



---

Publicly Accessible Penn Dissertations

---


Spring 5-16-2011

# A Molecular Anthropological Study of Altaian Histories Utilizing Population Genetics and Phylogeography

Matthew Dulik

University of Pennsylvania, [dulik@sas.upenn.edu](mailto:dulik@sas.upenn.edu)

Follow this and additional works at: <http://repository.upenn.edu/edissertations>

 Part of the [Archaeological Anthropology Commons](#), [Biological and Physical Anthropology Commons](#), [Genetics Commons](#), and the [Molecular Genetics Commons](#)

---

## Recommended Citation

Dulik, Matthew, "A Molecular Anthropological Study of Altaian Histories Utilizing Population Genetics and Phylogeography" (2011). *Publicly Accessible Penn Dissertations*. 1545.  
<http://repository.upenn.edu/edissertations/1545>

This paper is posted at ScholarlyCommons. <http://repository.upenn.edu/edissertations/1545>  
For more information, please contact [libraryrepository@pobox.upenn.edu](mailto:libraryrepository@pobox.upenn.edu).

---

# A Molecular Anthropological Study of Altaian Histories Utilizing Population Genetics and Phylogeography

## **Abstract**

This dissertation explores the genetic histories of several populations living in the Altai Republic of Russia. It employs an approach combining methods from population genetics and phylogeography to characterize genetic diversity in these populations, and places the results in a molecular anthropological context. Previously, researchers used anthropological, historical, ethnographic and linguistic evidence to categorize the indigenous inhabitants of the Altai into two groups – northern and southern Altaians. Genetic data obtained in this study were therefore used to determine whether these anthropological groupings resulted from historical processes involving different source populations, and if the observed geographical and anthropological separation between northern and southern Altaians also represented a genetic boundary between them. These comparisons were made by examining mitochondrial DNA (mtDNA) coding region single nucleotide polymorphisms (SNPs), control region sequences (including HVS1), and several complete mitochondrial genomes. Variation in the non-recombining portion of the Y-chromosome (NRY) was characterized with biallelic markers and short tandem repeat (STR) haplotypes. Overall, this work provided a high-resolution data set for both unipaternally inherited genetic marker systems. The resulting data were analyzed using both population genetic and phylogeographic methods. Northern Altaians (Chelkan, Kumandin and Tubalar) were distinctive from the southern Altaians (Altai-kizhi) with both genetic systems, yet the Tubalar consistently showed evidence of admixture with southern Altaians, reflecting differences in the origin and population history of northern and southern groups as well as between ethnic northern Altaian populations. These results complement the observation of cultural differences as noted by anthropological/ethnographic research on Altaian populations. These differences likely reinforced and maintained the genetic differences between ethnic groups (i.e., a cultural barrier to genetic exchange). Therefore, biological and cultural lines of evidence suggest separate origins for northern and southern Altaians. Phylogeographic analysis of mtDNA and NRY haplotypes examined the impact of different historical events on genetic diversity in Altaians, including Neolithic expansions, the introduction of Kurgan cultures, the spread of Altaic-speakers, and the intrusion of the Mongol Empire. These insights also allowed for a greater understanding of the peopling of Siberia itself. The cultures of Altaian peoples ultimately helped to shape their current genetic variation.

## **Degree Type**

Dissertation

## **Degree Name**

Doctor of Philosophy (PhD)

## **Graduate Group**

Anthropology

## **First Advisor**

Theodore G. Schurr

---

**Keywords**

mtDNA, Y-chromosome, genetic history, molecular anthropology

**Subject Categories**

Archaeological Anthropology | Biological and Physical Anthropology | Genetics | Molecular Genetics

**A MOLECULAR ANTHROPOLOGICAL STUDY OF ALTAIAN HISTORIES  
UTILIZING POPULATION GENETICS AND PHYLOGEOGRAPHY**

Matthew C. Dulik

A DISSERTATION

in

Anthropology

Presented to the Faculties of the University of Pennsylvania

in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy


2011

*Dissertation Supervisor*



Theodore G. Schurr  
Associate Professor, Anthropology

*Graduate Group Chairperson*



Deborah A. Thomas  
Associate Professor, Anthropology

*Dissertation Committee*

Janet M. Monge, Adjunct Associate Professor, Anthropology  
Victor H. Mair, Professor of Chinese Language and Literature, East Asian Languages and  
Civilizations

A MOLECULAR ANTHROPOLOGICAL STUDY OF ALTAIAN HISTORIES  
UTILIZING POPULATION GENETICS AND PHYLOGEOGRAPHY

COPYRIGHT

2011

Matthew C. Dulik

*To my parents*

## **Acknowledgments**

First of all, I need to thank my dissertation advisor, Theodore Schurr. He allowed me and others in his lab the chance to explore our ideas (the plausible and farfetched) under his auspices, giving us a secure environment to expand our theoretical and technical expertise without getting lost along the way. Aside from my dissertation, the opportunities he afforded to me (both in range and scale of research topics) gave me numerous occasions to broaden my experience as a molecular anthropologist. Under his direction, I have learned what it takes to manage multiple projects and run a busy lab simultaneously. Our exchanges in the field, in grant and publication writing and in everyday interactions will be lessons I carry with me throughout my career.

This dissertation would not have been possible had it not been for my dissertation committee members, Janet Monge and Victor Mair. I sincerely appreciate their continued support and diligence in reading every page I gave them. Our discussions helped to shape my ideas more clearly, challenged my perspective and, as a result, strengthened the arguments I wished to make. Undoubtedly, their efforts made for a stronger, more polished dissertation. It was an absolute pleasure working with them. I also owe my deepest gratitude to my first advisor, Richard Zettler, who provided me the opportunity to come to Penn and encouraged my interests in combining Near Eastern archaeology and molecular genetics. I will never forget the patience he showed me and his continued guidance and honesty throughout my time here.

None of this research would have been possible without the financial support of numerous institutions and agencies. In particular, the Department of Anthropology and the University of Pennsylvania, the Baikal Project and the Social Sciences and

Humanities Research Council of Canada, the Russian Basic Fund for Research, the National Geographic Society and the National Science Foundation for my Doctoral Dissertation Improvement Grant. In addition, I want to extend my thanks to the Altaian participants who volunteered to collaborate with us on this project.

I would like to express my appreciation to Joe Lorenz for helping to facilitate a goal of mine to blend my interests in archaeology and human genetics through our ancient DNA work (and for helping me gain a better appreciation of “the lab where things work”). I would also like to thank Art Washburn, first and foremost, for great friendship and camaraderie. The opportunity to teach anatomy and share experiences in forensic consulting with him allowed me to apply my knowledge and training as an anthropologist outside of the Department, which has been just as fulfilling as any of my other endeavors at Penn.

I am grateful for my friends and fellow graduate students (from my year and others) as well as current and past members of the Schurr lab who have helped make this process more rewarding, especially Samara Rubeinstein, Ömer Gökçümen, Paul Babb, Emily Renschler, and Klara Stefflova. Early on in our graduate school careers, Samara, Ömer and I brainstormed together and challenged each other on everything molecular anthropology, from grandiose theories to the minuscule details of statistical analysis. I still see the result of much of those collective efforts in our bodies of scholarship today. Ömer eloquently characterized the relationship that the three of us shared as that of intellectual siblings – I cannot convey this mutual sentiment more precisely.

Words alone cannot express the thanks I owe to my fiancée, Kim Kahle, who has been amazing throughout this whole process. Her encouragement was constant, and her



impeccable advice was given freely and always at the right time. She possessed seemingly endless patience and tolerance during all the working weekends, the inordinate stacks of books and papers, the inescapable attachment to my computer, and the many late-night lab runs. Having received her PhD several years ago, she knew what she was getting into, but despite everything, it didn't scare her away, and I am wholly grateful for that. I don't know how I would have made it through this process without her, and I can't wait to embark on the next phase of our lives together.

Finally, I owe everything to my family, especially my parents, brothers and sisters. There is no way that I would be where I am today without their unfaltering support and constant encouragement. They sacrificed to get me to this point as well and have done everything from proof-reading and computer support to providing required distractions. Most graciously, they stopped asking the dreaded question, "so when are you going to finish?" quite some time ago. My parents have provided me the best models to live my life and have always been there for me. None of this would have been possible without them, which is why I dedicate my dissertation to Anna Marie and Dan Dulik.

# **ABSTRACT**

## **A MOLECULAR ANTHROPOLOGICAL STUDY OF ALTAIAN HISTORIES UTILIZING POPULATION GENETICS AND PHYLOGEOGRAPHY**

Matthew C. Dulik

Theodore G. Schurr

This dissertation explores the genetic histories of several populations living in the Altai Republic of Russia. It employs an approach combining methods from population genetics and phylogeography to characterize genetic diversity in these populations, and places the results in a molecular anthropological context. Previously, researchers used anthropological, historical, ethnographic and linguistic evidence to categorize the indigenous inhabitants of the Altai into two groups – northern and southern Altaians. Genetic data obtained in this study were therefore used to determine whether these anthropological groupings resulted from historical processes involving different source populations, and if the observed geographical and anthropological separation between northern and southern Altaians also represented a genetic boundary between them. These comparisons were made by examining mitochondrial DNA (mtDNA) coding region single nucleotide polymorphisms (SNPs), control region sequences (including HVS1),

and several complete mitochondrial genomes. Variation in the non-recombining portion of the Y-chromosome (NRY) was characterized with biallelic markers and short tandem repeat (STR) haplotypes. Overall, this work provided a high-resolution data set for both unipaternally inherited genetic marker systems. The resulting data were analyzed using both population genetic and phylogeographic methods. Northern Altaians (Chelkan, Kumandin and Tubalar) were distinctive from the southern Altaians (Altai-kizhi) with both genetic systems, yet the Tubalar consistently showed evidence of admixture with southern Altaians, reflecting differences in the origin and population history of northern and southern groups as well as between ethnic northern Altaian populations. These results complement the observation of cultural differences as noted by anthropological/ethnographic research on Altaian populations. These differences likely reinforced and maintained the genetic differences between ethnic groups (i.e., a cultural barrier to genetic exchange). Therefore, biological and cultural lines of evidence suggest separate origins for northern and southern Altaians. Phylogeographic analysis of mtDNA and NRY haplotypes examined the impact of different historical events on genetic diversity in Altaians, including Neolithic expansions, the introduction of Kurgan cultures, the spread of Altaic-speakers, and the intrusion of the Mongol Empire. These insights also allowed for a greater understanding of the peopling of Siberia itself. The cultures of Altaian peoples ultimately helped to shape their current genetic variation.

# Table of Contents

Abstract .....	vii
List of Tables .....	xiv
List of Figures .....	xvi
Introduction.....	1
Dissertation Approach .....	4
Project Aims and Hypotheses .....	7
Overview of the Dissertation .....	11
Chapter 1: Archaeology, History and Genetics: the Altaian Context.....	13
1.1 First Inhabitants of Siberia.....	13
1.2 The Neolithic in Siberia: From Stone to Pottery .....	18
1.2.1 Baikal Neolithic .....	19
1.2.2 Neolithic in Western Siberia.....	21
1.3 Bronze Ages of Siberia: Pastoralism and Metallurgy.....	22
1.4 From Nomads of the Iron Age to Nomads of Today .....	29
1.5 Craniometry and Genetics: Prospects of Population Affinities .....	39
Chapter 2: Background to Methods .....	52
2.1 The Mitochondrial Genome.....	52
2.1.1 Non-recombination and Maternal Inheritance .....	54
2.1.2 Characteristics of Mutational Change: Clocklike Rates and Time Dependency.....	56
2.1.3 Natural Selection on the mtDNA genome .....	61
2.1.4 Climate and Positive Selection on mtDNA .....	70

2.2 The Y-chromosome .....	82
2.2.1 Y-STR Mutation Rates .....	86
2.2.2 Selection on the Y-chromosome .....	87
2.3 Chapter Conclusions .....	94
Chapter 3: Materials and Methods .....	97
3.1 Sample Collection .....	97
3.2 DNA Extraction .....	100
3.3 MtDNA Characterization .....	101
3.4 NRY Characterization .....	105
3.5 Statistical Analysis .....	109
3.5.1 Population Genetic Statistics – Within Population Estimates .....	109
3.5.2 Population Genetic Statistics – Between Population Estimates .....	114
3.5.3 Phylogenetics – mtDNA Data .....	116
3.5.4 MtDNA Coalescence Estimates .....	121
3.5.5 NRY Phylogenetics .....	123
Chapter 4 Mitochondrial DNA and Population Histories .....	127
4.1 Northern Versus Southern Altaian Genetic Variation – Haplogroup Level .....	129
4.2 Northern Altaian Genetic Variation .....	130
4.3 Haplotype-Sharing among Altaians .....	131
4.4 Genetic Structure in Altaian Populations .....	138
4.5 Within Population Variation .....	150
4.6 Altaian Local Context .....	153

4.7	Altaiian Diversity - Haplogroup Level .....	155
4.8	Altaiian Populations – Haplotype Level .....	157
4.9	Southern Siberians – Haplogroup Level .....	160
4.10	Southern Siberians – Haplotype Level .....	163
4.11	Altaiians from a Global Perspective .....	169
4.12	Chapter Conclusions .....	172
Chapter 5: Phylogeography of mtDNA Haplogroups.....		176
5.1	Basal mtDNA Haplogroup Phylogenies .....	177
5.2	Assessment of Natural Selection on mtDNA Haplogroups.....	184
5.3	$K_A/K_S$ Ratios .....	185
5.4	dN/dS Analysis .....	188
5.5	Tests for Clocklike Behavior .....	190
5.6	Coalescence Dating and Phylogeography.....	195
5.7	Haplogroup C – Coalescence Dating and Phylogeography.....	196
5.8	Haplogroup D – Coalescence and Phylogeography.....	209
5.9	Haplogroup U4 – Coalescence Dating.....	218
5.10	Haplogroup U4 - Phylogeography .....	221
5.11	Haplogroup U5 – Coalescence Dating.....	223
5.12	Haplogroup U5 - Phylogeography .....	226
5.13	Chapter Conclusions .....	227
Chapter 6: NRY Variation and Population Histories.....		232
6.1	Northern Versus Southern Altaian NRY Variation – Haplogroup .....	234
6.2	Northern Altaian NRY Haplogroup Variation.....	236

