one shown in Figure 15 except this tree is constructed from coding region motifs. The tree shows how the phylogeny of haplogroup C1 changes when only the coding region is used. In all, 36 sequences (15 haplotypes) can not be assigned to one of the five sub-clades. In fact sub-clade C1b can not been indentified on the bases of coding region motifs alone. Note that mC corresponds to multiple Cs (the length variation at site 5895-5899) (see Section 5.2.3

A median-joining network was generated from 61 complete C1 sequences and three Icelandic C1 sequences (see Figure 25). The network is constructed from the same sequences as used in the manually constructed tree shown in Figure 22. This network shows the phylogenetic relationship between complete C1 sequences from Iceland and elsewhere in a different way. This analysis was also done to verify the accuracy of the manually constructed trees. The network verifies the order of mutations shown in the phylogenetic tree (Figure 22).

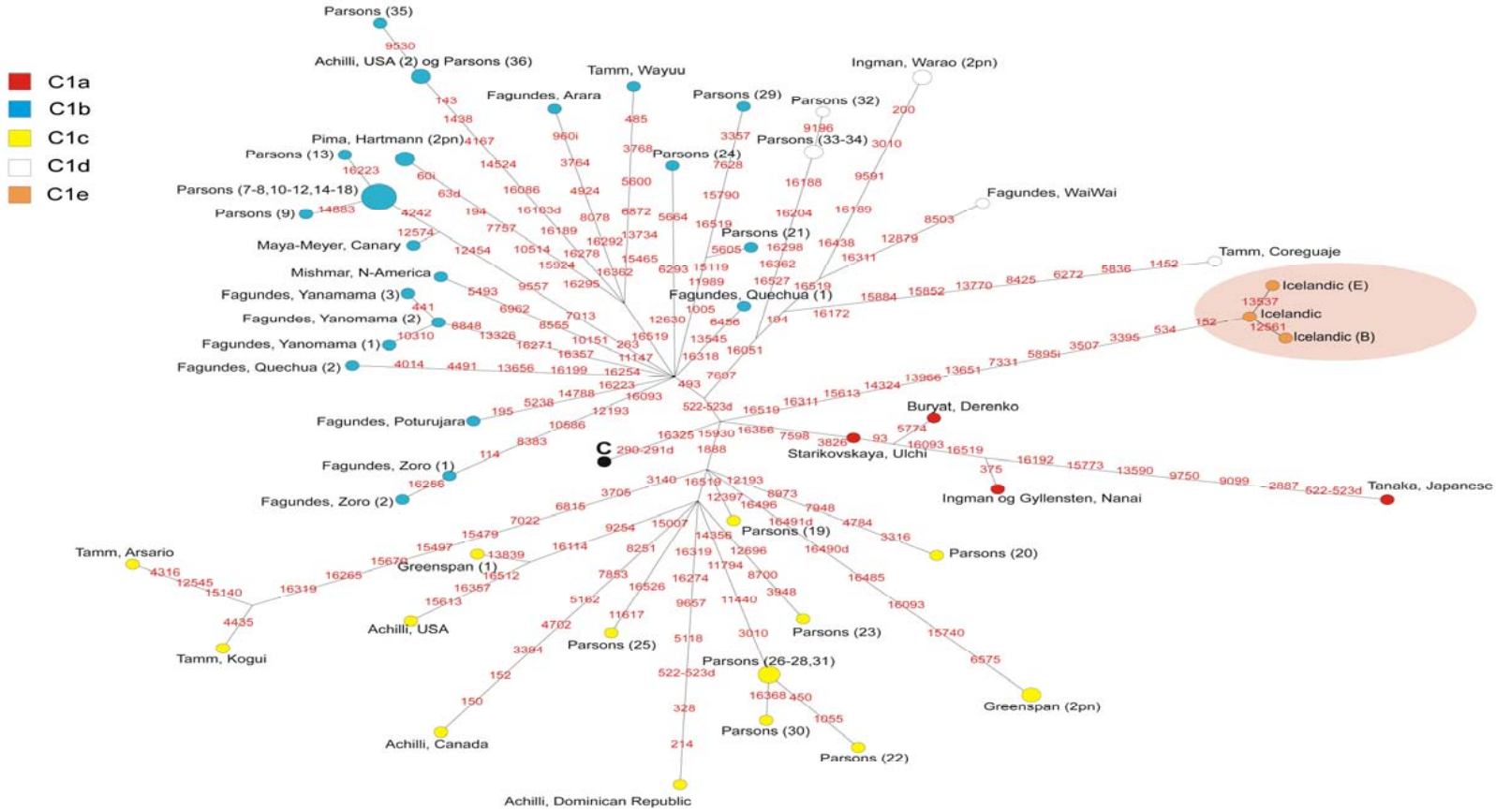

Figure 25. Network of complete mtDNA sequences of haplogroup C1. This median-joining network shows the complete mitochondrial genome motif for 64 sequences (47 haplotypes), the same sequences are shown in the Figure 14. The position of C (set in bold type) is indicated for reading off sequence motifs. Circles are proportional to haplotype frequency, the smallest representing single sequence and the largest 10 copies of the same haplotype. The colour of the circles represents the sub-clade the sequence belongs to. Lines represent mutational differences. The numbers by the lines represent the position of mutations (recurred mutations 16519 and 522-523 are also shown). Note that the recurred mutation 522-523d does not separate the C1b and C1d sub-clades form the others like the network implies. The Icelandic sequences that belong to sub-clade C1e are identified by an orange circle.

Because more complete C1 sequences were used in this study than have been done before we wanted to estimate the age of haplogroup C1 based on the current sequences, which we had assembled. To get a better idea about the age of C1e branch the coalescence age of C1 and three sub-clades were estimated. Coalescence-age calculations were estimated based on the phylogenies of coding sequences. The rate of one coding region mutation every 3533 years (Soares *et al.*, 2009) was used to calculate the coalescence age of haplogroup C1 and C1 sub-clades. The coalescence time of the Native American sub-clades C1b, C1c and C1c are shown in Figure 26, C1a is clearly under-represented (comprising only four mtDNAs) and thus its coalescence time was not calculated. These results show similar coalescence time for the three Native American sub-clades as described by Tamm, et al (2007) although different calculation methods were used in their study.

The age of haplogroup C1 is approximately 14,600±1,700 when all sequences except for the Icelandic are used. When the Icelandic sequence is included in the calculation then the age becomes approximately 14,800±1,650 years ago (mutations that occur within the Icelandic sequences were not included in the calculation). If the Icelandic sequence is in fact of a Native American origin the age of haplogroup C1 increases by 200 years.