6.3 Haplotype-Sharing Among Altaians.....	236
6.4 Genetic Structure among Altaian Populations.....	239
6.4.1 Haplogroup Analysis .....	239
6.4.2 Haplotype Analysis.....	240
6.5 Within Population Variation.....	246
6.6 Altaian Local Context.....	246
6.6.1 Altaian Diversity – Haplogroup (Biallelic Marker) Analysis.....	246
6.7 South Siberian and Regional Contexts .....	253
6.8 Altaians from a Global Perspective 6.3 Concluding Remarks .....	256
6.9 Population Haplotype Analysis.....	259
6.10 Chapter Conclusions .....	266
Chapter 7: NRY Phylogeography .....	271
7.1 Haplogroup N.....	272
7.1.1 Haplogroup N1b.....	274
7.2 Haplogroup R1.....	287
7.2.1 Haplogroup R1a1a .....	288
7.2.2 Haplogroup R1b1b1.....	298
7.3 Haplogroup Q1a3*.....	303
7.4 Haplogroup C3* and C3c.....	309
7.5 Chapter Conclusions .....	317
Conclusions.....	322
Appendix 1: Comparative Datasets for mtDNA and NRY Analyses .....	328
Appendix 2: Altaian Y-STR Data.....	333

Bibliography .....338



## List of Tables

Table 3.1	PCR-RFLP and deletion tests for mtDNA SNPs.....	102
Table 3.2	Control region amplification primers .....	104
Table 3.3	Control region sequencing primers.....	104
Table 3.4	NRY TaqMan assays .....	107
Table 3.5	NRY marker sequencing reactions .....	108
Table 4.1	MtDNA haplogroup frequencies for Altaian populations .....	128
Table 4.2	Haplotype-sharing among Altaian populations .....	133
Table 4.3	AMOVA of northern versus southern Altaian villages .....	141
Table 4.4	AMOVA of northern Altaian villages .....	146
Table 4.5	AMOVA of northern Altaian ethnic group.....	147
Table 4.6	AMOVA of northern Altaian ethnic groups (without small populations).....	148
Table 4.7	$F_{ST}$ values between Altaian ethnic groups .....	149
Table 4.8	Summary statistics for Altaian ethnic groups.....	151
Table 4.9	AMOVA of southern Siberians .....	168
Table 5.1	$K_A/K_S$ ratios for haplogroup C.....	186
Table 5.2	$K_A/K_S$ ratios for haplogroup D .....	187
Table 5.3	$K_A/K_S$ ratios for haplogroup U4 .....	187
Table 5.4	$K_A/K_S$ ratios for haplogroup U5 .....	188
Table 5.5	Coalescent estimates for haplogroup C .....	198
Table 5.6	Coalescent estimates of haplogroup C branches.....	202
Table 5.7	Coalescent estimates for haplogroup D .....	210

Table 5.8	Coalescent estimates of haplogroup D branches from complete genomes....	211
Table 5.9	Coalescent estimates for haplogroup U4 .....	220
Table 5.10	Coalescent estimates of haplogroup U4 branches .....	220
Table 5.11	Coalescent estimates for haplogroup U5 .....	225
Table 5.12	Coalescent estimates of haplogroup U5 branches .....	226
Table 6.1	High-resolution haplogroup frequencies of Altaian populations.....	233
Table 6.2	AMOVA of northern versus southern Altaian villages .....	242
Table 6.3	AMOVA of northern Altaians .....	245
Table 6.4	$R_{ST}$ values between Altaian ethnic groups.....	245
Table 6.5	Summary statistics for Altaian NRY variation.....	246
Table 6.6	Frequencies of 10-haplogroup profiles for Altaian populations.....	248
Table 6.7	Haplogroup diversities in Altaian populations with 10-haplogroup profiles .....	250
Table 6.8	AMOVA results among Altaian populations.....	252
Table 7.1	Intrapopulation variances of N1b haplotypes .....	282
Table 7.2	Intrapopulation variances for R1a1a haplotypes .....	290
Table 7.3	Intrapopulation variance for R1b1b1 haplotypes.....	302
Table 7.4	Intrapopulation variances for Q haplotypes.....	308
Table 7.5	Intrapopulation variance for C3* and C3c1 haplotypes .....	316

## List of Figures

Figure I.1	Map of southern Siberia .....	2
Figure 3.1	Locations of population sample collection in the Altai Republic .....	98
Figure 4.1	MDS plot of $F_{ST}$ values for Altaian villages .....	140
Figure 4.2	MDS plot of $F_{ST}$ values for each ethnic group by village .....	143
Figure 4.3	MDS plot of $F_{ST}$ values - ethnic groups and villages .....	145
Figure 4.4	Mismatch distributions of Altaian ethnic groups .....	153
Figure 4.5	MDS plot of Altaian ethnic group $F_{ST}$ values .....	158
Figure 4.6	MDS plot of Altaian ethnic group $F_{ST}$ values in three dimensions .....	159
Figure 4.7	Principal component analysis of Siberian populations (1st and 2nd components) .....	162
Figure 4.8	Principal components analysis of Siberian populations (1st and 3rd components) .....	163
Figure 4.9	MDS plot of $F_{ST}$ values for Siberian populations .....	165
Figure 4.10	MDS plot of $F_{ST}$ values for Siberian and Central Asian populations .....	171
Figure 5.1	Basal mtDNA haplogroup phylogenies .....	178
Figure 5.2	RM-MJ network of complete haplogroup C mtDNA genomes .....	199
Figure 5.3	RM-MJ network of only haplogroup C synonymous mutations .....	199
Figure 5.4	RM-MJ network of haplogroup C HVS1 from 16090-16365 .....	200
Figure 5.5	RM-MJ network of haplogroup C HVS1 from 16051-16400 .....	200
Figure 5.6	RM-MJ network of haplogroup C HVS2 from 68-263 .....	201
Figure 5.7	RM-MJ network of haplogroup C complete control regions .....	201
Figure 5.8	RM-MJ network of complete haplogroup D mtDNA genomes .....	210

Figure 5.9	RM-MJ network of complete haplogroup U4 mtDNA genomes .....	219
Figure 5.10	RM-MJ network of complete haplogroup U5 mtDNA genomes .....	224
Figure 6.1	MJ-RM network of northern Altaian NRY lineages .....	237
Figure 6.2	RM-MJ network of all Altaian NRY lineages.....	239
Figure 6.3	MDS plot of village $R_{ST}$ estimates villages.....	241
Figure 6.4	MDS plot of $F_{ST}$ values - ethnic groups and villages .....	243
Figure 6.5	MDS plot of conventional $F_{ST}$ values among Altaian populations .....	251
Figure 6.6	MDS plot of conventional $F_{ST}$ values for southern Siberian populations ....	255
Figure 6.7	MDS plot of $F_{ST}$ values for Siberian and Central Asian populations.....	257
Figure 6.8	MDS plot of $R_{ST}$ values for Siberian populations .....	261
Figure 7.1	RM-MJ network of haplogroup N (7-STRs).....	273
Figure 7.2	RM-MJ network of haplogroup N1b (7-STR).....	276
Figure 7.3	RM-MJ network of haplogroup N1b (10-STR).....	277
Figure 7.4	RM-MJ network of haplogroup N1b (15-STR).....	280
Figure 7.5	RM-MJ network of haplogroup R1a1a (7-STR) .....	291
Figure 7.6	RM-MJ network of haplogroup R1a1a (15-STR) .....	292
Figure 7.7	RM-MJ network of Altaian haplogroup R1a1a (15-STR) .....	296
Figure 7.8	RM-MJ network for haplogroup R1b1b1 (8-STR) .....	301
Figure 7.9	RM-MJ network of haplogroup Q (5-STR).....	306
Figure 7.10	RM-MJ network of haplogroup Q (8-STR).....	306
Figure 7.11	RM-MJ network for haplogroup C3* (7-STR) .....	315
Figure 7.12	RM-MJ network of haplogroup C3c1 (9-STR).....	317

## Introduction

To many, the mere mention of the word “Siberia” conjures images of bleak, desolate, harsh, and cold places. Some may think of punishing, primitive and poor living conditions. However, at its core, Siberia contains a history that is rich in character, intricacy and sheer volume (Forsyth, 1992; Gryaznov, 1969; Levin & Potapov, 1964; Okladnikov, 1964; Rudenko, 1970). At times, it was a frontier for a newly emerged species (Goebel, 1999; Okladnikov, 1964, 1990). Later, it belonged to vast cultural horizons that spanned two continents (Anthony, 2007; Golden, 1992; Kuzmina & Mair, 2008; Mallory, 1989). It even serves as a homeland – whether real or mythical – of languages, religion and peoples (Forsyth, 1992; Golden, 1992).

The Altai region of southern Siberia, in particular, has played a crucial role in human history (Figure I.1). It has served as an entry point into Siberia from Central Asia (and vice versa), as a junction for people of different cultures and civilizations for millennia, and as a possible homeland for Turkic-speaking populations (Forsyth, 1992; Golden, 1992; Levin & Potapov, 1964; Okladnikov, 1964). Thus, much of Eurasian history involves people living, or passing through, this region.

Since the Russian presence in this region, Altaian populations have been classified into northern and southern groups based on geographic, ethnographic, linguistic, historical and phenotypic evidence (Potapov, 1962, 1964a). The numerous southern Altaian populations consist of Altai-kizhi, Telenghits, Teleuts and Telesy, but were originally called Oirats or Kalmyks by the Russians, as they had closer ties to western Mongol Dzungars in the 17<sup>th</sup> and 18<sup>th</sup> centuries (Forsyth, 1992; Wixman, 1984). These Altaians live in the more mountainous southern region of the Altai and have

greater cultural similarities to groups in the Central Asian steppe. Southern Altaians generally practiced a pastoral nomadic lifestyle, with stock breeding as a major economic activity (Potapov, 1962, 1964a). Studies of cranial morphology show southern Altaians have a greater affinity with the “Central Asian” group of Siberians (Potapov, 1962, 1964a).



Figure I.1 Map of southern Siberia

The northern Altaians are also constituted by several ethnic groups, including Chelkans, Tubalars, and Kumandins (Forsyth, 1992; Potapov, 1964a). At times, the Shors have also been placed in the northern Altaian grouping (Potapov, 1962). These populations are found in the taiga regions in the northern part of the Altai Republic. Originally called Back Country Tatars by the Russians, these small groups lived a hunter-gatherer subsistence and economic lifestyle (Forsyth, 1992; Potapov, 1962, 1964a). Anthropometric assessment of northern Altaians showed that they shared affinities with populations living in northwestern Siberia, especially among the Khanty, Mansi, and Ket

populations (Potapov, 1962, 1964a). Clothing, artifacts and religious rituals were all cited as evidence of a shared common ancestry among northern Altaians, Shors and the Ugric and Samoyedic populations of northwestern Siberia (Potapov, 1962).

Both northern and southern groups speak Turkic languages, although closer examination reveals differences between the geographical groupings (Comrie, 1981; Menges, 1968). Tribal names found among the southern Altaians played a fundamental role in assigning close relationships between them and other Kipchak-speaking populations in the steppe (Potapov, 1962). The northern Altaians have a greater association with Yeniseian and Samoyedic speakers, similar to Turkic-speakers in the Abakan and Tuvan regions. This is likely due to the adoption of Turkic languages by Yeniseian and southern Samoyedics (Menges, 1968). Many of the place names in the northern Altai provide evidence that the area was once inhabited by Yeniseian-speaking populations (Vajda, 2001). It is assumed that those populations assimilated with Turkic populations to form the current northern Altaian and Shor ethnic groups (Forsyth, 1992; Menges, 1968; Potapov, 1962, 1964a).

The peoples of the Altai Republic, and the history of these populations as told through their DNA, are the focus of this dissertation. It is a history that, while largely lost to the memories of its people, was passed on in their customs, languages and biology. Thus, the purpose of this dissertation is to gain a better understanding of the history of these indigenous groups. It also, in a broader sense, clarifies how populations change over time, delineates the relationships and interactions between biology and culture, and determines how culture can ultimately influence genetic variation.

## **Dissertation Approach**

In pursuing my dissertation research, I used two separate but related methodological approaches that are employed in molecular anthropological studies. The first involves population genetics and the second phylogenetics. The first utilizes the mathematical properties espoused by population genetics, which has played a fundamental role in understanding human genetic variation. Population genetics reflects the synthesis between the Darwinian theory of evolution by natural selection and the Mendelian theory of inheritance, resulting in what is now known as neo-Darwinism. The discipline was not without its debates. It developed out of disagreements about selection on continuous and discontinuous variation that began after Darwin published *On the Origin of Species*, and later became the basis for the debates between biometricians and mutationists, with the latter using Mendel's work to support their view of discontinuous evolution (Provine, 1971). Subsequently, Fisher, Wright and Haldane produced this synthesis of Darwinian and Mendelian theories through their theoretical genetic work in the 1920s. By the early 1930s, each of them had presented his own version of this synthesis, although all are founded on the same mathematical principles.

From the very beginning, population genetics has had a pluralistic character (Gould, 2002; Provine, 1971). Fisher, Wright and Haldane promoted their own interpretations, although disagreeing with each other on various aspects of the synthesis. While Fisher and Haldane believed that selection played the most influential role, Wright emphasized the effects of sampling (random genetic drift) or genic interactions (epistasis). Wright also believed the genetic background on which the alleles were present could make a mutation either beneficial or deleterious (Provine, 1971; Wright,



1930). For Fisher, the emphasis was on a single gene replacement theory (Fisher, 1930; Provine, 1971).

In the following decades, a strict adaptationist (or selectionist) approach was promulgated narrowing the scope of population genetics (Gould, 2002). According to this perspective, mutations must be selected for if they are present in a population. Thus, nearly every instance of mutation could be correlated with an associated adaptation (Gould, 2002; Mayr, 1963).

It was in this context that Kimura promoted his neutral theory of molecular evolution (Kimura, 1968) and later, the nearly neutral theory by Ohta (Ohta, 1973). The neutral theory provides a null hypothesis and a model with which to compare to population genetic results. Central to this theory is that most new mutations are effectively neutral with regard to their selective value. Selection can act on the existing pool of mutations when a mutation produces a new phenotype that alters the relative fitness or survival of the organism. Most new phenotypes are deleterious. Thus, an expectation exists where most natural selection comes in the form of purifying selection. Positive selection can occur when a particular phenotype increases the fitness or survival of an organism. While the latter is not as common as the former, it is still of critical importance in evolution.

Given this characteristic, a number of other features are expected (Kimura, 1983; Nei, 1987). First, the rate of substitution will be equal to the rate of non-deleterious mutations because most substitutions are caused by random fixation. Second, the amount of genetic diversity at a locus in a population depends on the rate of new mutations and the size of the population. Third, mutations move through a process of extinction or

fixation; polymorphisms in populations are just one phase of this process. Therefore, a mutation that is deleterious can persist in a population, but ultimately may not contribute to the overall genetic variation of a species. Fourth, “neutral alleles are not functionless genes but are generally of vital importance to the organism. A pair of alleles are called neutral if they are functionally equivalent and do not differentially affect the fitness of the organism” (Nei, 1987, 411). Finally, the neutrality of alleles depends on the effects of genetic drift in a population.

The second methodological approach employed in my dissertation uses phylogenetic analyses to infer the relationships between haplotypes and assess their geographical origins and expansions (Avice et al., 1987; Hey & Machado, 2003). Molecular phylogenetics developed from studies of taxonomy and systematics, where the major objective was to understand the genetic relationships among organisms, often in the form of phylogenies (Hillis, 1987). Today, molecular phylogenetics mostly depends on DNA sequences to construct gene trees that can be used to infer the processes involved in the evolution of a particular genetic locus.

Phylogeography is the application of gene trees to geographical distributions of DNA sequences to explore population structure (Avice et al., 1987). This methodology has been particularly successful in mitochondrial DNA (mtDNA) and Y-chromosome studies due to the high-resolution phylogenies that can be generated from these genetic systems. It must be remembered, however, that gene trees are not necessarily the same as species trees, or even population trees (Nichols, 2001).

In the 1980s, a mathematical model describing the mutational and evolutionary processes of genetic data and demographic factors was devised. This model, called the n-

coalescent, helped create a branch of applied mathematics focused on understanding population-level phenomena (Kingman, 1982). Coalescent theory is, at its core, a model that is used to assess observed data to infer the processes involved in shaping gene genealogies and genetic parameters in population models, whether it be population structure, natural selection or neutral evolution (Wakeley, 2009). While a phylogenetic-based approach relied on accurate construction of gene trees to infer patterns of genetic relationships and evolutionary change without an explicit population model, the coalescent-based approach attempts to estimate model parameters, such as population mutation rates and effective population size (Hey & Machado, 2003). Furthermore, the coalescent was derived from population genetic models, but can be used hand-in-hand with phylogenetic methods, as the two are not necessarily mutually exclusive (Knowles & Maddison, 2002).

### **Project Aims and Hypotheses**

The project has several specific aims, which dictate the hypotheses tested throughout the dissertation. The first objective is to characterize mtDNA and Y-chromosome variation in Altaian populations, as this will reveal the genetic diversity of male and female lineages in the Chelkan, Tubalar, Kumandin, Altai-kizhi and Altaian Kazakh. These data will, in turn, help to elucidate their possible roles in the settlement of southern Siberia and provide a larger, more integrated picture of genetic variation in the Altai region.

The population histories of the indigenous ethnic groups will also be compared to that of the Altaian Kazakhs. Currently, the Altaian Kazakhs and indigenous Altaians

have distinctive cultures; these may have served as a barrier to gene flow between groups. Religion plays a significant role in this respect, with Altaian Kazakhs practicing Islam and indigenous Altaians traditionally practicing some form of Shamanism and/or Buddhism. In fact, Burkhanism, which is a religion that has influences from both Buddhism and Shamanism, has been cited as a unifying force for indigenous Altaians (Halemba, 2003).

The effects of these cultural differences will be investigated with the resulting genetic data and tests of admixture. The hypothesis, then, is that genetic variation in the mtDNA and NRY will be differentiated along ethnic group membership, and that the Altaian Kazakhs will not show a great degree of similarity to indigenous Altaians. Furthermore, it is expected that geographic and/or linguistic classifications will not explain the structuring of genetic variation as well as classifications based on ethnic group membership.

The next aim is to examine the genetic diversity in relation to the tribe and clan structure of these populations. It is possible to test differences in mtDNA and NRY variation resulting from different social practices. For these populations, patrilocality is the social norm. These groups were also semi-nomadic in the recent past, and much of their tribal and clan systems are based on the relationships between fathers, sons and brothers (Potapov, 1964a). In this regard, a study of Y-chromosome markers in Central Asian populations has shown that individuals of the same lineages and clans share recent common ancestry, but no such correlation exists at the tribal level (Chaix et al., 2004).

This pattern will be tested with the NRY data from Altaian populations and compared to those obtained from the analysis of mtDNA variation, in which haplotype

sharing between ethnic groups and between villages within them, will be assessed. It is expected that marriage and residence patterns will have shaped patterns of genetic variation in the Altaian populations such that there is not a correspondence between mtDNA (maternal lineages) diversity and geography but there is between Y-chromosomes (paternal lineages) and geography. In other words, if clan membership is most important, then the NRY genetic variation will be highly differentiated, whereas that from mtDNA will not. The influence that a forced sedentary lifestyle has had on the traditional clan structures will also be assessed in this way, as the government-imposed restrictions could have affected these genetic patterns.

A third goal is to compare genetic variation in Altaian groups to that observed in Native American tribes to determine whether the Altai Region has a possible connection with indigenous American populations. This idea has been supported by mtDNA studies, which show that all of the mtDNA haplogroups present in the Americas (haplogroups A, B, C, D, and X) are also found in our sampled Altaian populations (Dulik, Zhadanov, Osipova, & Schurr, 2006). Haplogroup Q makes up the majority of Native American Y-chromosome lineages (Bortolini et al., 2002; Bortolini et al., 2003). Such lineages are also present in Siberia and Central Asia, and, thus, they will be used to elucidate the Altaians' possible contribution to New World populations. Thus, at least a subset of the mtDNA and NRY lineages in Altaians should closely match those found in Native Americans.

The final goal is part of a long-term objective for the Laboratory of Molecular Anthropology, which is the investigation of the spread of Altaic speaking populations, with a more specific focus on Turkic-speaking populations. Today, populations from

Turkey through Central Asia and into Siberia use Turkic languages. Therefore, this aim will not be fully realized with this dissertation alone. However, the genetic data collected from Siberian populations will be an integral piece of the larger puzzle that includes samples from Turkey and the Caucasus, as well as previously published data for Central Asian populations. In particular, if there was a demic diffusion linked to the spread of Turkic languages, then we would expect similar lineages to be found in all Turkic-speaking populations. Altaian populations should therefore have a subset of these lineages.

From a practical standpoint, genetic analysis of these populations provides a different perspective on a people undergoing cultural integration and assimilation. Historically, the Russian and Chinese governments have marginalized these populations. They are located in areas considered peripheral in the past several centuries and have undergone significant shifts in lifestyle and cultural practices in the past couple of decades. These populations, which at one time were nomadic hunter-gatherer groups, are now sedentary, with large settlements being formed by the merging of many hunting and gathering communities. As these populations change and different economic factors influence their decisions to maintain traditional lifestyles, or move to larger towns and cities, the imprint of the hunter-gatherer or pastoral nomadic lifestyles that they once utilized will disappear from the population history of these people (Karafet et al., 2002). Given the importance of Siberia in terms of migrations of modern human populations for the Americas, northern Asia and historical steppe populations, it is important to gain a better understanding of these historical threads before they are erased due to admixture and migration.

## **Overview of the Dissertation**

Chapter 1 explores Siberia's prehistoric and historic past from the earliest human inhabitations to the present day. This chapter provides the necessary background information to place the genetic analyses into the proper historical context. This framework was essential for understanding the development of cultural groups, the origins of indigenous archaeological horizons, and the identification of potential immigrant populations. It also serves as a starting point for integrating the genetic data – both ancient and modern – into the histories of Altaian populations.

Chapter 2 delves into the mechanical workings of the mtDNA and Y-chromosome as they relate to the neutral theory of molecular evolution. In essence, this chapter provides a justification for using these genetic systems. In doing so, this chapter first examines the mitochondrial genome and its characteristics. Molecular clocks, tests of neutrality and selection are also discussed. Similarly, the Y-chromosome is introduced and examined. Ultimately, it is determined that both genetic systems are ideal for studies of population histories.

Chapter 3 provides the methodologies used in this dissertation. First, details pertaining to the samples used – locations, numbers, etc. – are provided, followed by methods used for DNA extraction, purification and quantitation. Statistical analyses for within population and between population genetic diversity are then described. Finally, the chapter concludes with a discussion of phylogenetic methodologies.

Chapters 4 through 8 present the results and discussion of the genetic and statistical analyses. Chapter 4 focuses on the mtDNA population data, the variation found within populations, and comparisons between them. This analysis explains the

way that Altaian populations are related to one another via maternal lineages and determines how these populations fit into the broader southern Siberian context.

Chapter 5 presents the phylogenetic analysis of mtDNA data in an attempt to assess the evolution of Altaian populations from a diachronic perspective. In doing so, each relevant haplogroup is analyzed. This analysis necessitates that haplogroups are evolving in a neutral manner. Therefore, in addition to the phylogenetic analysis, this chapter includes assessments of natural selection.

Chapters 6 and 7 include the results and discussions of the Y-chromosome analysis. These chapters mirror the previous two chapters in structure and content, but focus on the paternal lineages present in Altaian populations.

Chapter 8 concludes the dissertation with a summary of findings and an assessment of the project objectives.



## **Chapter 1: Archaeology, History and Genetics: the Altaian Context**

The purpose of this chapter is to provide a context for the genetic studies undertaken with Altaian populations. In particular, it covers background information on the Altai region as understood through archaeological and historical evidence. The chapter first discusses the human presence in Siberia from the Paleolithic to the present, and briefly describes studies that have delineated genetic relationships of indigenous Siberian populations. Finally, the chapter turns to the molecular level and summarizes work done with modern Siberians since the 1970s.

This introduction necessitates a discussion of southern Siberia and, at times, Siberia as a whole. Much of the material must unavoidably be summarized as an exhaustive analysis goes beyond the scope of this dissertation. A substantial part of Altaian and southern Siberian prehistory comes from the mortuary archaeology of the region. Very few dwellings have actually been excavated and, as such, burials generally provide sufficient information regarding lifestyles, subsistence behavior, and material culture. Critical for a comprehensive understanding of Altaian population histories, this information creates the context from which the genetic data can be interpreted. It is anticipated that the genetic data will complement the archaeological, historical, linguistic and skeletal evidence to provide a coherent picture of Altaian history and facilitate a greater understanding of it in broader anthropological studies.

### **1.1 First Inhabitants of Siberia**

Archaeological analysis of Siberian material culture provides little evidence of hominin activity in southern Siberia during the Lower Paleolithic. Of the handful of sites

with claimed antiquity – some even dated to 700 kya – none are generally accepted, and artifacts from them are often of questionable context (Chlachula, 2001; Larichev, Khol'ushkin, & Laricheva, 1987). Currently, most of these data have been dismissed as being mere geofacts or were obtained from sites considered chronologically problematic (Goebel, 1999).

Human presence in the Altai during the Middle Paleolithic is far more definitive. The earliest Middle Paleolithic site is located in the western portion of the Altai Republic, Russia, at Denisova Cave and may have been inhabited as early as 130 kya (Goebel, 1999; Okladnikov, 1990). Most sites date to between 75 and 40 thousand years ago (kya) (Goebel, 1999; Vasil'ev, 1993). These sites are generally located in the mountainous regions of southern Siberia, where local raw materials were available for making stone tools. The associated lithics were classified as Mousterian in type and were produced using a Levalloisian technique (Goebel, 1999). Such tools are also similar in design and technology to others found in Mongolia, China and eastern Siberia along the Amur River (Okladnikov, 1990; Vasil'ev, 1993). The absence of local lithic varieties was typical for Middle Paleolithic assemblages.

Skeletal remains from Middle Paleolithic sites are also rare, but several teeth and long bones have been recovered (Turner, 1990). Despite the previous debate regarding whether skeletal remains are from anatomically modern humans, Neanderthals or even *Homo erectus* (Dolukhanov, Shukurov, Tarasov, & Zaitseva, 2002; Turner, 1990), the consensus now is that these remains belong to Neanderthals, who showed morphological similarities to specimens recovered from the Near East and Europe (Krause et al., 2007; Turner, 1990). Analysis of ancient mtDNA from a subadult excavated from Okladnikov

Cave in the Altai confirmed the presence of Neanderthals (Krause et al., 2007). Remains from an adult were also tested, but failed to provide evidence for Neanderthal mtDNA. Radiocarbon dates show that these remains were 5,000 years younger than the Neanderthal subadult, which was dated to about 37.2 kya.

A recent ancient DNA analysis of one phalanx from Denisova Cave provided drastically different and unexpected results (Krause, Fu et al., 2010). These remains were deposited between 48 and 30 kya, dating roughly to the Middle Paleolithic or during the transition to the Upper Paleolithic. The complete mtDNA genome from this specimen was sequenced and compared with that of Neanderthals and modern humans. The mtDNA of the Denisova hominin was unique, showing twice as much variation as there is between Neanderthals and modern humans. Assuming a 6 million year ago (mya) split between chimpanzees and modern humans, the Denisova hominin was estimated to have had a common ancestor with Neanderthals at about 1 mya (Krause, Fu et al., 2010). This finding does not prove that the phalanx came from a different species, only that the mtDNA was outside the known range of variation detected in less than 20 Neanderthal mtDNAs, even though this sample is significantly different from other Neanderthal mtDNAs from Siberia (Krause, Fu et al., 2010; Krause et al., 2007). Confirmation from nuclear loci can help to confirm this sample as an outlier, but those data have yet to be published.

The transition to the Upper Paleolithic around 40 kya is characterized by a change in lithic technology marked by an increase in blade, burin and scrapper stone tools (Dolukhanov et al., 2002; Goebel, 1999; Okladnikov, 1990; Vasil'ev, 1993). The raw material that was used in lithic manufacturing changed as well. The use of antler,

bone and ivory appeared in this period, as did artistic representations of animals, Venus figurines and jewelry, although some aspects of the early phase of the Upper Paleolithic remained similar to the Middle Paleolithic – for example, locations of sites, use of local resources, and generalized hunting (Goebel, 1999). In addition, a continuation of the use of older lithic technologies in later periods has been noted, and may be the result of developments from earlier Mousterian techniques (Okladnikov, 1990; Vasil'ev, 1993).

Climatic changes between 26 and 19 kya brought about new environments exploitable by Paleolithic humans – the mammoth steppes (Goebel, 1999). Like preceding periods, the habitation and camping sites of the middle Upper Paleolithic were concentrated in river valleys ( the Upper Angara and Upper Lena Rivers, the Yenisei River, and the Trans-Baikal region, in particular) (Vasil'ev, 1993). The middle Upper Paleolithic saw an expansion in the number of sites, including ones located farther to the north than previously, reflecting the human effort to take advantage of the large mammal herds of bison, mammoth and horse (Goebel, 1999). The sites of Malta and Buret' are typical for this period. Found at those sites are tool assemblages based on small blades, borers, small points and a pebble technology (Goebel, 1999; Okladnikov, 1964; Vasil'ev, 1993). These sites provided evidence of a sedentary lifestyle dependent on the large mammals available in the Siberian subarctic region. Semi-subterranean structures were constructed using bones of mammoth or bison and often covered with antler (Okladnikov, 1964). Other locations were used specially for butchering animals, usually one specific large mammal (for example, either mammoth or bison) (Goebel, 1999).