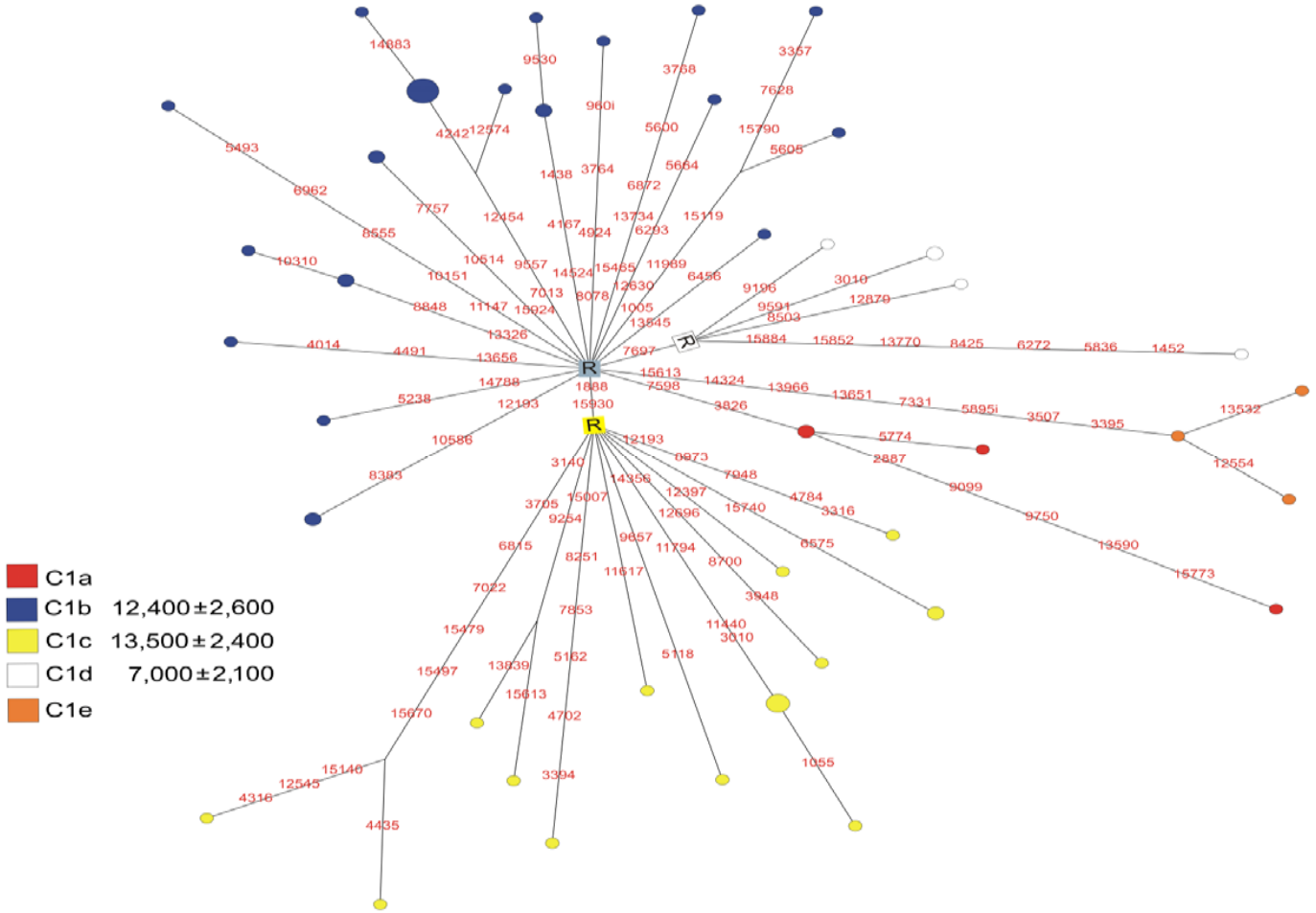

Figure 26. Network of coding region mtDNA sequences of haplogroup C1.

This median-joining network show the coding region motif for 64 sequences (41 haplotypes). Circles are proportional to haplotype frequency. The colors of the circles represent the sub-clade the sequence belongs to. The R-squares represent the roots of the sub-clades. Time estimated shown for clades, C1b, C1c and C1d are averaged distance (ρ) of each haplotype to respective root.

5.3 Analyses of control region sequences

5.3.1 Finding the closest match to the Icelandic C1 sequences

As there are many sequences from haplogroup C1 in the literature that have only been sequenced for the control region we wanted to examine them to see if that would give us any additional information about the origin of the Icelandic sequence. A database of 28,514 mtDNA control region sequences from the literature has been compiled at deCODE Genetics. The database comprises many of the mtDNA sequences that have been published from 1990 to the present day along with unpublished sequences from deCODE Genetics. There are 964 haplogroup C control region sequences in the database, thereof 447 individuals have been sequenced for the whole control region (HVS1 and HVS2), but only a part of the control region (HVS1) has been sequenced for the remaining 517 individuals. The start and stop sites of both HVS1 and HVS2 differ among sequences. We found ten sequences that have the same HVS1 motif as the Icelandic sequence, four come from the Yungay population in Peru (Lewis *et al.*, 2007), four from ancient mtDNA analysis form the now extinguished Tainos population of the Caribbean (dated 670±70 to 1680±100 A.D.) (Lalueza-Fox *et al.*, 2001), one from Germany (Pfeiffer *et al.*, 2001) and one from Chile (see Table 6). Two additional ancient Native American mtDNA sequences have a similar HVS1 motif as the Icelandic sequences, the only difference is that they have an additional mutation, 16126C. One of these sequences was found in the Norris Farm cemetery in Illinois. The skeletal remains found at this location belong to the Oneota culture (dated 1300 AD) (Stone & Stoneking, 1998). The other sequence is also Pre-Columbian, found in the Ciboney population from Cuba (dated 380 AD) (Lalueza-Fox *et al.*, 2003). At last, one sequence found in Chile has a similar HVS1 motif as the Icelandic sequence, the only difference between the two is that the Chilean sequence has one additional mutation at site 16343 (Horai, 1994). A few other sequences had the same HVS1 motif as the Icelandic one, but showed differences in HVS2. One Swedish sequence has the same motif as the Icelandic sequence in HVS1 (Tillmar *et al.*, 2009). However, this sequence does not have the "Icelandic" 152C and 534T mutations and it has three additional mutations in HVS2, two of which suggest that

it belongs to sub-clade C1b, i.e. the transition at site 493 and the 522-523 deletion. From this it can be concluded that the Swedish sequence is likely of Native American origin, probably as the result of recent admixture. Two haplotypes found in the FBI database (Monson K.L., 2002) have similar control region motifs as the Icelandic haplogroup C1 haplotype. However, both of them have three additional mutations, 16129A, 16127C and 146C, and one also has a C insertion at site 302 (see Table 6). Based on the control region motifs alone we cannot tell whether these Hispanic C1 sequences belong to the Native-American sub-clades, C1c or C1b, or if they belong to the same sub-clade as the Icelandic sequence, C1e (see Figure 22).

Table 6. The control region motif for the Icelandic haplogroup C1 sequence and the most similar sequences from other populations. Note that three of the sequences differ by a handful of mutations that are shown in bold.

Country (number of sequences)	Population	Region	HVS1 motif	Sites	HVS2 motif	Sites
Iceland	Icelandic	North Europe	16223T 16298C 16311C 16325C 16327T	15971-16569	73G 152C 249d 263G 290-291d 315_1C 489C 534T	1-599
Peru (4)	Yungay	South America	16223T 16298C 16311C 16325C 16327T	16024-16365	not determined	none
Caribbean (4)	Tainos	South America	16223T 16298C 16311C 16325C 16327T	16055-16410	not determined	none
Germany (1)	German	Northwest Europe	16223T 16298C 16311C 16325C 16327T	16040-16362	not determined	none
Chile (1)	Chilean	South America	16223T 16298C 16311C 16325C 16327T	16129-16400	not determined	none
United States (1)	Oneota	North America	**16126C** 16223T 16298C 16311C 16325C 16327T	16055-16410	not determined	none
Cuba (1)	Ciboney	South America	**16126C** 16223T 16298C 16311C 16325C 16327T	16055-16041	not determined	none
Chile (1)	Chilean	South America	16223T 16298C 16311C 16325C 16327T **16343G**	16129-16400	not determined	none
Sweden (1)	Swedish	North Europe	16223T 16298C 16311C 16325C 16327T	15971-16569	73G **258T** 249d 263G 290-291d 315_1C 489C **493G 522-524d**	1-599
United Stated (1)	Hispanic	North America	**16129A 16172C** 16223T 16298C 16311C 16325C 16327T	16024-16365	73G **146C** 152C 249d 263G 290-291d 315_1C	73-340
United Stated (1)	Hispanic	North America	**16129A 16172C** 16223T 16298C 16311C 16325C 16327T	16024-16365	73G **146C** 152C 249d 263G 290-291d **302_1C** 315_1C	73-340

This database search conforms that the Icelandic C1 sequence is in fact very rare. The ten sequences that have the same HVS1 motif as the Icelandic sequence could potentially belong to the Icelandic C1e sub-clade and same goes for the three sequences that have one additional mutation in the HVS1 region. However, the Swedish sequence is also identical to the Icelandic sequence to for the HVS1, although it does not belong to the same sub-clade as the Icelandic sequence. In order to locate these sequences in the phylogeny of haplogroup C1 and thus, give us information on how similar they are to the Icelandic C1e sequence, further sequencing, of HVS2, preferably along with a few coding region SNPs, would have to be performed.