The last glacial maximum (LGM) occurred in Siberia during what is called the Sartan glacial and started between 19 and 18 kya (Goebel, 1999). Approaching the

LGM, pollen deposits changed from a high frequency of arboreal pollen to more grass pollen between 22 and 18 kya in different regions of Siberia (Vasilev, Kuzmin, Orlova, & Dementiev, 2002). Southern Siberia was a refuge for those populations that lived farther north, as the climate in the subarctic became inhospitable. Several researchers have noted a general reduction in the number of sites dating to the period between 22 and 18 kya and thus, a decrease in overall human population sizes (Dolukhanov et al., 2002; Goebel, 1999), although others claim that such an interpretation is not warranted given the radiocarbon evidence (Kuzmin & Keates, 2005; Vasilev et al., 2002). Certainly, climatic conditions played a fundamental role in determining the extent of habitable locations throughout the Paleolithic (Chlachula, 2001).

The general trend in Paleolithic Siberia is one of continuity among lithic technologies and types. While the earliest colonization of Siberia is still being debated, a hominin presence in the Altai as early as 45 kya is clear. Neanderthals were likely the first inhabitants, while anatomically modern humans arrived shortly thereafter. There is little evidence to suggest *Homo erectus* lived in southern Siberia, although it is certainly not out of the question. The enigmatic aDNA results from the Denisova hominin suggest high levels of mtDNA diversity among these initial residents. At this point, the data are insufficient to know conclusively what kinds of interactions occurred among these hominin populations. It is clear, however, that modern Siberians do not contain either mtDNAs or Y-chromosomes from Siberia's earliest populations. Furthermore, dental characteristics of these ancient groups are not found among Siberia's modern inhabitants (Turner, 1990). The exact role that Paleolithic anatomically modern humans may have played in the settlement of Siberia will be explored in later chapters.

The Altai region witnessed the same trend in numbers and locations of sites throughout the Paleolithic, except during the middle Upper Paleolithic where sites were lacking (Vasil'ev, 1993). The density of middle Upper Paleolithic sites shifted to the north and east (the mammoth steppes). As the LGM set in, populations shifted back to southern Siberia, increasing the number of sites in the Altai again. With warmer temperatures at the end of the LGM, the late Upper Paleolithic was marked by the reclaiming of inhabitation areas in central and northern Siberia, resulting in a continuous human presence that has lasted to the present. Microblade technology emerged during this period, but some of the older lithic-making techniques were maintained (Okladnikov, 1990; Vasil'ev, 1993).

## **1.2 The Neolithic in Siberia: From Stone to Pottery**

The Neolithic in Siberia is characterized as a fundamental change in human subsistence patterns, resource usage, and their worldviews (Okladnikov, 1990). Overall, these communities shared a common suite of traits. The early Neolithic in Siberia was generally seen as a continuation of the previous Upper Paleolithic cultures, although new artifact types emerged despite the continued use of microliths, flint and Levallois technologies. The most fundamental of these were the development of ceramic technologies and the use of the bow (Okladnikov, 1990). Regional variations of these technologies soon appeared, and western and eastern Siberia tended to follow different stylistic trajectories, with the Yenisei River being the rough boundary between them (Chard, 1958). These differences were observed in ceramic use, lithic technology, and ornamentation and sedentary versus hunter-forager lifestyles.

Much more is known of this period from the Baikal region than from any other area of Siberia (Weber, Katzenberg, & Schurr, 2010). As such, the Baikal cultures play an important role in placing the other Neolithic cultures of Siberia into relative and absolute chronologies. Therefore, a brief description of these material cultures is warranted.

### 1.2.1 Baikal Neolithic

The Neolithic in the Baikal area is characterized by essentially two phases – the Kitoi and Isakovo / Serovo-Glazkovo (Weber, Katzenberg et al., 2010). The first represents an early Neolithic culture, while the latter is a late Neolithic / early Bronze Age culture – both of which were first classified from cemeteries found in the region (Weber, Link, & Katzenberg, 2002). Researchers have noted a gap in the archaeological record between these periods, during which there is an absence of mortuary archaeological material (Weber et al., 2002). Radiocarbon dating of human bone from burials in cemeteries of both phases provided the absolute chronology for the region. The calibrated ages for the cultural phases are 8000-7000/6800 BP for the Kitoi, 6000/5800-5200 BP for the Isakovo/Serovo, and 5200/5000-4000 BP for the Glazkovo (Weber, McKenzie, & Beukens, 2010).

The Kitoi culture seems to have originated on the southern coast of Lake Baikal and/or Angara River areas, which possessed the oldest dates for the Kitoi (Weber, McKenzie et al., 2010). Pit graves without stone architecture were characteristic for the Kitoi cemeteries, and bodies were typically laid extended in supine position and oriented in rows facing north (Bazaliiskii, 2010). Most are single interments, but regional

variation was found. Sometimes crania were removed, indicating that either bodies were initially interred elsewhere or bodies were exposed for some time before burial in a cemetery. Grave goods were found more often among men, but the amounts varied among burials (Bazaliiskii, 2010). The most common artifacts were fishhooks, arrowheads, bone or antler harpoons and a variety of other lithics (axes, knives and scrapers), while the art and ornamentation of pendants and utensils often included representations of elk heads (Bazaliiskii, 2010). Burials from other areas of the Baikal region (Upper Lena River and Small Lake Sites) were differentiated from the classic Kitoi with regards to body position (some arranged on their side and all had flexed legs in eastern Baikal sites) and stone use for burial construction (in Upper Lena sites) (Bazaliiskii, 2010). Both Upper Lena and eastern Baikal sites lacked the distinctive Kitoi fishhooks and arrowheads. In addition, little to no ceramics were found in any Kitoi cemetery.

The late Neolithic Isakovo/Serovo culture appeared region-wide at roughly 5900 BP (Weber, McKenzie et al., 2010). Archaeological investigations of burials dating to the late Neolithic provide evidence of multiple burial cultures coexisting in the Baikal region. For the Isakovo and Serovo, these cultures even shared the same cemeteries. In all cases, the use of stones for building burials was common to the different cultures. Ritual fires were typical in the Isakovo and Ol'khon Group of the late Serovo, but not in the other Serovo cultures (Bazaliiskii, 2010). Bodies tended to be extended in the supine position, with most being single burials, but heads were positioned upstream of the Angara for the Isakovo, and bodies were perpendicular to the river for Serovo. Of the Isakovo burials, 60% possessed remains of children (Bazaliiskii, 2010). Grave goods



often consisted of a variety of points, harpoons and ceramics, but the ceramics differed between the Isakovo and Serovo cultures. In addition, the Serovo burials usually had bifacial spearheads (Bazaliiskii, 2010). The artifacts indicated a switch to hunting, with a lesser emphasis on fishing, appearing to be the main source of subsistence for the Kitoi.

### 1.2.2 Neolithic in Western Siberia

In comparison to what is known of the early Neolithic from the Baikal area, archaeological evidence from the Altai region is lacking (Jettmar, 1951). The little material that has been described seemed to share similarities in lithic and ceramic types with the Kelteminar culture found in Kazakhstan and western Siberia. One burial had both Kelteminar lithics and fishhooks very similar to those found in the Kitoi burials of the Baikal region (Jettmar, 1951). Thus, it seems that during the Neolithic, the Altai region represented an intermediate zone between the Baikal and Kelteminar steppe cultures, like that of the Minusinsk and the Krasnoyarsk region to the north of the Altai.

Despite the overall scarcity of archaeological material from the late Neolithic in southern Siberia, one point is clear – local variations in material cultures began to take shape. There are at least four or five such varieties identified in and around southwestern Siberia (Bobrov, 1988). Ceramics show influences in stylistic decorations from the Kelteminar and the Serovo (Okladnikov, 1990), which is why these archaeological complexes were assigned to the late Neolithic. In the Altai, however, little evidence for the Neolithic has been found. Of the few artifacts that have been studied, connections with both the Baikal and Ural regions have been noted (Chard, 1958). Therefore, at this time, the evidence is far too sparse to make conclusive remarks.

The transition from early to late Neolithic witnessed the development and differentiation of cultures in western and eastern Siberia (Chard, 1958). From the Ural Mountains to the Yenisei River, sedentary or semi-sedentary hunter-forager groups eventually came to share a similar style of comb-marked pottery (Okladnikov, 1990). Some of this pottery included stylized water and ducks containing motifs that have been associated with later Finno-Ugric speaking cultures, even though these artifacts are thousands of years older than the emergence of historically attested Finno-Ugric groups (Okladnikov, 1990). From the Yenisei River to the Baikal region, the material culture also continued to have similar connections to its Paleolithic past. From the Mesolithic Khin to later Neolithic Isakovo and Serovo periods, pottery became more developed, with increased ornamentation and varying styles (Okladnikov, 1990; Weber et al., 2002). Unlike the groups in the western Siberia, however, populations in the Baikal region appeared to be more mobile. No settlements dating to the Neolithic have yet been uncovered in the Baikal (Weber, Katzenberg et al., 2010).

### **1.3 Bronze Ages of Siberia: Pastoralism and Metallurgy**

The cultures appearing in the next prehistoric phase of southern Siberia mark a pivotal change in the development of Siberian cultures, as they exhibit a distinctive suite of characteristics and new technological elements. The first of these periods was assigned to the Eneolithic – comparable to the Copper Age or Chalcolithic in the Near East. The culture has been named the Afanasievo Culture after the site in which the first burials were excavated in the 1920s – Afanasieva Gora (Gryaznov, 1969). This period began at roughly 3500 BCE (Anthony, 2007). The Afanasievo burials provide a

distinguishable break in the continuity of material culture witnessed in Siberia up to this point in time. The burials were of the kurgan mound style, which is generally identifiable as a stone or earthen mound of varying size that was popular throughout Eurasia during the Eneolithic, Bronze and Iron Ages (Anthony, 2007; Gimbutas & Hencken, 1956; Mallory, 1976, 1977, 1989). All used stone slabs to cover inhumations, with smaller stones circling around the mound's perimeter. Bodies were often laid on their backs with the knees bent, and ochre was used extensively (Jettmar, 1951). Multiple burials were not typical in the Altai, but some were found in Minusinsk (Gryaznov, 1969).

Most of the Afanasievo burials were found in Minusinsk, with only a fraction being found in the (northern) Altai (Gryaznov, 1969; Jettmar, 1951). Therefore, the Afanasievo was clearly an intrusive cultural complex. It first appeared in the Altai on the previously uninhabited Ukok plateau and was built on virgin soil. Antecedents have been identified in the Repin material culture from the Volga-Ural region (Anthony, 2007). Similarities with the Yamnaya or Pit-Grave culture in the western steppe lands have also been identified (Anthony, 2007; Kuzmina & Mair, 2008; Mallory, 1989). Both cultures are based on the use of a pastoral economy probably facilitated by horseback riding and wagons (Anthony, 2007). As such, the Afanasievo introduced domesticated cattle, sheep and goat to southern Siberia. Evidence of both domesticated and wild animals was found in burials, indicating that hunting was still an important means of obtaining food (Gryaznov, 1969).

Along with domesticated animals came the use of metals. Copper wire was mostly used to repair wooden implements, but needles, awls, small knives and curved fishhooks have also been found (Gryaznov, 1969; Jettmar, 1951; Okladnikov, 1964,

1990). Sledges for mining the metals were found in burials in the Altai, indicating that metals were mined from local sources (Jettmar, 1951). The copper implements were used alongside lithic tools that showed resemblances to the flaked and retouched Neolithic tools produced by a new technique of selective hammering called *tochechnaya tekhnika* by Russian archaeologists (Gryaznov, 1969). In this way, the Eneolithic resembled that of the preceding period (Okladnikov, 1964, 1990). Although the impetus for such a long migration cannot be discerned, the metal resources in the Altai and Minusinsk region would certainly have been of interest to those groups in the steppe already using these materials for household items and weapons (Kuzmina & Mair, 2008).

Already in the early prehistory of the Altai, there is evidence that at least two different cultural complexes met and interacted. Cemeteries of the hunter-forager groups in the Altai region (like the Lebed II site) provide insight into the groups indigenous to the region for that time (Bobrov, 1988). Presumably, these groups are the descendents of Neolithic southern Siberians. Their cemeteries were not built as kurgans and lacked stone slabs as capstones over inhumations. The material culture consisted of antler and lithic tools, bone carvings of elk and bear, and necklaces made of bear teeth (Anthony, 2007; Bobrov, 1988). Over time, stylistic elements of the Afanasievo material culture were emulated by these forager groups (Anthony, 2007). It is, however, not clear how much interaction occurred between the indigenous foragers and the pastoralist immigrants or whether these groups intermarried.

Another Eneolithic period following the Afanasievo Culture was characterized in the Minusinsk. This culture – the Okunev – shows no continuity with the preceding period. Kurgans were no longer constructed. Instead, burials were cysts made out of

stone slabs built in rectangular form. Low laying stone enclosures were built around cemeteries, and these cemeteries had higher densities of burials compared to the Afanasievo (Gryaznov, 1969). The skeletal remains were placed in the supine position with knees bent. Often, the burials were reused with the oldest remains being pushed to the side for the newest interment. Some tools were very similar to the Afanasievo in type, but others showed significant departures, specifically with the Okunev knives (Gryaznov, 1969).

Hunting and fishing were still integral means of obtaining food, along with the use of domesticated animals. Teeth from wild animals were found in abundance, especially sable, and were used as ornamentation in clothing (very similar to the Lebed II cemetery mentioned above) (Gryaznov, 1969). In addition, typical for the Okunev (and also unique for southern Siberians) were stone steles with female faces carved into them. Small (3-5 cm) stones with similarly carved faces were found in many of the Okunev burials (Gryaznov, 1969). These have been attributed to goddess/ancestress worship.

It is generally believed that the Okunev was the material culture of indigenous Siberian groups moving into the region, possibly from the Baikal area, or from northwestern Siberia. Features of the Okunev were shared with the Samus culture, which is found in western Siberia, centered on the Tom River, and is contemporaneous with the Okunev (Kovtun, 2008). In a re-analysis of Okunev sites, Sokolova (2007) categorized the Okunev into four chronological stages. The earliest of these stages corresponds to a pre-Afanasievo period, which has similarities with Siberian Neolithic cultures. The second dates to a period contemporaneous with the Afanasievo, where the classical Okunev burials mentioned above can be found along with “hybrid” burials containing

features of both cultural traditions. The third stage retained the cyst burial structure, but advances were made in the technology used to make ceramics. The final stage is characterized by ceramics similar to the Andronovo type (see discussion below). Thus, the Okunev tradition actually represents an indigenous culture that merged with the immigrant Afanasievo culture and subsequently developed along its own path (Sokolova, 2007).