5.3.2 The phylogeny of C1 control region sequences

Median-joining networks were generated to infer the phylogenetic relationship between the Icelandic haplogroup C1 sequence and other control region sequences belonging to haplogroup C1. The first network was generated from a total of 154 HVS1 (sites 16024-16383) and HVS2 (sites 1-534) haplogroup C1 sequences, obtained from the deCODE Genetics database (see Figure 27). The resulting sequencing data were compared with the revised Cambridge reference sequence (rCRS) (Andrews et al., 1999).

A total of 77 different HVS1 haplotypes were observed among the 154 sequences. According to this network the Asian-specific branch, C1a, forms a separate cluster, defined by transition at site 16356. Sub-clade C1b also forms a separate cluster, defined by one control region mutation at site 493. Furthermore, haplotypes that belong to sub-clade C1d can be distinguished by one control region mutation at site 16051. One haplotype has both the characteristic mutation for sub-clade C1b and sub-clade C1d and therefore we cannot say in which of the two it belongs, based only on control region sequence data (see Figure 27, labeled as C1b or C1d). Sub-clade C1c is defined only by coding region mutations and therefore cannot be defined on the basis of control region motifs.

The network does not give us any additional information about the possible origin of the Icelandic sequence, it only tells us that it does not belong to sub-clades C1a, C1b or C1d, which we already knew from the analyze of complete sequences.

It is clear that HVSI and HVS2 data alone do not provide sufficient information about the phylogenetic relationships of sub-clades within haplogroup C1. The reason for this is that the control region experiences a high frequency of recurrent mutations that makes it difficult to examine the phylogeny of haplotypes at such a fine-scale. The variation found by sequencing just the control region is often insufficient to tell whether mutations that are identical by state are also identical by descent. For example, it is known that site 152 has multiple mutation events, but this is not represented in the network and therefore the Icelandic sequence, which has this mutation, falls into the same cluster as other sequences that have it although they are most likely only identical by state, but not by descent.

The second network was generated from a total of 280 HVS1 (sites 16040-16400) haplogroup C1 sequences, obtained from the deCODE Genetics database, thereof six sequences were from Icelandic individuals (see Figure 28). The resulting sequencing data were compared with the revised Cambridge reference sequence (rCRS) (Andrews et al., 1999). A total of 84 different HVS1 haplotypes were observed. The Asian-specific branch of C1 (C1a) forms a separate cluster, defined by transition at site 16356, and the Native American C1d branch is defined by a transition at site 16051. However, it is clear that we do not get sufficient information about the phylogenetic relationship between C1 sequences from HVS1 information alone. Three sequences have the same motif as the Icelandic sequence for HVS1 (sites 16040-16400), the rapidly mutating site 16311 separates them from other C1 sequences. When we took a closer look at the motifs for these three sequences we found out that one of the sequences has an additional mutation at site 16456 that the Icelandic sequence does not have. The HVS2 had also been sequenced for the two other samples and they did not carry the “Icelandic” 152 mutation. Thus, it is not clear that these sequences are close relatives of the Icelandic sequence. It is clear that HVS1 data alone cannot provide information about the origin of the Icelandic sequence in haplogroup C1 in view of the number of highly mutable sites. However, these analyses confirm that the Icelandic type is rare in other populations and indicates, albeit weakly, that the closest relatives are to be found in the Americas or in Europe.

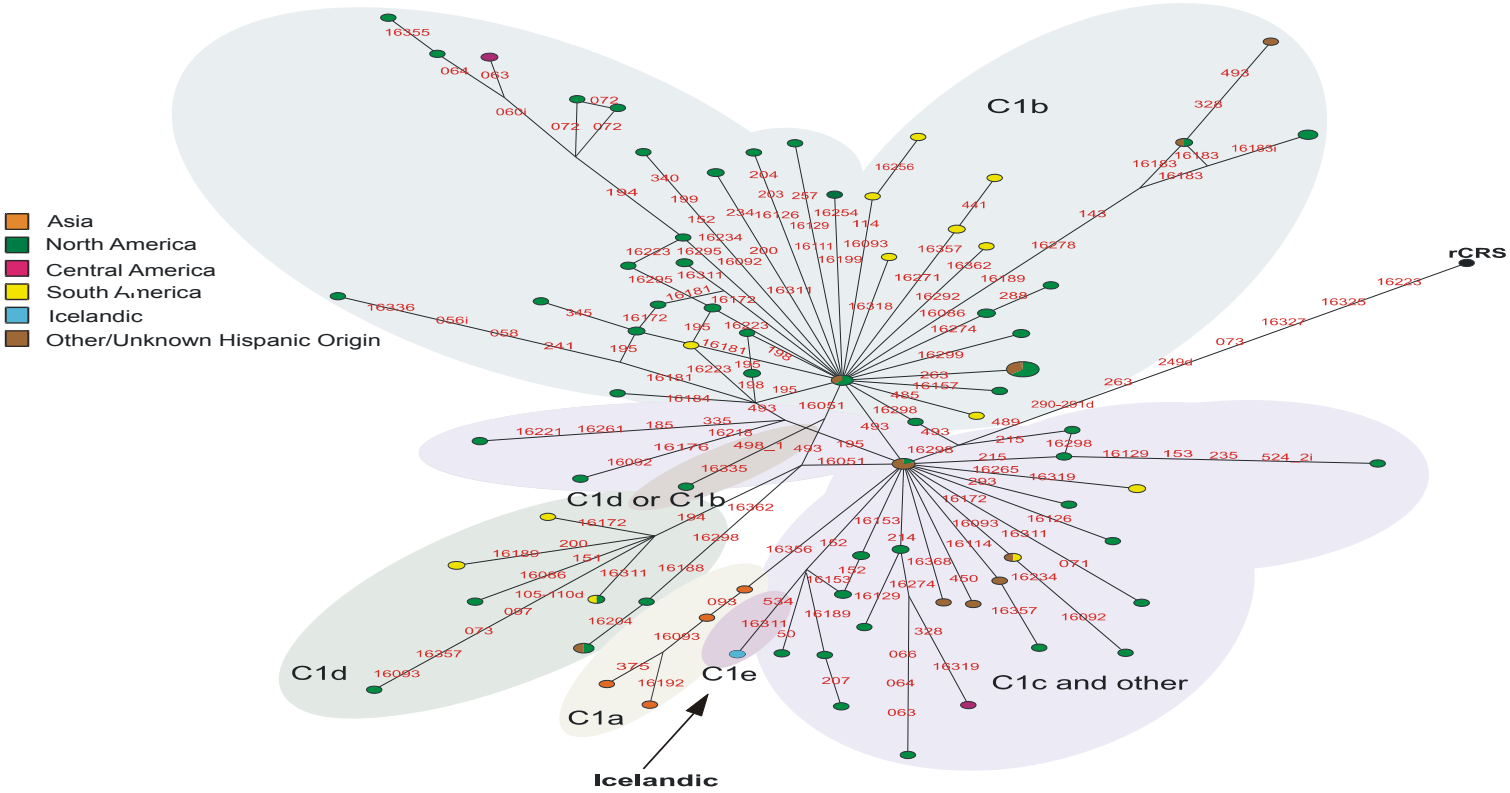

Figure 27. Network of mtDNA control region (HVS1 and HVS2) sequences from haplogroup C1.

This median-joining network shows the control region motif for 145 sequences (77 haplotypes), belonging to haplogroup C1. One Icelandic haplogroup C1 sequences is included in the network. The position of the rCRS (set in bold type) is shown in the network. The circles represent the individual haplotypes and their relative sizes denote the frequency of the haplotype. The geographical origin of each individual haplogroup is indicated by the color of the circle. Lines between haplotypes represent mutational differences. The numbers by the lines represent the position of the mutations (listed relative to rCRS) that distinguish the haplotypes. Membership of sequences in sub-clades of C1 is indicated by background shading and labels (.i.e. C1a, C1b, etc.).

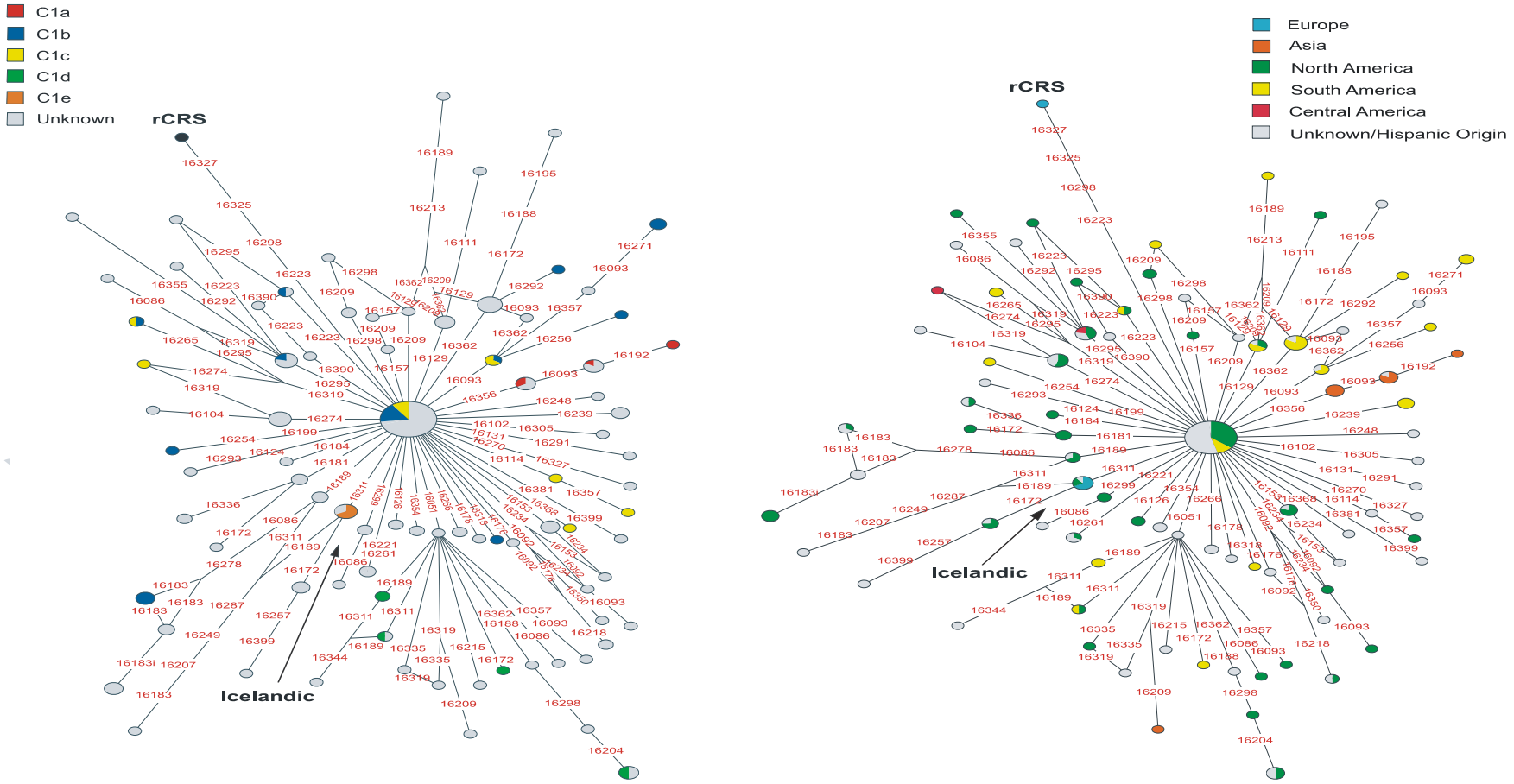

Figure 28. Networks for hypervariable segment 1 (HVS1). These median-joining networks show the HVS1 (16040-16400) for 280 haplogroup C1 sequences (47 haplotypes). Six Icelandic haplogroup C1 sequences are included in the networks. The position of the rCRS (set in bold type) is shown in the network. The two networks are identical apart from the colour of the circles: for the network on the left the colour of the circles represents the sub-clade the sequences belong to, but for the network on the right the colour of the circles represent the geographic or ethnic origin of the sequences. The circles represent the individual haplotypes and their sizes denote the frequency of the haplotype. Lines represent mutational differences. The numbers by the lines represent the site of mutations.

6 DISCUSSION

6.1 The genealogical history of the Icelandic C1 lineage

The fact that we have access to comprehensive genealogical information for the past ten centuries in Iceland allows unusually detailed conclusions to be drawn about the genealogical history of the Icelandic haplogroup C1 lineage. Our combined data of sampled Icelanders with control region sequences set included 2,266 contemporary individuals, who could be traced back to a total of 1,096 matrilineal ancestors. These ancestors, in turn, left a total of 296,708 matrilineal descendants, which amounts to 72.56% of the modern Icelandic mtDNA pool which consists of 408,915 individuals (defined as all individuals born 1895-2005). Of those 2,266 sampled individuals, seven were haplogroup C1 carriers, corresponding to frequency of haplogroup C1 in Iceland of about 0.3 %. We note that there is a bias in the age distribution of individuals included in our initial sample, which includes a disproportionate number of old or middle-aged individuals. However, it is unlikely that the frequency of haplogroups has changed much over the last three or four decades, therefore our estimate is likely to apply to the modern population. Based on a frequency of 0.3% we would expect to find about 1,227 haplogroup C1 carriers in the contemporary Icelandic population. The seven haplogroup C1 carriers found among the sampled individuals were traced back to four matrilineal ancestors, named A, B, D and E. Those four ancestors have 671 contemporary matrilineal descendants, who are therefore haplogroup C1 carriers. It can be surmised that there is at least one, currently unidentified lineage of haplogroup C1 carriers that correspond to the estimated 556 individuals (1227-671) given a frequency of 0.3%. As the sample covers 72.56% of the contemporary Icelandic mtDNA pool we could also surmise that we may have uncovered the same percent of haplogroup C1 carriers in which case we would expect 337

unsampled C1 carriers. It is also possible (albeit unlikely) that our sample covers all Icelandic haplogroup C1 carriers found in the contemporary Icelandic mtDNA pool.

6.2 The founding age of the Icelandic C1 lineage

From the genealogical analysis, it could be concluded that the presence of a haplogroup C1 lineage in the Icelandic mtDNA pool was not the result of recent admixture. Haplogroup C1 carriers identified in this study could be traced back at least eight generations to four matrilineal ancestors, named A, B, D and E. These female ancestors were born between 1710 and 1740 AD. The latest possible date for the MRCA of the Icelandic haplogroup C1 lineage is 1690 AD, if it is assumed that ancestors A, B, D and E were all sisters. The settlement of Iceland has been dated to the period 874-930 AD. We can conclude that if a single founding woman, located in Iceland, gave rise to the Icelandic haplogroup C1 lineage she must have arrived sometimes between 874 and 1690 AD. A study of the geographic ancestry of Icelanders indicates that individuals tended to live in the same region as most of their ancestor five generation previously (Helgason *et al.*, 2005). The limited geographic spread of the four Icelandic C1 lineages supports the notion that they derive from a single female ancestor in Iceland.

During the period from settlement to the year 1262 AD Icelanders were relatively prosperous. During this time Icelandic Vikings travelled far and wide, trading and raiding. However, during the end of the 13th century deteriorating climate imposed limits to the growing Icelandic population. Icelanders started to lag behind in the development of trade and lost their own foreign trade to others. During the next four centuries voyages to and from the country were infrequent, which hindered post-settlement immigration to the island (Karlsson, 2000). Although other European populations were developing and exploring new territories during this period in history, America being one of them, Icelanders were, for the most part, isolated from the outside world. Therefore the Icelandic population has not been affected by constant gene flow like many other populations. Thus, it can be assumed that most mtDNA lineages observed in the contemporary Icelandic population are descended from the original set of female

settlers (Helgason et al., 2000a). Accordingly, it is likely that the earliest ancestor that carried the haplogroup C1 lineage in Iceland was among the original settlers 1,100 years ago. Supporting this notion is the fact that one of the four lineages, E, has one mutation that sets it apart from the other lineages. When the age of the MRCA for all four lineages was estimated, based on the mutation rate, it fell within the period of settlement. However, this estimate is not very robust.