Unlike the Minusinsk region, the Afanasievo transitioned directly to the Andronovo in the Altai (Jettmar, 1951). Although relatively short-lived (lasting between 18<sup>th</sup> and 12<sup>th</sup> centuries BCE), the Andronovo was widespread, stretching from the Minusinsk and Altai regions to Kazakhstan, and down to the Amu Darya in the south (Anthony, 2007; Kuzmina & Mair, 2008; Mallory, 1989; Mallory & Mair, 2000). Local varieties of this culture have been identified (Tkacheva & Tkachev, 2008). For example, the Alakul and Fedrovo were found in northern Kazakhstan and southern Siberia, respectively. Different styles of the Andronovo were also found in Kyrgyzstan and the Semirechye regions, and its influences were seen in the pastoralist Tazabagyab culture and urban centers of Transoxiana (Kuzmina & Mair, 2008). Many of the features were uniform (ceramics in particular), allowing these local varieties to be included in an overall Andronovo horizon.

The people belonging to the Andronovo culture lived sedentary lives (Anthony, 2007; Mallory, 1989). Villages of around a dozen houses and up to 40 houses have been excavated in Central Asia and southern Siberia. They continued to use domesticated animals, but now almost to the complete exclusion of wild game, with the vast majority of faunal remains coming from sheep and cattle (Anthony, 2007). They also employed

extensive use of metals, with bronze-tin replacing copper in most instances (Jettmar, 1951). Burials were in the form of kurgans again, but a wide variety of types was employed. A minority of the Andronovo burials (25-33%) contained two bodies – one male and one female. These finds were viewed as being a married couple buried together (Okladnikov, 1964, 1990). Because of this interpretation, there was speculation that the society transitioned from a matriarchal to a patriarchal society (Gryaznov, 1969).

It is assumed that the language spoken by Andronovo communities was a branch of Indo-Iranian (Anthony, 2007; Kuzmina & Mair, 2008; Mallory, 1989; Mallory & Mair, 2000). Words associated with patriarchal societies are common to most Indo-European languages (Anthony, 2007). The Afanasievo cultural community is also suspected of using a Proto-Indo-European language. Thus, it could be speculated that patriarchal societies have existed in southern Siberia as early as the Eneolithic.

The Andronovo also brought new styles of ceramics to southern Siberia that differed from those of the Okunev and Afanasievo cultures. These ceramics have greater affinities with archaeological material from the steppe (Anthony, 2007; Mallory, 1989). Findings from recent investigations make it clear that the Andronovo was yet another immigrant culture originating west of the Altai from the archaeological complexes of the Sintashta and Polktovka in northern Kazakhstan (Anthony, 2007).

The next cultural phase was called the Karasuk culture by Teploukhov, which immediately succeeded the Andronovo period (Jettmar, 1950, 1951). The Karasuk culture began around the 14<sup>th</sup> or 13<sup>th</sup> century BCE and is viewed as a continuation of the preceding period (Gryaznov, 1969; Jettmar, 1951; Legrand & Bokovenko, 2006; Okladnikov, 1964). Beginning in the Minusinsk basin and spreading through the Altai,

this culture was typified by advancements in metallurgical practices. The Karasuk flourished in the Minusinsk, where distinctive types of knives were made that were later spread throughout southern Siberia, Mongolia, China and eastern Central Asia (Chernykh, 2008; Okladnikov, 1964).

Local varieties of ceramic styles increased in number during this time. Most pottery had rounded bottoms in Minusinsk, but the earliest versions had flat bottoms and designs similar to the Andronovo. In the Altai, however, flat bottomed pots were the norm (Jettmar, 1950). Local varieties were also obvious in different burial customs. The Karasuk in the Altai region tended to possess characteristics distinct from either the Minusinsk or the various forms found among the lower stretches of the Ob' and Tom rivers. Burials in the Altai continued to be constructed in kurgan form, but there is no indication that a stone structure was used for construction (Gryaznov, 1969). Inhumations were small, oval pits where the flexed bodies were placed on their sides. Bronze and copper implements were found in most burials, but in meager quantities. Gryaznov believed these differences could be due to the presence of a different ethnic group living in the Altai (Gryaznov, 1969).

Few dwellings associated with the Karasuk have been found, but those that have suggest that structures served multi-purpose functions (Jettmar, 1951). The spatial organization of cemeteries from this period was described as representing clan or family units and was considered evidence for a patriarchal society (Jettmar, 1951; Legrand & Bokovenko, 2006; Okladnikov, 1964). By the end of the Karasuk, only men are found in burials (Jettmar, 1951). Given the increased number of Karasuk burials found throughout the Minusinsk and the increased numbers of sheep and horse remains, it is understood



that population sizes were expanding during this period (Legrand & Bokovenko, 2006). It is not clear, however, whether this was the result of local populations increasing in size, or if the population increase was because of immigrant (Andronovo peoples) settlers.

#### **1.4 From Nomads of the Iron Age to Nomads of Today**

The next phase in the prehistory of southern Siberia has been classified as the Iron Age or the Early Nomad Age (Gryaznov, 1969). As the name implies, pastoral nomadism became widespread in southern Siberia. The Minusinsk region provides much of the evidence from this period where it has been labeled the Tagar culture by Teploukhov and later subdivided into four stages (Bainovo, Podgornovo, Saragash and Tes') by Gryaznov (Bokovenko, 1995c; Legrand & Bokovenko, 2006). The Tagar dates from the 7<sup>th</sup> to 1<sup>st</sup> centuries BCE, and is seen as a continuation of the preceding Karasuk period (Bokovenko, 1995b). In the Altai region, this period is divided into three phases: Maiemir, Pazyryk and Shibe (Jettmar, 1951). These date to the 7<sup>th</sup> – 5<sup>th</sup> centuries BCE, the 5<sup>th</sup> – 3<sup>rd</sup> centuries BCE and the 2<sup>nd</sup> century BCE – 1<sup>st</sup> century CE, respectively (Bokovenko, 1995a; Jettmar, 1951). Although the presence of this culture in the Altai has been labeled as “Scytho-Siberian,” it does not necessarily provide a definitive characterization of the ethnic identity of these people.

One general feature of this period is the increasing use of iron, which in the Altai began during the Pazyryk and was completed by the Shibe phase (Jettmar, 1951). Burials are of the kurgan style, often built with stone and a central burial chamber of wooden logs. The kurgans vary in size, with the larger of these belonging to the elite of the community. In addition, the horse came to play a central role in the cultures of the Iron

Age, as evidenced by the burial of horses with the dead (multiple horses were found in some of the larger kurgans and often, in the more elaborate burials, the horses were of a different breed) (Anthony, 2007; Rudenko, 1970). This feature was widespread throughout the steppes of Central Asia and Siberia. The discovery of horse bits and chariots in some kurgans provide support for the idea that horseback riding and domestication of the horse had been accomplished by this point in time, although it has been argued that these elements could have arrived in Siberia as early as the Afanasievo (Anthony, 2007; Kuzmina & Mair, 2008).

In addition to horse remains and horse riding accoutrement, a drastic increase in the number of weapons was uncovered from these burials indicating greater levels of warfare (Gryaznov, 1969). In fact, evidence of violence has also been noted from the osteological remains (Jordana et al., 2009; Murphy, Gokhman, Chistov, & Barkova, 2002). Trade and exchange must also have been an important component of the economy of Iron Age Siberia, as Pazyryk burials showed items originating from India, Persia and China (Rudenko, 1970). The local artistic style (the Scytho-Siberian animal style) has also been found throughout Eurasia (Okladnikov, 1964), suggesting the influence southern Siberian culture had on cultures of the Eurasian steppe.

In the Altai, the Maiemir period showed regional differences between northern and southern Altai sites. In the north, the culture is called the Bolsherechensk culture (Jettmar, 1951). The reliance on wild animals and fish equaled that of domesticated animals, which is a departure from the earlier Karasuk. Tools for processing agricultural products were also found at these sites, indicating a rather generalized subsistence strategy. It appears these people lived a more sedentary lifestyle (Jettmar, 1951). The

archaeological sites of the Maiemir from the High Altai were different in that dwelling sites were lacking. All of the burials included weapons and horses, leading researchers to believe this was a more nomadic, warlike group (Jettmar, 1951).

The following phase was differentiated by the use of iron and the shape of bronze mirrors and horse bits, but differences continued to persist between northern and southern sites (Jettmar, 1951). In the northern Altai, two types of kurgans were found, namely – the Berezovsk and Tuiakhta groups. The mountainous regions of southern Altai provided at least three different types of kurgans. The first of these was classified as “simple burials.” They contained elements that were reminiscent of the Tuiakhta group of the northern Altai and of Tagar burials of Minusinsk (Jettmar, 1951). The second type was the “middle kurgans” that had some similarities with the third type – the so-called “princely kurgans” (Jettmar, 1951; Rudenko, 1970).

One set of the princely kurgans are the famous frozen burials found in the Altai where extraordinary finds were preserved in ice (Rudenko, 1970) . These tombs were looted in antiquity, but because they were opened, water was allowed to seep into them then later froze, preserving much of the material that would otherwise have been lost to decay. Human remains (including some with tattoos), textiles, and wooden artifacts were found very well preserved. These royal burials provided insights into the Iron Age period of the Altai. Surprisingly, some of the artifacts in the Pazyryk burials match items also found in eastern Europe and were mentioned by Herodotus in his description of the Scythians (Herodotus, Waterfield, & Dewald, 1998; Rudenko, 1970).

Above all else, the major defining feature of this period is the adoption of a nomadic lifestyle – the uniformity of which stretched over the expanse of the Eurasian

steppe. While local varieties certainly existed, obvious similarities in culture, food production and warfare among the Eurasian steppe nomads were enough to prescribe appellations of “Scythian” or “Saka” to most of them by ancient historians. These cultural elements (particularly pastoral nomadism) had lasting impacts on Eurasian populations where they were shared across two continents and persisted for over two thousand years, this despite the presumably heterogeneous nature of Eurasian populations (Comas et al., 1998; Comas et al., 2004; Kozintsev, 2007; Wells et al., 2001; Zerjal, Wells, Yuldasheva, Ruzibakiev, & Tyler-Smith, 2002).

It is not at all clear that the Eurasian steppe populations shared a common ancestry. Regional variations in material culture and changes in these cultures over time indicate separate histories and trajectories for them. In southern Siberia, the Scytho-Siberian animal style is shared with other regional cultures, but the burial customs are rather different from their neighbors. The structure of the Central Asian Saka cemeteries were also distinctive, consisting of common and special kurgans (Yablonsky, 1995). The special kurgans consisted of a large kurgan and two smaller ones with a line of stones connecting them. This characteristic structure has been called a “kurgan with mustache,” and is typical in Saka cemeteries throughout the Kazakh steppes (Yablonsky, 1995, 201). Despite these differences, similarities in grave goods and kurgan construction were found among burials in central Kazakhstan, the lower Syr Darya region and the Altai mountains, which, in turn, were different from burials in Semirechye and Tien Shan regions (Yablonsky, 1995). Furthermore, the Altai region is the only one in which there is evidence of mummification and embalming. Mummification occurred naturally in some cases but, in others, evidence of efforts to mimic the body of the dead is clear. This

usually occurred through the use of masks, straw and even at times the construction of nearly complete dolls (Mallory & Mair, 2000; Rudenko, 1970).

Following the “Scythian” phases of the Maiemir, Pazyryk and Shibe cultures in the Altai, the archaeological record becomes scanty. It is possible that the few burials found in the northern Altai represent a movement of people coming down from the forest steppe and taiga of northwestern Siberia. The finds from this time were similar to those of the presumably Finno-Ugric peoples living just to the north and west of the Altai region (Jettmar, 1951). There is, in fact, a dearth of archaeological information from the Altai between the 2<sup>nd</sup> and 5<sup>th</sup> centuries CE.

Nearby in Minusinsk, the Tashyk culture prevailed. Chinese influence is obvious in the material culture from this period (Okladnikov, 1964), and it has long been associated with Xiongnu hegemony (Martynova, 1988). It was during this time that the nomadic Xiongnu conquered surrounding nomadic peoples, including the Yuezhi (presumably Indo-European speakers), forcing them west away from the Tarim Basin (Mallory & Mair, 2000). Chinese historical records note the advance of the Xiongnu into their lands at this time. The subsequent struggle between Xiongnu and Chinese empires for control of the Tarim Basin ensured ample records for later historical analysis (Barfield, 1989; Di Cosmo, 1994; Grousset, 1970; Mallory & Mair, 2000).

In the Altai, there is little evidence to suggest that the Xiongnu had much control or even interest in the region. It has been suggested that with the defeat of the Yuezhi by the Xiongnu, the Yuezhi moved towards the Altai, resulting in the Pazyryk culture (Okladnikov, 1964), although there is also evidence to suggest that the Pazyryk developed in the Altai before the 2<sup>nd</sup> century BC (i.e., before the Yuezhi were defeated

and forced to move west) (Jettmar, 1951). In addition, it is fairly well attested that the Yuehzi helped to form the Kushan empire in Afghanistan and northwestern India (Golden, 1992; Grousset, 1970). There is no question, however, that Pazyryk represented a heterogeneous population, as the variety of burial structures and cranial morphological analyses suggest.

By no means is the ethnographic information for populations living in southern Siberia during the Bronze and Iron Ages clearly resolved. Historical accounts from ancient China name the people of this general region (southern Siberia) as the Ting-ling and describe them as having blond or reddish hair and fair skin (Menges, 1968; Potapov, 1962). Apparently, the Xiongnu conquered the Ting-ling and the Kyrgyz (Ko-k'un and Chien-k'un) in southern Siberia around 200 BCE and again in 49 BCE, possibly resulting in intermixture between these tribal groups (Golden, 1992).

The region over which the Xiongnu reigned was continuously and successively dominated by various "nomadic hordes" (Barfield, 1989; Grousset, 1970). From the 3<sup>rd</sup> to 6<sup>th</sup> centuries CE, the Xiongnu, Xianbei, Tabgatch/Toba, Jou-Jan, and Ephthalite Huns each controlled portions of Central Asia, Mongolia and northern China. All of these groups apparently spoke either Mongolic or Turkic languages or precursors of these languages (Golden, 1992; Grousset, 1970). During the 6<sup>th</sup> century CE, the Jou-Jan had power over the regions surrounding the Gobi desert in modern day Mongolia, with their power extending as far west as Turfan. The Ephthalite Huns controlled the regions to the west and south from Lake Balkhash in modern day Kazakhstan. Meanwhile, according to ancient Chinese historians, the Altai was inhabited by the T'u-chüeh, who are also referred to as Turks (Golden, 1992; Grousset, 1970).