6.3 The phylogeny of the Icelandic C1 sequence

All complete mtDNA C1 sequences observed prior to this study fall into one of four C1 sub-clades, C1a, C1b, C1c or C1d. Sub-clade C1a is Asian-specific, whereas the other three sub-clades have only been found among individuals with Native American ancestry. Phylogenetic analysis of individuals from Siberia suggest that the founding haplotype of the Native American C1 mtDNA haplogroup originated in Siberia (Starikovskaya *et al.*, 2005). The vast majority of complete C1 sequences that have been identified fall into one of the Native American branches. In light of this, combined with the fact that there is firm archaeological evidence for a Viking settlement in America sometime between 980 and 1020 AD, we originally suspected that the Icelandic C1 sequence had a Native American ancestry and thus, would be placed within one of the three Native American C1 sub-clades. Interestingly, the complete mtDNA sequences of Icelandic haplogroup C1 carriers did not belong to any of the four previously identified C1 sub-clades. Rather, it belongs to a distinct sub-clade, which we named C1e. This sub-clade has no other members at present. Consequently, we were unable to determine the geographical origin of the Icelandic C1 haplotype in this study. In all, eleven mutations differ between the Icelandic C1e sub-clade and the other four C1 sub-clades. That is approximately the same number of mutation found within the other sub-clades which indicates divergence near the time of the MRCA for haplogroup C1.

The finding that the Icelandic C1 sequence belongs to a previously unidentified sub-clade of haplogroup C1 was unexpected. Small populations like the Icelanders have been more heavily affected by genetic drift than larger populations and therefore we would not expect that such rare sequence would

survive in Icelanders, but not in target source populations (Helgason *et al.*, 2009). On the other hand, the Icelanders are one of the most studied populations in human genetics, thus detection of rare haplotypes in the Icelandic gene pool is expected to be more comprehensive than in any other European gene pool.

Instead of answering the question about the origin of the Icelandic C1 sequence the results from phylogenetic analysis only increased the mystery surrounding the Icelandic C1 lineage. However, the number of complete mtDNA sequences from C1 is still limited, thus it is likely that further sampling will reveal haplotypes that will fall into the same branch as the Icelandic sequence. The important question remains, in which population will these sequences be found? Haplogroup C1 is found both in Native America and in Asia, thus these new C1e haplotypes could come from either of these geographical regions. Although, haplogroup C1 sequences are found in more frequency in Native American populations, the founding C1 haplogroup originated in Asia. Furthermore, the fact that Icelandic settlers were of European origin combined with the fact that the Icelandic C1 sequence does not fall into any of the Native American sub-clades raises the possibility that C1e is a very rare European branch. However, until further sampling reveals haplotypes that fall into C1e, this must remain the subject of speculation.

6.4 A Native American origin for the Icelandic C1 lineage

One interpretation of the possible origin of the Icelandic C1 sequence based on the results from phylogenetic analysis, is that it probably does not have a Native American origin, since if it did we would expect it to belong to one of the two most frequently observed Native American C1 sub-clades, C1b or C1c (see Figure 22, section 5.2.4). Furthermore, the mtDNA variation found in contemporary Native American populations is limited and thus if the C1e sub-clade originated in America, one might expect that it would have been found by now. However, if we assume that sub-clade C1e was relatively rare in the first Native American populations, then it might now be almost lost from the mtDNA pool of contemporary Native Americans as a result of genetic drift, and thus not get sampled.

Most haplogroup C1 sequences from the literature have only been sequenced for the control region. The majority of these sequences cannot be classified into sub-clades, because the control region does not contain most of the key mutations that define the C1 sub-clades. An examination of haplogroup C1 control region HVS1 sequences from the literature revealed thirteen that could belong to C1e(see Section 5.3.1, Table 6). Twelve of those thirteen sequences were found among Native Americans populations; one was found in a German but none in Asian populations. This points, albeit weakly to a Native American origin for the Icelandic C1 sequence. The majority of the potentially C1e sequences are found in South America, five in the Caribbean, four in Peru and two in Chile, whereas one is found in Midwest North America. Interestingly, six of those sequences were found in ancient Native American remains, dated from around 380 to 1680 AD. Thus, if it turns out that those sequences belong to sub-clade C1e it could be inferred this clade was found in a higher frequency a few centuries ago, but has been almost lost as a result of genetic drift. This could explain why the Native American sequences that would fall into sub-clade C1e have not been discovered yet.

Another intriguing observation is that Beothuk Indians, which Vikings most likely encountered during their brief colonization of America, were completely wiped out in the late 1820s (Kuch *et al.*, 2007; Williamsen, 2005). There is little known about many aspects of the Beothuk culture, its origin and there relationship with neighboring native groups. However, some scholars believe that the Beothuk Indians were isolated from other groups and linguistic differences have been regarded as evidence for this (Donald, 2003). Although, farfetched there is a possibility that C1e sequences were found in a higher frequency in this particular native group and following its extinction the sub-clade it self became lost as well. Ancient DNA studies will hopefully be able to shed light on this in the future.

Only 61 complete haplogroup C1 sequences have been published in the literature to date. Thus, there is a chance that sequences that would fall into the C1e sub-clade have been sampled, but that only their control region have been sequenced, which is not enough to establish that they belong to C1e . There is a precedent for recently discovered rare Native American mtDNA haplogroups. For example, haplogroup X2a, was first identified in a study by Forster *et al.*, (1996)

and haplogroup D4h3, first identified a decade ago by Rickards *et al* (1999). Since then only twelve complete sequences from haplogroup X2a and two from haplogroup D4h2 have been reported in the literature (Bandelt *et al.*, 2003; Fagundes et al., 2008b; Reidla et al., 2003; Tamm et al., 2007).

If further sampling reveals that the Icelandic C1 sequence is of Native American origin, the founding ancestor would have had to arrive to Iceland either during the initial settlement (870-930AD) or after Columbus re-discovered America, i.e. after the year 1492 AD. Explorers and adventurers have long returned home with living humans for display, as Columbus did when he took Native American back to Spain in 1493 (Columbus, 1992). Thus, there is a possibility that a Native American woman brought to Europe during the colonial period in America (1492-1776 AD) could have ended up in Iceland. However, this seems unlikely considering that the voyages to and from Iceland were infrequent during this time. Moreover, Native American haplotypes are rarely found in European populations, suggesting limited admixture, in Europe, between Europeans and Native Americans. Therefore it can be said that if the Icelandic C1 sequence is of Native American ancestry, the most likely scenario is that a single Native American woman came here during the initial settlement and that her mtDNA has been transmitted through matrilineal descendants to cotemporary Icelanders. This would indicate contact between Icelanders and Native Americans prior to Columbus's re-discovery of the continent and thus, provide an important insight into the history of contact between Europeans and Americans.

6.5 A European or Asian origin for Icelandic C1 lineage

The larger size of most European and Asian populations, and the fact that only a small portion of the populations found in these areas have been sampled for mtDNA variation, could explain why C1e sequences have not been found yet, even though they might exist at low frequency in either Asia or in Europe. If further sampling reveals C1e sequences from Europe, then this would most likely European origin of C1e. The fact that one of the thirteen HVS1 sequences that potentially belong to sub-clade C1e was found in Germany at least raises this possibility. If C1e haplotypes are found in Asia, we can not say whether it has

arrived in Iceland via America or Europe. Although, since the original Icelandic settlers came from Europe that would be the most likely route of passage. If the Icelandic lineage is of European or Asian origin, then the founding ancestor could have migrated to Iceland anytime during the period 874 to 1690 AD, although it most likely that she arrived with the original settlers 1100 years ago.

The eleven mutations that distinguish the "Icelandic" C1e sequences from the other four C1 sub-clades suggest divergence near the time of the MRCA for haplogroup C1. This could be taken to support the hypothesis of a European or Asian origin, since we would expect to see a distinctive European or Asian C1 sub-clade that could be distinguished from other C1 sub-clades. This is the pattern that can be observed in the case of haplogroup X (Reidla et al., 2003). Haplogroup X2 is found in low frequency ~4% in western Eurasians, but a distinctive sub-clade named X2a is one of the minor Native American founding lineages. This is the only known case to date of a haplogroup that is native to both Europe and the Americas. The C1e sub-clade might be another example of this kind of diffusion.

6.6 Further research

The results of this study are inconclusive about the origin of the Icelandic C1e sequence, and leave us with an even greater mystery than before. The Icelandic C1e haplotype could either be of Native American, European or Asian origin. Further research is necessary to determine which of these regions gave rise to C1e, but at present the strongest evidence points to a Native American origin.

In order to provide us with more evidence on the geographical origin of the Icelandic haplogroup C lineage some further analysis would be informative. There are a few existing mtDNA haplogroup C HVS1 sequences in the literature (see Table 6) that match the Icelandic sequence and might therefore belong to the C1e sub-clade. If extracts from these samples could be obtained it would be worthwhile to examine them in more detail to test for the coding-region mutations that define the Icelandic C1e or test for branch defining coding-region mutations for the other C1 sub-clades. This would determine whether these HVS1 sequences in fact belong to the C1e or another C1 sub-clade.

The implications of a positive result are potentially direct genetic evidence of pre-Colombian contact between Vikings and Native Americans. The findings of this thesis support that the founding ancestor of the Icelandic C1 lineage arrived around the time of settlement of Iceland (874 -930 AD), which is around the same time as Vikings discovered America (ca.1000 AD). However, there is no conclusive evidence for this and it is unlikely that such evidence will ever come to light. To gain more information about the founding age of the C1e lineage in Iceland, we could attempt to find more Icelandic haplogroup C carriers and thus, new haplogroup C1 lineages, by analysing a larger sample size of modern Icelanders. This could potentially reveal additional mutations that have occurred within the Icelandic C1 haplogroup, which could give us further information about the age of the Icelandic haplogroup C1 lineage. Moreover, through new matrilines we might be able to identify the matrilineal ancestors of the C1 haplotype in Iceland that lived prior to 1710 AD, the birth year of the earliest known matrilineal ancestor known to date.

If it turns out, thought the analysis of complete sequences that the German haplogroup C HVS1 sequence (see Table 6) that matches the HVS1 motif of the Icelandic C1e sequence also showes coding region muations that confirm membership in C1e then that would indicate a European origin. However there would still be the mystery of how the sequence the C1e branch survived in Europe for so long and so rare, and where it came from originally. If the Icelandic C1 sequence originated in Europe, then the haplotype most likely came to Iceland with the original settlers from Scandinavia or the British Islands. Sequence data from individuals from these areas might yield a C1e haplotype, if the sample is large enough. However, if only a single C1e sequence were to be discovered that would not give us any conclusive information about the origin of the Icelandic C1 lineages, as it might just as well have been brought to Europe from Iceland. Thus, we would have to find a few European haplotypes across Europe to be able to make the conclusion that the Icelandic lineage has a European origin. Other factors, such as the location of these new C1e sequences in the phylogenetic context of C1e sub-clade would also play an important part.

Further studies on complete mtDNA, will hopefully provide us with data that can confirm the geographic origin of the Icelandic sequence. Whatever the conclusion will be, it will not only shed light on the population history of Iceland, it would also provide an important insight into the phylogeny and distribution of haplogroup C1.

REFERENCES

Achilli, A., Perego, U. A., Bravi, C. M., Coble, M. D., Kong, Q. P., Woodward, S. R., et al. (2008). The phylogeny of the four pan-American MtDNA haplogroups: implications for evolutionary and disease studies. *PLoS One, 3*(3), e1764.

Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H., Coulson, A. R., Drouin, J., et al. (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. *Nat Genet, 23*(2), 147.

Bandelt, H. J., Forster, P., Sykes, B. C., & Richards, M. B. (1995). Mitochondrial portraits of human populations using median networks. *Genetics, 141*(2), 743-753.

Bandelt, H. J., Herrnstadt, C., Yao, Y. G., Kong, Q. P., Kivisild, T., Rengo, C., et al. (2003). Identification of Native American founder mtDNAs through the analysis of complete mtDNA sequences: some caveats. *Ann Hum Genet, 67*(Pt 6), 512-524.

Barton, N. H., Briggs, D. E. G., Eisen, J. A., Goldstein, D. B., Patel, N. P. (2007). *Evolution*. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press.

Behar, D. M., Rosset, S., Blue-Smith, J., Balanovsky, O., Tzur, S., Comas, D., et al. (2007). The Genographic Project public participation mitochondrial DNA database. *PLoS Genet, 3*(6), e104.

Bendall, K. E., Macaulay, V. A., & Sykes, B. C. (1997). Variable levels of a heteroplasmic point mutation in individual hair roots. *Am J Hum Genet, 61*(6), 1303-1308.

Bergþórsson, P. (2000). *The Wineland Millenium. Saga and Evidence*. Reykjavík: Mál og menning.

Bonatto, S. L., & Salzano, F. M. (1997). Diversity and age of the four major mtDNA haplogroups, and their implications for the peopling of the New World. *Am J Hum Genet, 61*(6), 1413-1423.

Brandstatter, A., Niederstatter, H., Pavlic, M., Grubwieser, P., & Parson, W. (2007). Generating population data for the EMPOP database - an overview of the mtDNA sequencing and data evaluation processes considering 273 Austrian control region sequences as example. *Forensic Sci Int, 166*(2-3), 164-175.

Brown, M. D. (1998). mtDNA Haplogroup X: An ancient link between Europe/Western Asia and Noth America? *The American Journal of Human Genetics, 63*, 1852-1861.

Cavalli-Sforza, L. L., & Feldman, M. W. (2003). The application of molecular genetic approaches to the study of human evolution. *Nat Genet, 33 Suppl*, 266-275.

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

Columbus, C. (1992). *The voyage of Christopher Columbus: Columbus' own journal of discovery* (J. Cummins, Trans.): Weidenfeld & Nicolson.

Derenko, M., Malayarchuk, B., Grzybowski, T., Denisova, G., Dambueva, I., Perkova, M., et al. (2007). Phylogeographic Analysis of Mitochondrail DNA in Nothern Asian Populations. *The American Journal of Human Genetics, 81*.

Donald, H. H. J. (2003). A Historiography of an Ahistoricity: on the Beothuk Indians. *History and Anthropology, 14*(2), 127-140.

Fagundes, N. J., Kanitz, R., & Bonatto, S. L. (2008a). A reevaluation of the Native American mtDNA genome diversity and its bearing on the models of early colonization of Beringia. *PLoS One, 3*(9), e3157.

Fagundes, N. J., Kanitz, R., Eckert, R., Valls, A. C., Bogo, M. R., Salzano, F. M., et al. (2008b). Mitochondrial population genomics supports a single pre-Clovis origin with a coastal route for the peopling of the Americas. *Am J Hum Genet, 82*(3), 583-592.

Fergusson, B. W. (2001). L´Anse aux Meadows and Vínland. In A. Wawn & Sigurðardóttir (Eds.), *Approaches to Vínland* (pp. 134-146). Reykjavík: Sigurður Nordal Institute.

Forster, P. (2004). Ice Ages and the mitochondrial DNA chronology of human dispersals: a review. *Philos Trans R Soc Lond B Biol Sci, 359*(1442), 255-264; discussion 264.

Forster, P., Harding, R., Torroni, A., & Bandelt, H. J. (1996). Origin and evolution of Native American mtDNA variation: a reappraisal. *Am J Hum Genet, 59*(4), 935-945.

Goebel, T., Waters, M. R., & O'Rourke, D. H. (2008). The late Pleistocene dispersal of modern humans in the Americas. *Science, 319*(5869), 1497-1502.

Griffiths, A. J. F. (2008). *Introduction to genetic analysis* (9th ed.). New York: W.H. Freeman and Co.

Halldórsson, Ó. (2001). The Vínland Sagas. In A. Wawn & Sigruðardóttir (Eds.), *Approaches to Vínland*. Reykjavík: Sigurður Nordal Institute.