By the mid-6<sup>th</sup> century CE, revolts by tribes from within the Jou-Jan Empire succeeded in removing its hegemony, and a new power emerged in the form of the first Turkic Khanate. The Turks were described as the blacksmiths of the Jou-Jan (Golden, 1992), giving circumstantial evidence that Turks were from Minusinsk and the Upper Yenisei region, which had continuously supplied the metallurgical needs of the cultures throughout the Bronze and Iron Ages (Okladnikov, 1964, 1990). Currently, the debate concerning the origins of the Turks is still unresolved (Golden, 1992; Sinor, 1990). Yet, the success of the Turkic Khanate is evident considering its rapid political expansion and the vast number of peoples reportedly subservient to it. Among those, several are known from the Yenisei and Altai regions, such as the Az, Čik, Kyrgyz and İzgil; but it is not known what languages they spoke, nor how they were related to one another (Golden, 1992).

After the death of the Turk's first khan, Bumen, the territory was split into eastern and western khanates with their borders delineated by the Altai (Grousset, 1970). Internal strife disallowed a union of the two groups. In the 7<sup>th</sup> century CE, both T'u-chüeh (Turk) states collapsed, in part, by Chinese military action. Near the end of the 7<sup>th</sup> century, the T'u-chüeh regained power, although their rule was short lived.

The Uyghur state emerged in the 8<sup>th</sup> century and lasted through the 9<sup>th</sup> century CE. In 758, the Uyghurs conquered the Yenisei region and, presumably, the Altai (Menges, 1968). During this time, it is believed the Yenisei Kyrgyz formed around the Yenisei River, while a separate cultural group formed on the other side of the Ob' River (Okladnikov, 1964). The non-Kyrgyz group had a hunter-forager subsistence economy, and was found throughout the forest and taiga, while the Kyrgyz were confined to the

steppe lands. Cranial morphology of these groups showed differentiation between the hunter-foragers, who had a greater affinity with populations to the west, versus those of the steppe, who had greater influences from the populations in eastern Siberia and Mongolia (Okladnikov, 1964).

Based on records kept by Chinese historians, it appears the Yenisei Kyrgyz were already speaking a Turkic language in the 8<sup>th</sup> century CE. However, it is not clear whether they were entirely assimilated or if they spoke another language in the past, such as a Samoyedic or Yeniseian tongue (Menges, 1968). The Kyrgyz formed their own khanate and, in the 10<sup>th</sup> century, succeeded in overthrowing the Uyghur Empire. They were later defeated by the Khitans and presumably driven back to the Yenisei region (Grousset, 1970). By this time, the area between the Yenisei and the Ob' was probably inhabited by Samoyedic-speaking populations (Okladnikov, 1964). Thus, the Turkic-speaking Kyrgyz likely ruled over the Samoyedic and Yeniseian populations of the Altai-Sayan region (Forsyth, 1992).

In the 12<sup>th</sup> century, the Jurchids, who lived to the north of the Khitans in the far eastern areas of Siberia, overthrew the Khitan empire (Grousset, 1970). Yet, in the west, the Kara-Khitan Empire continued blossoming. The Altai region was under the control of either the Kara-Khitans or Naiman confederation in the 13<sup>th</sup> century, when Genghis Khan's Mongol empire rose to power (Golden, 1992; Grousset, 1970; Potapov, 1964a). The Altai was brought under Mongol control in 1207 during Jöchi's conquests north of Mongolia, at which time numerous groups were listed as living there. These include the Čhinos, Tö'ölös, and Telenggüd of the Adarkin, as well as the Oyirad, Buriyad, Bargun, Ursud, Qabqanas, Qangqas, Tubas, and Kirgisuds (Cleaves, 1982; Golden, 1992). The



descendents of Genghis Khan continued to rule Central Asia for the next several centuries.

From the 15<sup>th</sup> to 17<sup>th</sup> centuries, southern Siberia was under the influence of western Mongols (Dzungarians and Oirats). It is at this time that the ethnogenesis of current populations in the Altai (and more generally, Central Asia) is believed to have begun (Golden, 1992; Potapov, 1964a). The Kipchak tribes are thought to have begun differentiating into various ethnic groups after the demise of the Golden Horde, including southern Altaian populations (Potapov, 1964a).

The Russians soon advanced into the region. The Russian colonization of Siberia had greater impact on the Altai in the 17<sup>th</sup> and 18<sup>th</sup> centuries (Collins, 1991; Forsyth, 1991, 1992; Naumov & Collins, 2006; Potapov, 1964b). It is at this time that some of the current ethnic groups, as they are known today (and/or their historical antecedents), were first recorded. The Altaians were differentiated based on where and how they lived. Those hunter-foragers who lived in the taiga were placed in the northern Altaian group. The other group consisted of pastoral nomads living on the steppe very much like other nomadic groups in Kazakhstan and Mongolia. The former consisted of Chelkan, Tubalar, Kumandin and Shor, all of whom were already speaking Turkic languages. They were believed to be related to their Yeniseian or Samoyedic speaking neighbors across the Yenisei River – Asans, Baikots, Kamasins, Karagas, Kotts and Motors (Forsyth, 1992). The latter group likely consisted of the ancestors of modern day southern Altaians and included Arin, Chats, Eushta, Kachin, Koibal, Kyzyl and Teleut communities (Forsyth, 1992). Like the northern group, some of these populations may be descendents of

Samoyedic or Yeniseian groups, but all spoke Turkic when first encountered by the Russians.

In 1703, a local elite that was made up of Kyrgyz (consisting of Altyr, Altysary, Tuba and Yezer clans) lost all political power and were forced into Dzungaria, after which the Russian government had free reign to colonize the Altai, finally annexing the region in 1756 (Forsyth, 1992). With the political vacuum caused by the absence of the Kyrgyz and the settlement of the Russian peasants around the Yenisei, pastoral nomadic indigenous populations moved south and west to the Abakan, Kuznetsk and Altai ranges. Those groups that did not move quickly were assimilated into Russian culture. In the northern Altai region, the Shor and Teleut were particularly affected by Russian settlements (Forsyth, 1992). Southern Altaians (Altai-kizhi, Telenghit and Tölös) were affected less by the Russians as they lived in the more inaccessible mountainous region of the Altai. They also had greater contact with populations in Mongolia. Nonetheless, the anti-colonist sentiment was strong throughout the Altai and Yenisei areas, and frequent revolts, although often small, occurred sporadically throughout the late 1800s and early 1900s (Forsyth, 1992).

Possibly the most dramatic change imposed upon the indigenous people of the Altai and Yenisei regions was the forced collectivization ordered by the Russian Communist Party. Collectivization dramatically affected local economies, and also had significant effects on indigenous culture by abolishing clan structures and nomadism (Bobrick, 1992; Conquest, 1986; Forsyth, 1992; Naumov & Collins, 2006; Potapov, 1964b). The Yenisei and Altai were extremely hard hit by these policies, resulting in some Altaian and Kazakh groups moving to Xinjiang. By 1933, around 87% of Altaian

farms were collectivized. After 1945, some collectives were combined to form even larger state farms. It was during this period that the clans and various groups living in the Altai regions began acknowledging themselves as “Altaian” (Forsyth, 1992).

### **1.5 Craniometry and Genetics: Prospects of Population Affinities**

A central component of the archaeology and history of southern Siberia cultures is the assessment of genetic relatedness among groups. The examination of morphological variability in crania to assess population affinities has a long history in Siberian archaeology. Several studies focused specifically on the Bronze and early Iron Age periods of southern Siberia, but anecdotal comments on cranial shape are typical in discussions of prehistoric Siberia (Okladnikov, 1964). The general trend has been to assess the levels of “Caucasoid” and “Mongoloid” characteristics in crania from a particular location and discuss how they relate to remains from variously dated burials (Debets, 1962; Gryaznov, 1969; Jettmar, 1951; Levin, 1964; Levin & Potapov, 1964; Okladnikov, 1964). In particular, attempts were made to characterize crania as belonging to one of several Caucasoid races (Khodzhayov, 2008; Kozintsev, 2009 and references therein). However, this practice is problematic from historical and scientific perspectives (Brace, 2005; Mallory & Mair, 2000).

Typically, the Afanasievo and Andronovo cultures were classified as Caucasoid, with ongoing debates as to whether these were “Mediterranean” types or “Proto-Europoids” (Kozintsev, 2009). Greater “Mongoloid” components were often found in Siberian forest belt populations throughout the Neolithic and Eneolithic (Levin 1964). Mongoloid components also increase in the Altai during the Okunev and Pazyryk

periods, although remains from the Pazyryk kurgans were quite heterogeneous (Okladnikov, 1964; Potapov, 1962, 1964a). This trend continued through the reign of the Turkic Khanate, Uyghurs and Kyrgyz as late as the 10<sup>th</sup> century CE, but affected the southern more so than the northern Altai regions (Potapov, 1964a). In fact, the recent use of Europoid races (particularly the Mediterranean type) as designations for southern Siberian morphology has led some to hypothesize that Bronze Age cultures from this region had an origin in the Near East or Trans-Caucasia (Kozintsev, 2000 and references therein). Nevertheless, the problems associated with the subjective classification of ancient Siberian or Central Asian crania into typologies that were created for living Europeans have resulted in confusion; researchers cannot even agree about the category to which some of these remains belong, as Kozintsev and Mallory and Mair have noted (Kozintsev, 2009; Mallory & Mair, 2000).

Several studies have instead used cranial measurements to assess the population affinities among Siberian, Central Asian and European populations. Hemphill and Mallory (2004) examined a number of Siberian crania while examining the origins of Bronze Age Tarim Basin populations. Mahalanobis  $D^2$  distances were estimated and used in conjunction with weighted pairwise average linkage and neighboring joining methods (Hemphill & Mallory, 2004). Additionally, multi-dimensional scaling, minimum spanning trees and principal coordinate analysis were used to assess population affinities (Hemphill & Mallory, 2004). The Siberian populations included two Afanasievo, two Andronovo and one Karasuk sample sets, totaling 215 males and 178 females. After data from both sexes were standardized, craniometric data showed similarities between Afanasievo samples from the Altai and Minusinsk, which were

similar to Andronovo samples from Kazakhstan and Minusinsk. This observation verifies the descriptive analysis which noted similarities in cranial morphology of these groups (Okladnikov, 1964). Here, the Karasuk were the most divergent of the Siberian samples (Hemphill & Mallory, 2004).

In a series of separate analyses, Kozintsev (2008, 2009) examined populations spanning Eurasia, including numerous ones from Siberia. The sample set includes 220 male crania dating from Neolithic through Iron Ages. Data was obtained from nine Afanasievo samples (with six samples coming from the Altai and three from Minusinsk), five Okunev samples (four from Minusinsk and one from Tuva), and seven Andronovo samples (two from western Kazakhstan representative of the Alakul variety and five representing the Fedrovo type – one from Minusinsk, one from Rudny Altai, two from the Upper Ob, and one from northeastern Kazakhstan). In addition, a couple of Siberian forest belt cultures were represented, including Samus and Yelunino cultures (Kozintsev, 2008, 2009). All groups were compared in a pairwise manner with Mahalanobis  $D^2$  values (Kozintsev, 2008, 2009).

When compared to remains from eastern European archaeological sites, the Afanasievo crania had the greatest affinities with the Katakombnaya (Catacomb) culture (Kozintsev, 2009). Some also have closer affinities with the Yamnaya (Pit Grave) culture, but the Afanasievo from the Altai generally did not. Furthermore, there was also an association with the Scrubnaya (Timber Grave) culture, which postdates the Afanasievo. Thus, although the evidence suggests an eastern European origin for the Afanasievo, a single conclusive match for the people of this culture was not found.

Following the Afanasievo in Minusinsk, the Okunev was considered an immigrant population with greater Mongoloid characteristics (Okladnikov, 1990). The Okunev sample from Tuva proved to be an interesting case. As part of a larger study on Scythian origins, it was consistently noted that Scythian groups had the closest affinities to the Okunev of Tuva (Kozintsev, 2007). As a result, it was stated that the Okunev probably migrated from Europe (Kozintsev, 2008). This was an unexpected result, given the greater Mongoloid characteristics reported for Okunev crania.

The most recent analysis instead classified the Okunev from Minusinsk as an indigenous Siberian population that had similarities with Neolithic crania from Krasnoyarsk (Kozintsev, 2009). Meanwhile, Kozintsev maintained the relationship between Okunev from Tuva with eastern European groups. Chikisheva (2008) examined several cranial series from Tuva, but came to a different conclusion regarding the Okunev of Tuva. She used cluster analysis to show that the Okunev from Tuva have affinities with the Karasuk from Khakassia and Andronovo from Minusinsk and western Kazakhstan (Chikisheva, 2008).

The two Andronovo varieties (Alakul and Fedrovo) showed similarities with eastern European groups and Yamnaya (Pit-Grave) series in particular (Kozintsev, 2009). The eastern Alakul Andronovo group showed this pattern, but the western Alakul had greater affinities with late Catacomb series from the Ukraine. The Fedrovo Andronovo crania from the Upper Ob' were most similar to the Afanasievo of the Altai, while the Fedrovo from Rudny Altai had greatest affinities with the Samus culture, consistent with the notion that there is continuity between the Afanasievo and Andronovo cultures in

Siberia. The remaining Fedrovo from southern Siberia and northeastern Kazakhstan appeared to have a closer affinity with the Yamnaya (Pit Grave) and Catacomb cultures.

The general trend for cranial morphology reveals the similarities among Afanasievo and Andronovo populations and similarities between these and eastern European populations. Given the archaeological evidence, I would argue that the Okunev is representative of the indigenous Siberian groups, most likely resembling the Neolithic populations of Siberia west of the Yenisei. The Karasuk was shown to be more similar to the Afanasievo and Andronovo, suggesting that some continuity in populations existed throughout the Bronze Age of southern Siberia. These findings also are supported by non-metric cranial features (Moiseyev, 2006). However, there is a problem in comparing all of these studies because each uses different data sets, different batteries of measurements and different methods for determining relationships between populations.