Hauswirth, W. W., & Laipis, P. J. (1982). Mitochondrial DNA polymorphism in a maternal lineage of Holstein cows. *Proc Natl Acad Sci U S A, 79*(15), 4686-4690.

Helgason, A., Hickey, E., Goodacre, S., Bosnes, V., Stefansson, K., Ward, R., et al. (2001). mtDNA and the islands of the North Atlantic: estimating the proportions of Norse and Gaelic ancestry. *Am J Hum Genet, 68*(3), 723-737.

Helgason, A., Hrafnkelsson, B., Gulcher, J. R., Ward, R., & Stefansson, K. (2003a). A populationwide coalescent analysis of Icelandic matrilineal and patrilineal genealogies: evidence for a faster evolutionary rate of mtDNA lineages than Y chromosomes. *Am J Hum Genet, 72*(6), 1370-1388.

Helgason, A., Lalueza-Fox, C., Ghosh, S., Sigurethardottir, S., Sampietro, M. L., Gigli, E., et al. (2009). Sequences from first settlers reveal rapid evolution in Icelandic mtDNA pool. *PLoS Genet, 5*(1), e1000343.

Helgason, A., Nicholson, G., Stefansson, K., & Donnelly, P. (2003b). A reassessment of genetic diversity in Icelanders: strong evidence from multiple loci for relative homogeneity caused by genetic drift. *Ann Hum Genet, 67*(Pt 4), 281-297.

Helgason, A., Palsson, G., Pedersen, H. S., Angulalik, E., Gunnarsdottir, E. D., Yngvadottir, B., et al. (2006). mtDNA variation in Inuit populations of Greenland and Canada: migration history and population structure. *Am J Phys Anthropol, 130*(1), 123-134.

Helgason, A., Sigurdardottir, S., Gulcher, J. R., Ward, R., & Stefansson, K. (2000a). mtDNA and the origin of the Icelanders: deciphering signals of recent population history. *Am J Hum Genet, 66*(3), 999-1016.

Helgason, A., Sigurdardottir, S., Nicholson, J., Sykes, B., Hill, E. W., Bradley, D. G., et al. (2000b). Estimating Scandinavian and Gaelic ancestry in the male settlers of Iceland. *Am J Hum Genet, 67*(3), 697-717.

Helgason, A., Yngvadottir, B., Hrafnkelsson, B., Gulcher, J., & Stefansson, K. (2005). An Icelandic example of the impact of population structure on association studies. *Nat Genet, 37*(1), 90-95.

Herrnstadt, C., Elson, J. L., Fahy, E., Preston, G., Turnbull, D. M., Anderson, C., et al. (2002). Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences for the major African, Asian, and European haplogroups. *Am J Hum Genet, 70*(5), 1152-1171.

Hey, J. (2005). On the number of New World founders: a population genetic portrait of the peopling of the Americas. *PLoS Biol, 3*(6), e193.

Horai, S. (1994). Peopling of the Americas, founded by four major lineages of mitochondrial DNA. *Tanpakushitsu Kakusan Koso, 39*(15), 2759-2767.

Howell, N., Smejkal, C. B., Mackey, D. A., Chinnery, P. F., Turnbull, D. M., & Herrnstadt, C. (2003). The pedigree rate of sequence divergence in the human mitochondrial genome: there is a difference between phylogenetic and pedigree rates. *Am J Hum Genet, 72*(3), 659-670.

Ingman, M., & Gyllensten, U. (2001). Analysis of the complete human mtDNA genome: methodology and inferences for human evolution. *J Hered, 92*(6), 454-461.

Ingman, M., & Gyllensten, U. (2007). Rate variation between mitochondrial domains and adaptive evolution in humans. *Hum Mol Genet, 16*(19), 2281-2287.

Irwin, J. A., Saunier, J. L., Niederstatter, H., Strouss, K. M., Sturk, K. A., Diegoli, T. M., et al. (2009). Investigation of heteroplasmy in the human mitochondrial DNA control region: a synthesis of observations from more than 5000 global population samples. *J Mol Evol, 68*(5), 516-527.

Hagstofa Íslands. (2009). *Mannfjöldi 1. Janúar, 2009*. Reykjavík.

Jobling, M. A., Hurles, M., & Tyler-Smith, C. (2004). *Human evolutionary genetics: origins, peoples & disease*. New York: Garland Science.

Just, R. S., Diegoli, T. M., Saunier, J. L., Irwin, J. A., & Parsons, T. J. (2008). Complete mitochondrial genome sequences for 265 African American and U.S. "Hispanic" individuals. *Forensic Sci Int Genet, 2*(3), e45-48.

Karlsson, G. (2000). *Iceland´s 1100 Years: The History of a Marginal Society*. London: Hurst & Company.

Kelly, R. L. (2002). Maybe we know when people first cam to North America; and what does ot mean if we do? *Quaternary International, 109-111*, 133-145.

Kitchen, A., Miyamoto, M. M., & Mulligan, C. J. (2008). A three-stage colonization model for the peopling of the Americas. *PLoS One, 3*(2), e1596.

Kivisild, T., Shen, P., Wall, D. P., Do, B., Sung, R., Davis, K., et al. (2006). The role of selection in the evolution of human mitochondrial genomes. *Genetics, 172*(1), 373-387.

Kolman, C. J., Sambuughin, N., & Bermingham, E. (1996). Mitochondrial DNA analysis of Mongolian populations and implications for the origin of New World founders. *Genetics, 142*(4), 1321-1334.

Kuch, M., Grocke, D. R., Knyf, M. C., Gilbert, M. T., Younghusband, B., Young, T., et al. (2007). A preliminary analysis of the DNA and diet of the extinct Beothuk: a systematic approach to ancient human DNA. *Am J Phys Anthropol, 132*(4), 594-604.

Lalueza-Fox, C., Calderon, F. L., Calafell, F., Morera, B., & Bertranpetit, J. (2001). MtDNA from extinct Tainos and the peopling of the Caribbean. *Ann Hum Genet, 65*(Pt 2), 137-151.

Lalueza-Fox, C., Gilbert, M. T. P., Martínez-Fuentes, A. J., Calafell, F., & Bertranpetit, J. (2003). Mitochondrial DNA from pre-columbian Ciboneys from Cuba and the prehistoric colonization of the Caribbean. *American Journal of Physical Anthropology, 121*, 97-108.

Lewis, C. M., Jr., Lizarraga, B., Tito, R. Y., Lopez, P. W., Iannacone, G. C., Medina, A., et al. (2007). Mitochondrial DNA and the peopling of South America. *Hum Biol, 79*(2), 159-178.

Lutz, S., Weisser, H. J., Heizmann, J., & Pollak, S. (1999). Mitochondrial heteroplasmy among maternally related individuals. *Int J Legal Med, 113*, 155-161.

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

Magnusson, M., & Pálsson, H. (Eds.). (1965). *The Vinland Sagas: The Norse Discovery of America.* London: Penguin Books.

Malhi, R. S., Eshleman, J. A., Greenberg, J. A., Weiss, D. A., Schultz Shook, B. A., Kaestle, F. A., et al. (2002). The structure of diversity within New World mitochondrial DNA haplogroups: implications for the prehistory of North America. *Am J Hum Genet, 70*(4), 905-919.

Marchington, D. R., Hartshorne, G. M., Barlow, D., & Poulton, J. (1997). Homopolymeric tract heteroplasmy in mtDNA from tissues and single oocytes: support for a genetic bottleneck. *Am J Hum Genet, 60*(2), 408-416.

McGuire, E. H. (2006). Archaeology in Iceland: Recent Developments. *Schandinavian-Canadian Studies, 16*, 10-26.

Mishmar, D., Ruiz-Pesini, E., Golik, P., Macaulay, V., Clark, A. G., Hosseini, S., et al. (2003). Natural selection shaped regional mtDNA variation in humans. *Proc Natl Acad Sci U S A, 100*(1), 171-176.

Monson K.L., M. K. W. P., Wilson M.R., DiZinno J.A., Budowle B. (2002). The mtDNA population database: an integrated software and database resource for forensic comparison, *Forensic Science Communications* (Vol. 4).