Regardless of this issue, one point taken from the Siberian craniometric data is especially clear, although not explicitly noted. Samples from multiple sites composing the same cultural group do not cluster together to the exclusion of others. Therefore, the heterogeneity of cranial features present in each cultural group is significant. This is true for the Afanasievo, Andronovo, Okunev and Scythian sample sets. This heterogeneity must therefore be considered when discussing the implications of any migrations. In particular, one needs to know whether the migrations are consistent with a single homogenous entity or a by-product of interactions between recent immigrant and existing populations. Historical records show that migrations across the steppe are complicated and intricate, sometimes involving genetic exchange, and at other times just cultural innovations (Golden, 1992). Furthermore, nomadic groups can control large expanses

without necessarily contributing DNA to the existing sedentary populations (Grousset, 1970).

The boundary between populations resembling eastern European groups versus indigenous Siberian groups appears to be where the steppe meets the forest belts/taiga in southern Siberia. Because a pastoral economy is closely tied with the environmental conditions of the steppe, and because the hunter-forager economy is equally dependent on the taiga, it will be crucial to determine whether a genetic boundary such as this persists throughout the history of the region. Such boundaries can also impede the spread of languages and maintain cultural barriers. Anthony calls these barriers “persistent cultural frontiers” (Anthony, 2007).

Anthropometrical studies of modern Siberians are generally lacking. Instead, the literature often cites the racial typologies popularized in the early to mid 1900s. Southern Siberian populations are generally classified into three types: the Central or Middle Asian, Ural and Baikal types (Levin, 1964). While these descriptive analyses were illustrative at the time they were initially made, they lack the scientific validity necessary for understanding the biological relationships among phenotypes.

Modern Siberian populations have been studied far more extensively using molecular genetics methodologies. The earliest molecular anthropological studies conducted on Siberian populations largely involved protein polymorphisms including blood group markers, serum proteins and immunoglobulin allotypes. Ethnic groups from northeastern Siberia were often the subjects of inquiry, in particular the Chukchi, Koryak, and Siberian Eskimo (Rychkov Iu & Udina, 1985; Sheremet'eva & Gorshkov, 1977, 1981; Solovenchuk & Avanesova, 1977; Solovenchuk, Deviatkina, & Avanesova, 1976;



Sukernik, Lemza, Karaphet, & Osipova, 1981; Sukernik & Osipova, 1982). At the same time, studies published by Sukernik and colleagues added information on northern Altaians, Nganasans and Forest Nentsi (Karafet & Sukernik, 1978; Osipova & Sukernik, 1978; Sukernik, Abanina, Karafet, Osipova, & Galaktionov, 1979; Sukernik, Gol'tsova, Karafet, Osipova, & Galaktionov, 1977; Sukernik, Karafet, Abanina, Korostyshevskii, & Bashlai, 1977; Sukernik, Karafet, & Osipova, 1977; Sukernik, Karafet, Osipova, & Posukh, 1985; Sukernik & Osipova, 1976).

Several generalizations can be made from the results of these and other earlier studies (Crawford, Williams, & Duggirala, 1997). The results provide a clear connection between modern Siberian and Native American populations. Although no one particular group can be assigned as the ancestral population for Native Americans, there is a closer affinity between populations from southern Siberia and Mongolia and indigenous groups in the Americas. In addition, genetic diversity in Siberia is structured by geography and, in some cases, language. This structure is due largely to stochastic and migratory processes, which influenced indigenous Siberian populations, most of which had small effective population sizes and are often geographically isolated (Crawford et al., 1997). Because of the characteristics of these populations, it is argued that evidence for adaptation could be difficult to detect, as random genetic drift will play a significant role in determining current genetic diversities.

The objectives of the first studies examining mtDNA variation in Siberian populations were also to understand the origins of Native American populations. The first significant use of mtDNA in molecular anthropology for answering these questions occurred in the mid 1980s and early 1990s (Schurr et al., 1990; Torroni et al., 1992;

Wallace, Garrison, & Knowler, 1985; Ward, Frazier, Dew-Jager, & Paabo, 1991). This work appeared at about the same time that the landmark study of the worldwide distribution of human mtDNA variation was published (Cann, Stoneking, & Wilson, 1987; Vigilant, Stoneking, Harpending, Hawkes, & Wilson, 1991). These studies provided the groundwork for investigating Siberian mtDNA variation. They also popularized two different methodologies for characterizing mtDNAs in molecular anthropology – restriction fragment length polymorphisms (RFLPs) and control region (CR) sequencing (Torrioni, Schurr et al., 1993). The central focus of many studies on Siberian populations became the correlation of mtDNA results from Siberian populations with those from American populations. A search for the origins of the four mtDNA haplogroups initially identified in Native American populations helped to foster research on Siberian populations, which also led to the recognition of additional mtDNA haplogroups not present in the Americas.

The common ancestry of mtDNAs from Siberian and Native American populations were confirmed in this earlier phase of mitochondrial-based molecular anthropology (Torrioni, Sukernik et al., 1993). Haplogroups A, B, C and D were indeed found in Siberian populations. The focus, however, remained on populations in the northeastern section of Siberia, closest to the Bering Strait (Derbeneva, Sukernik et al., 2002; Forster, Harding, Torrioni, & Bandelt, 1996; Malyarchuk, Derenko, Balmysheva, Lapinskii, & Solovenchuk, 1994; Malyarchuk & Derenko, 1995; Malyarchuk, Derenko, & Solovenchuk, 1994; Malyarchuk, Lapinskii, Balmysheva, Butorina, & Solovenchuk, 1994; Rubicz, Schurr, Babb, & Crawford, 2003; Rychkov, Naumova, Falunin, Zhukova, & Rychkov Iu, 1995; Schurr, Sukernik, Starikovskaya, & Wallace, 1999; Shields et al.,

1993; Starikovskaya, Sukernik, Schurr, Kogelnik, & Wallace, 1998; Sukernik, Schurr, Starikovskaia, & Uolles, 1996; Torroni, Neel, Barrantes, Schurr, & Wallace, 1994). As early as the 1700s, theories on the origins of America's indigenous populations focused on the physical and cultural similarities between the populations of northeastern Siberia and the Americas (Thomas, 2000). Given the prevailing theory that Native Americans entered the New World at a time when the two continents were connected via the Bering land bridge, a focus on the populations of northeastern Siberia were critical in these early molecular anthropological studies.

One study departed from the others in examining Kets and northern Altaians in northwestern and southern Siberia, respectively (Sukernik et al., 1996). They found that northern Altaians were the only population to have all four of the haplogroups (A, B, C, and D) that were found in the Americas. A fifth haplogroup (X) was later also confirmed in Altaians (Derenko, Grzybowski et al., 2001). The focus eventually shifted to southern Siberia and Mongolia, where two studies found that Mongolians also had all four Native American haplogroups, making them potential ancestral populations for New World populations (Kolman, Sambuughin, & Bermingham, 1996; Merriwether, Hall, Vahlne, & Ferrell, 1996). However, as studies began examining populations along the border of Siberia and Mongolia/northern China, it became clear that many of these peoples shared similar haplogroups, although some possessed the "Native American" haplogroups at higher frequencies (Derenko et al., 2000).

Multiple studies were conducted on populations of southern Siberia (Derenko, Denisova et al., 2001; Derenko, Maliarchuk, Denisova, Dambueva et al., 2002; Derenko, Maliarchuk, & Zakharov, 2002; Derenko et al., 2000). RFLP analysis and control region

sequencing of these populations culminated in a large study of nearly 500 samples from seven ethnic groups inhabiting the regions across southern Siberia (Derenko et al., 2003). Several themes could be identified from these publications. The first is that the origin of Native American mtDNAs was still the primary focus of the Siberian molecular anthropology studies. Second, genetic boundaries could be found within regional sets of populations, in this case, between Tuvinians and Buryats in southern Siberia. At the same time, differences between western and eastern Siberian populations were noted in a larger comparison of Siberian populations (Schurr & Wallace, 2003). These were the first instances where differentiation was examined between large populations within the same general region of Siberia.

For the southern Siberian populations, the clusters were explained by invoking relative amounts of “Caucasoid” versus “Mongoloid” specific mtDNA haplogroups (Derenko, Denisova et al., 2001; Derenko, Maliarchuk, & Zakharov, 2002). Even regional differences within an ethnic group were discussed in these terms (Golubenko, Puzyrev, Saliukov, Kucher, & Sanchat, 2001). Ultimately, it was concluded that, despite the high amounts of migration into and out of southern Siberia, population differentiation was still significant among ethnic groups of this region (Derenko et al., 2003).

Today, we have a greater understanding of the mitochondrial genetic diversity of Siberia. While studies were first explicitly undertaken to understand Native American origins, the trend slowly switched to examining relationships among Siberian populations buried in historical, cultural, linguistic and anthropological contexts (e.g. Schurr & Wallace, 2003). Although this general trend persists, the importance of understanding Native American origins from a Siberian perspective has not lost its significance and

continues to be a fruitful line of research (Achilli et al., 2008; Derenko, Malyarchuk, Grzybowski et al., 2007; Fagundes et al., 2008; Starikovskaya et al., 2005; Tamm et al., 2007; Volodko et al., 2008).

The Y-chromosome played a far narrower role in molecular anthropological studies of Siberian populations in the late 1980s and early 1990s. Early studies focused on a few fragment length polymorphisms and one SNP (Lin et al., 1994; Nazarenko & Puzyrev, 1985; Pena et al., 1995; Santos, Pena, & Tyler-Smith, 1995; Zerjal et al., 1997). While important because they utilized the Y-chromosome, these studies demonstrated only very broad relationships, as this genetic system could be assayed at only a very low resolution, unlike the mtDNA. After seminal works identifying Y-chromosome polymorphisms (mostly SNPs) and their geographic distributions, the Y-chromosome was primed as an ideal complement to the mtDNA (Hammer, 1995; Hammer et al., 2001; Hammer et al., 1997; Hammer & Zegura, 1996; Underhill et al., 1997; Underhill, Jin, Zemans, Oefner, & Cavalli-Sforza, 1996; Underhill et al., 2001; Underhill et al., 2000).

Studies using the Y-chromosome mirrored those using the mtDNA in terms of the populations examined and the questions asked. The first three of these studies were concerned with the origins of Native American Y-chromosomes and in which populations in Siberia that these Y-chromosomes could be found (Karafet et al., 1997; Lell et al., 1997; Santos et al., 1999). These studies quickly expanded from using 3 to 5 loci to around thirty, thereby increasing the resolution of Y-chromosome haplotype trees and networks. The general consensus from these three studies was that Native American populations derived from a Siberian source, most likely southern Siberia, possibly along the Yenisei River. Other populations in Siberia and Mongolia were investigated for

“Native American” Y-chromosomes (Bortolini et al., 2003; Lell et al., 2002; Zegura, Karafet, Zhivotovsky, & Hammer, 2004). The number of polymorphic loci as well as the type of markers used (SNPs and STRs) helped to further clarify the relationships between Siberian and American populations. Later studies involving Siberian populations and their relationships eventually focused on southern Siberian ethnic groups (Derenko, Maliarchuk, Denisova, M. et al., 2002; Derenko et al., 2006; Kharkov et al., 2009; Stepanov & Puzyrev, 2000a, 2000b, 2000c). Differentiation among southern Siberian ethnic groups was quite evident, as had been seen with the mtDNA.

One paper looked at Siberia as a whole (Karafet et al., 2002). This study was set apart from the others in its objectives and geographical scope. In addition to creating a context for understanding the peopling of the New World and Japan, it aimed to comprehend the genetic Y-chromosome signatures of populations living a hunter-gather subsistence lifestyle. By including 28 populations from all over Siberia and parts of Central Asia, a more precise picture of paternal genetic histories was attained. Furthermore, researchers uncovered two new Y-chromosome polymorphisms that now serve as critical markers for Siberian and Native American Y-chromosomes.

It is from these contexts that the questions being examined in this dissertation were considered. By characterizing mtDNA and NRY lineages at the highest resolution possible, important historical and anthropological questions can be more fully addressed, in particular those that previous studies have inadequately investigated or even ignored. Clearly, the search for origins of Native Americans played a fundamental part in studying Siberian populations. As such, this data can also be used to address this question in greater detail than previous studies.

Just as important, these analyses will address genetic relationships at a finer anthropological scale. In particular, it can be determined how closely northern Altaian populations are related to southern Altaian populations and how northern Altaian ethnic groups have interacted with each other to persevere against new cultural elements brought to the region by Russians – including a Russian sedentary way of life, education and economy. Furthermore, by placing the genetic results into the archaeological and historical framework, it may be possible to gain a more complete understanding of the influences of past cultures on the current populations of the Altai.

## **Chapter 2: Background to Methods**

The methodological approach used in this dissertation employs two genetic systems used extensively for exploring human origins, migrations and population histories. The mitochondrial DNA and Y-chromosome share similar characteristics that have made them powerful tools for investigating these types of questions. Based on the vast quantity of papers and books published in the past 15 years, it is safe to say that the majority of geneticists believe that these systems are neutrally evolving molecules. Yet, these interpretations have been criticized. Throughout this chapter, I will assess the usefulness of mtDNA and Y-chromosome polymorphisms as neutral genetic markers. I will also discuss critiques made against the use of these genetic systems for population and phylogenetic studies, including issues of recombination, uniparental inheritance and natural selection.

### **2.1 The Mitochondrial Genome**

Mitochondria are organelles located in the cytoplasm of nearly every cell in eukaryotes and are the primary source of energy production (adenosine triphosphate – ATP). As organelles, they are unique in that they possess their own genomes. Mitochondrial DNA is substantially different from nuclear DNA in that it is circular in form and replicates independently from nuclear DNA without regard to the cell cycle. The genome consists of 37 tightly packed genes, including 13 protein-coding genes, 22 transfer RNA (tRNA) genes and 2 ribosomal RNA (rRNA) genes. The 13 polypeptides encoded in the mtDNA, along with nuclear encoded genes, are all involved in the process of generating ATP through oxidative phosphorylation (OXPHOS). The mitochondrial



genome is also unique in that it utilizes a slightly different genetic code than nuclear DNA (e.g., UGA codes for tryptophan instead of a termination codon, AUA can be an initiation codon).

In this regard, it has been argued that the origin of the mitochondrion organelle was the result of an endosymbiotic event between early prokaryotic and eukaryotic organisms (Chang, Wang, Hao, Li, & Li, 2010; Gray, Burger, & Lang, 1999; Margulis, 1970; Sagan, 1967; Wallace, 2007). After endosymbiosis, many of the genes involved in oxidative phosphorylation were transferred to the nucleus. The separate and independently replicating genome and different genetic code may be features retained from this symbiotic event (Barrell, Bankier, & Drouin, 1979; Wallace, 1999).