Monson, K. L., Miler, K. W. P., Wilson, M. R., DiZinno, J. A., & Budowle, B. (2002). The mtDNA population database: an integrated software and database resource for forensic comparison: Forensic Science Communications.

Perego, U. A., Achilli, A., Angerhofer, N., Accetturo, M., Pala, M., Olivieri, A., et al. (2009). Distinctive Paleo-Indian migration routes from Beringia marked by two rare mtDNA haplogroups. *Curr Biol, 19*(1), 1-8.

Pfeiffer, H., Forster, P., Ortmann, C., & Brinkmann, B. (2001). The results of an mtDNA study of 1,200 inhabitants of a German village in comparison to other Caucasian databases and its relevance for forensic casework. *Int J Legal Med, 114*(3), 169-172.

Rafnsson, S. (1997). The Atlantic Islands. In P. Sawyer (Ed.), *The Oxford Illustrated History of the Vikings* (pp. 110-134). Oxford: Oxford Univeristy Press.

Reidla, M., Kivisild, T., Metspalu, E., Kaldma, K., Tambets, K., Tolk, H. V., et al. (2003). Origin and diffusion of mtDNA haplogroup X. *Am J Hum Genet, 73*(5), 1178-1190.

Relethford, J. (2005). *The human species: an introduction to biological anthropology* (6th ed.). Boston: McGraw-Hill.

Rickards, O., Martinez-Labarga, C., Lum, J. K., De Stefano, G. F., & Cann, R. L. (1999). mtDNA history of the Cayapa Amerinds of Ecuador: detection of additional founding lineages for the Native American populations. *Am J Hum Genet, 65*(2), 519-530.

Rosenblad, E., & Sigurðardóttir-Rosenblad, R. (1993). *Iceland: From Past to Present* (A. Crozier, Trans.). Reykjavík: Mál og menning.

Salas, A., Bandelt, H. J., Macaulay, V., & Richards, M. B. (2007). Phylogeographic investigations: the role of trees in forensic genetics. *Forensic Sci Int, 168*(1), 1-13.

Sawyer, P. (1997). The Age of the Vikings, and Before. In P. Sawyer (Ed.), *The Oxford Illustrated History of the Vikings* (pp. 1-19). Oxford: Oxford Univeristy Press.

Schurr, T. G. (2004). The peopling of the New World: perspectives from molecular Anthropology. *Annual Review of Anthropology, 33*, 551-583.

Sigurðardóttir, S., Helgason, A., Gulcher, J. R., Stefansson, K., & Donnelly, P. (2000). The mutation rate in the human mtDNA control region. *Am J Hum Genet, 66*(5), 1599-1609.

Smith, K. (1995). Landnám: the settelment of Icelend in archaeological and historical perspective. *World Aechaeology, 26*(3).

Soares, P., Ermini, L., Thomson, N., Mormina, M., Rito, T., Rohl, A., et al. (2009). Correcting for purifying selection: an improved human mitochondrial molecular clock. *Am J Hum Genet, 84*(6), 740-759.

Stanyon, R., Sazzini, M., & Luiselli, D. (2009). Timing the first human migration into eastern Asia. *J Biol, 8*(2), 18.

Starikovskaya, E. B., Sukernik, R. I., Derbeneva, O. A., Volodko, N. V., Ruiz-Pesini, E., Torroni, A., et al. (2005). Mitochondrial DNA diversity in indigenous populations of the southern extent of Siberia, and the origins of Native American haplogroups. *Ann Hum Genet, 69*(Pt 1), 67-89.

Stein, P. L., & Rowe, B. M. (2003). *Physical anthropology* (8th ed.). Boston: McGraw-Hill.

Stone, A. C., & Stoneking, M. (1998). mtDNA analysis of a prehistoric Oneota population: implications for the peopling of the New World. *Am J Hum Genet, 62*(5), 1153-1170.

Stoneking, M. (1998). Women on the move. *Nat Genet, 20*(3), 219-220.

Strachan, T., & Read, A. P. (2004). *Human molecular genetics 3* (3rd ed.). London; New York: Garland Press.

Sun, C., Kong, Q. P., & Zhang, Y. P. (2007). The role of climate in human mitochondrial DNA evolution: a reappraisal. *Genomics, 89*(3), 338-342.

Sveinbjarnadóttir, G. (2004). Landnám og Elsta Byggð: Byggðamunstur og Búsetuþróun. In Á. Björnsson & H. Róbertsdóttir (Eds.), *Hlutavelta Tímanns: Menningararfur á Þjóðminjasafni* (pp. 38-48). Reykjavík: Þjóðminjasafn Íslands.

Swindell, S. R., & Plasterer, T. N. (1997). SEQMAN. Contig assembly. *Methods Mol Biol, 70*, 75-89.

Tamm, E., Kivisild, T., Reidla, M., Metspalu, M., Smith, D. G., Mulligan, C. J., et al. (2007). Beringian standstill and spread of Native American founders. *PLoS One, 2*(9), e829.

Tanaka, M., Cabrera, V. M., Gonzalez, A. M., Larruga, J. M., Takeyasu, T., Fuku, N., et al. (2004). Mitochondrial genome variation in eastern Asia and the peopling of Japan. *Genome Res, 14*(10A), 1832-1850.

Thomasson, R. (1980). *Iceland: The First New Society*. Reykjavík: Icelandi Review.

Thomson, E. (1973). The Icelandic admixture problem. *Annuals of Human Genetic, 37*(69), 69-80.

Tillmar, A. O., Coble, M. D., Wallerstrom, T., & Holmlund, G. (2009). Homogeneity in mitochondrial DNA control region sequences in Swedish subpopulations. *Int J Legal Med*.

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

Torroni, A., Schurr, T. G., Cabell, M. F., Brown, M. D., Neel, J. V., Larsen, M., et al. (1993). Asian affinities and continental radiation of the four founding Native American mtDNAs. *Am J Hum Genet, 53*(3), 563-590.

Vésteinsson, O. (1998). Patterns of Settlement in Iceland: A study in Prehistory. *Saga-Book of the Viking Society for Northern Research, 25*(1), 1-29.

Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K., & Wilson, A. C. (1991). African populations and the evolution of human mitochondrial DNA. *Science, 253*(5027), 1503-1507.

Volodko, N. V., Starikovskaya, E. B., Mazunin, I. O., Eltsov, N. P., Naidenko, P. V., Wallace, D. C., et al. (2008). Mitochondrial genome diversity in arctic Siberians, with particular reference to the evolutionary history of Beringia and Pleistocenic peopling of the Americas. *Am J Hum Genet, 82*(5), 1084-1100.

Vona, G., Falchi, A., Moral, P., Calo, C. M., & Varesi, L. (2005). Mitochondrial sequence variation in the Guahibo Amerindian population from Venezuela. *Am J Phys Anthropol, 127*(3), 361-369.

Wang, S., Lewis, C. M., Jakobsson, M., Ramachandran, S., Ray, N., Bedoya, G., et al. (2007). Genetic variation and population structure in native Americans. *PLoS Genet, 3*(11), e185.

Waters, M. R., & Stafford, T. W., Jr. (2007). Redefining the age of Clovis: implications for the peopling of the Americas. *Science, 315*(5815), 1122-1126.

Westley, K., & Dix, J. (2008). The Solutrean Atlantic Hypothesis: A view from the ocean. *Journal of the North Atlantic, 1*, 85-98.

Wijsman, E. (1984). Techniques for estimating genetic admixture and applications to the problem of the origin of the Icelanders and the Ashkenazi Jews. *Human Genetics, 67*, 441-448.

Williamsen, E. A. (2005). Boundaries of Difference in the Vínland Sagas. *Scandinavian Studies, 77*(4), 451-478.

Þorláksson, H. (2001). The Vínland Sagas in Contemporary Light. In A. Wawn & Sigurdardóttir (Eds.), *Approaches to Vínland* (pp. 63-77). Reykjavík: Sigurdur Nordal Istitute.

Þorsteinsson, B., & Jónsson, B. (1991). *Íslands Saga*. Reykjavík: Sögufélagið.