The characteristics of the mtDNA genome made it desirable for use in initial molecular evolutionary studies. Its high copy number made it easily accessible for research purposes before the development of the polymerase chain reaction (PCR). It also accumulates new mutations relatively quickly and is a haploid genetic system (Brown, 1980; Brown, George, & Wilson, 1979). The lack of recombination and its mode of transmission (from mother to offspring) make the construction of maternal haplotypes and their phylogenies rather clear cut (Awise et al., 1987; Giles, Blanc, Cann, & Wallace, 1980). Furthermore, as a neutral marker, mtDNA sequence evolution occurs in a predictable manner; it has a molecular clock.

Anthropologists and biologists have extensively used the mitochondrial genome in the study of evolution. Portions of the genome are conservative in nature and are useful for defining differences between species, while hypervariable segments of the genome (with a propensity for accumulating new mutations) have made it the molecule

of choice for differentiating haplotypes within species (Irwin, Kocher, & Wilson, 1991; Kocher et al., 1989). Much of this work focused on constructing molecular phylogenies and verifying taxonomies created from morphological characteristics (Cropp & Boinski, 2000; Cropp, Larson, & Cheverud, 1999; Ferris, Brown, Davidson, & Wilson, 1981; Ferris, Wilson, & Brown, 1981; Gagneux et al., 1999; Hixson & Brown, 1986; Horovitz & Meyer, 1995; Nei & Tajima, 1985; Pastorini, Martin, Ehresmann, Zimmermann, & Forstner, 2001; Ruvolo, Disotell, Allard, Brown, & Honeycutt, 1991; Ruvolo et al., 1994; Yoder, Cartmill, Ruvolo, Smith, & Vilgalys, 1996). MtDNA has been used to address similar questions concerning humans, starting in the mid-1980s (Cann et al., 1987; Johnson, Wallace, Ferris, Rattazzi, & Cavalli-Sforza, 1983; Wallace et al., 1985). The power of haplotype differentiation along with the high copy number made the mtDNA an ideal molecular tool in the field of forensic genetics, although its most popular use is in its application to questions of human origins and migration history.

### 2.1.1 Non-recombination and Maternal Inheritance

Two of the mtDNA's fundamental characteristics (non-recombination and uniparental inheritance) have been questioned. Several studies suggested that recombination plays a significant role in mtDNA evolution, as shown by a number of mutations shared across haplogroups and a reduction in linkage disequilibrium in mtDNA sequences (Awadalla, Eyre-Walker, & Smith, 1999; Eyre-Walker & Awadalla, 2001; Eyre-Walker, Smith, & Smith, 1999; Hagelberg et al., 1999).

However, evidence in support of this hypothesis is presently lacking. Instead, reanalysis of the data from these studies has shown that inappropriate statistical methods

were used (Jorde & Bamshad, 2000; Kumar, Hedrick, Dowling, & Stoneking, 2000), and errors were likely present in the data sets (Kivisild & Villems, 2000). Additionally, the use of the same analytical techniques with other data sets could not replicate evidence for recombination (Jorde & Bamshad, 2000; Kumar et al., 2000; Parsons & Irwin, 2000). In one instance, the conclusions of a study arguing in favor of mtDNA recombination was redacted because the authors discovered that the evidence for recombination was actually caused by a sequence alignment error (Hagelberg et al., 2000).

In the end, the homoplasies shared between multiple haplotypes could easily have occurred as recurrent mutation in the highly mutable mtDNA. Further study of the recombination issue has led to little, if any, evidence for recombination in mtDNAs. Thus, if recombination does occur, it does so rarely, making its impact at the population level an insignificant factor (Elson et al., 2001; Elson & Lightowers, 2006; Eyre-Walker & Awadalla, 2001; Innan & Nordborg, 2002; Pakendorf & Stoneking, 2005).

Only one instance of paternal leakage has ever been discovered (Schwartz & Vissing, 2002). The paternal mtDNA was found in high proportions in muscle tissue only, while mtDNA from blood, hair, and fibroblasts were exclusively maternal in origin. The mtDNA from the sperm are targeted for elimination (Schwartz & Vissing, 2003), yet this mechanism appeared to fail in this instance. Since the 2002 study, there have been no other reports of paternal leakage in humans, and it is considered so rare an event that maternal inheritance remains the general rule (Pakendorf & Stoneking, 2005).

### 2.1.2 Characteristics of Mutational Change: Clocklike Rates and Time Dependency

The “molecular clock” aspect of mtDNA has been assumed based on the expectation that the mitochondrial genome is a neutral locus and therefore accumulates mutations in a clocklike fashion (Ingman, Kaessmann, Paabo, & Gyllensten, 2000; Kivisild et al., 2006; Mishmar et al., 2003; Soares et al., 2009; Torroni, Neel et al., 1994). The time to the most recent common ancestor (TMRCA) cannot, of course, be accurately ascertained if the molecular evolution of haplogroups deviates from predictable clocklike rates. Likelihood ratio tests (LRTs) are the best estimators of differences in lengths of phylogenetic branches and can, therefore, be used to differentiate between branches evolving at different rates. This is precisely how the clocklike characteristics of mtDNA has been verified (Ingman et al., 2000). If a particular lineage shows evidence of non-clocklike evolution, then there is a possibility that selection is the cause of this deviation, and the mtDNA lineage must be examined more closely.

Still, violations of the molecular clock have been reported for certain mtDNA lineages (Howell, Elson, Howell, & Turnbull, 2007; Howell, Elson, Turnbull, & Herrnstadt, 2004; Torroni, Rengo et al., 2001). Branch length differences in the mtDNA haplogroup L2 caught the attention of researchers who have investigated this haplogroup for non-clocklike sequence evolution. L2 is found mostly in Africa (and the Americas due to the trans-Atlantic slave trade) (Chen et al., 2000; Pereira et al., 2001; Salas, Carracedo, Richards, & Macaulay, 2005; Salas et al., 2002; Salas et al., 2004; Salas, Richards et al., 2005). It is made up of at least four distinct branches (Howell et al., 2004; Torroni, Rengo et al., 2001). These four L2 subhaplogroups were found to have different branch lengths. Two rare branches (L2b and L2d) have been described as

overly derived and, thus, likely victims or beneficiaries of natural selection (Torrioni, Rengo et al., 2001). In another study, the other two branches (L2a, L2c) were implicated, although the original two (L2b and L2d) were not (Howell et al., 2004).

This asymmetry was first noted by Torrioni et al. (2001), who concluded that (among a number of possibilities) differences in selective pressures could cause the observed pattern. Howell et al. (2004) suggested that negative selection acting on the L2a and L2c subhaplogroups caused these differences. Demographic scenarios and population structure were also considered in trying to explain the current genetic variability in these lineages (Howell et al., 2004). That both studies came to different conclusions while using the same methods (LRTs) is interesting.

In general, it does not seem likely that the patterns for L2a and L2c would be due to negative selection, as L2a is one of the most widespread lineages in Africa. It also seems contradictory that the patterns for L2b and L2d would be the result of positive selection, given that they are rather rare lineages. If the accumulated mutations they possess were beneficial, then it is expected that they occur at higher frequencies. These studies show that more analysis is often necessary when violations of the clocklike mutation rates are found within a particular haplogroup, and that definitive conclusions may not necessarily be drawn from the resulting data.<sup>1</sup>

The mutation rate is of critical importance for implementing a molecular clock. Early studies of protein polymorphisms suggested that this clock is roughly linear over time between species (Zuckerandl & Pauling, 1961, 1965); nevertheless evidence of

---

<sup>1</sup> Both studies also commented that the patterns could be due to inaccurate tree construction. This seems unlikely, but the possibility remains that recurrent mutations were counted multiple times per each branch, thus creating a long, seemingly over-derived branch. In fact, this was a criticism of the Howell et al. (2004) paper.

relaxed clocks casts doubt on this linearity. Instead of linear universal clocks, local clocks were proposed to account for apparent differences in phylogenies or between investigated species (Yoder & Yang, 2000). For mtDNA, another aspect that must be considered is the effect of rate heterogeneity among sites (Yang, 1996). Some nucleotide positions tend to mutate faster than others. As a consequence, mutational “hotspots” have been identified in the mtDNA genome, particularly in the hypervariable regions, which are locations for recurrent mutation and sequencing artifacts (Bandelt, Quintana-Murci, Salas, & Macaulay, 2002; Brandstätter et al., 2005; Hagelberg, 2003; Malyarchuk, Rogozin, Berikov, & Derenko, 2002; Stoneking, 2000).

The mutation rates used for human mtDNA studies were estimated from essentially two methods. The first method produced what has been called the “evolutionary” or “phylogenetic” mutation rate. This conventional approach used the divergence between human and other hominoid species to calibrate the mutation rate (Cann et al., 1987; Hasegawa & Horai, 1991; Hasegawa, Kishino, Hayasaka, & Horai, 1990; Hasegawa, Kishino, & Yano, 1985; Stoneking, Sherry, Redd, & Vigilant, 1992; Vigilant et al., 1991). Because of apparent differences in the rate of mutation between the mitochondrial hypervariable and coding regions, substitution rates were estimated from both sections of the mtDNA (Hasegawa, Di Rienzo, Kocher, & Wilson, 1993; Horai, Hayasaka, Kondo, Tsugane, & Takahata, 1995; Kondo, Horai, Satta, & Takahata, 1993; Stoneking et al., 1992; Tamura & Nei, 1993; Torroni, Neel et al., 1994; Wakeley, 1993; Ward et al., 1991). The consensus from these early studies was that the divergence rate of the human mtDNA (which is twice the substitution rate) is roughly 30% per million years (Macaulay et al., 1997). This estimate matches closely with that obtained

from an extensive phylogenetic analysis of Native American mtDNAs (Forster et al., 1996). The substitution rate that was estimated from the Forster et al. (1996) study is one transition per 20,180 years ( $\pm 1,000$  years). This estimate relies on an assumed Beringian expansion around 11,300 years ago (Forster et al., 1996; Saillard, Forster, Lynnerup, Bandelt, & Norby, 2000). The Forster et al. (1996) rate became the standard mutation rate (for sequences between positions 16090 and 16365) for human mtDNA population studies.

The second method used genealogical records to construct pedigrees that included genotype information (Heyer et al., 2001; Howell, Kubacka, & Mackey, 1996; Howell et al., 2003; Parsons et al., 1997; Sigurdardottir, Helgason, Gulcher, Stefansson, & Donnelly, 2000). In all of these studies, the pedigree rate was calculated by counting the number of new mutations between mother and offspring. These studies produced mutation rates that were much higher than the typical evolutionary substitution rates (Heyer et al., 2001; Howell et al., 1996; Howell et al., 2003; Parsons et al., 1997; Sigurdardottir et al., 2000). The ensuing debate between the different rates focused on issues of selection, rate heterogeneity, somatic versus germ line mutations, and even recombination (Heyer et al., 2001; Meyer, Weiss, & von Haeseler, 1999; Sigurdardottir et al., 2000).

The inconsistency between pedigree and evolutionary rates seem problematic at first glance, but this disparity can be explained as differences between mutation and substitution rates (Henn, Gignoux, Feldman, & Mountain, 2009). The pedigree rates should more closely reflect the mutation rate, but not every mutation (or mtDNA lineage) gets fixed in a population. Only women can pass on their mtDNAs, thus any (germ line)

mutation gained by male children will be lost. Furthermore, not every person will have children. In addition, purifying selection can also remove private mutations from the gene pool.

Recently, the time dependency of mutation rates has been investigated in an attempt to uncover a mutation rate that more accurately reflects the accumulated sequence data (Endicott & Ho, 2008; Endicott, Ho, Metspalu, & Stringer, 2009; Henn et al., 2009; Ho & Larson, 2006; Ho, Phillips, Cooper, & Drummond, 2005; Howell et al., 2007; Howell et al., 2004; Ingman et al., 2000; Kivisild et al., 2006; Loogvali, Kivisild, Margus, & Villems, 2009; Soares et al., 2009; Torroni, Rengo et al., 2001). Time dependent mutation rates change over time, but do so in a predictable, non-linear manner. These studies indicate that neither the pedigree nor the phylogenetic rate is appropriate to use in all situations. In reality, the pedigree rates are fastest, reflecting recent historical events, while the substitution rates are slower, reflecting molecular evolution after longer time spans.

The shape of the rate of change is roughly a J-shaped curve (Henn et al., 2009; Ho & Larson, 2006; Ho et al., 2005; Soares et al., 2009). More precisely, the “relationship can be described by a vertically translated exponential decay curve, with the y-axis intercept representing the instantaneous rate of non-lethal mutations and the asymptote representing the substitution rate” (Ho et al., 2006, 80). An estimate of haplotype diversity can then be compared to this curve to obtain a better approximation of the TMRCA. Thus, the difference between a pedigree (mutation) rate and evolutionary (substitution) rate is not problematic from a theoretical perspective. Instead, they merely represent two separate portions of the decay curve.



### 2.1.3 Natural Selection on the mtDNA Genome

The neutrality of mtDNA sequence evolution has been questioned in a number of recent studies (Balloux, Handley, Jombart, Liu, & Manica, 2009; Bazin, Glemin, & Galtier, 2006; Elson, Turnbull, & Howell, 2004; Howell et al., 2007; Kivisild et al., 2006; Mishmar et al., 2003; Moilanen, Finnila, & Majamaa, 2003; Nachman, Brown, Stoneking, & Aquadro, 1996; Ruiz-Pesini, Mishmar, Brandon, Procaccio, & Wallace, 2004; Soares et al., 2009; Stewart et al., 2008; Torroni, Rengo et al., 2001). The mtDNA genome is made up of genes that are involved in the oxidative phosphorylation process that produces cellular ATP. It would make sense then that selection is maintaining the function of these genes that are so critical for survival. The fact that there are mitochondrial diseases implies that some mutations can produce distinctly different products such that selection may act on the molecule. Even so, it is believed that genetic drift has a greater role than selection in passing heteroplasmic pathogenic mutations on to the next generation (Chinnery et al., 2000; see Tuppen et al. 2010 for a review of mtDNA diseases). Furthermore, some have suggested that mtDNA has played a role in adapting to different environments (but see section below).

Two types of analysis have been developed to evaluate whether a given locus is being acted upon by selection. These include tests of neutrality and consideration of nonsynonymous versus synonymous mutations. The two most common tests of neutrality (Tajima's  $D$  and Fu's  $F_S$  statistics) assess DNA sequences for excesses of rare variants. If a particular haplotype has beneficial characteristics and selection increases the frequency of that haplotype, then a proliferation of singleton mutations (rare variants) will be observed, thus producing a significant value for these tests. The problem,

however, is that for tests of neutrality to identify the action of selection, populations must be at mutation-drift equilibrium. When this assumption is not met, the excess of rare alleles could actually be indicative of population expansion. Therefore, the properties that are being investigated with neutrality indices are not just the neutrality or selection of alleles in a population but also whether a population is structured or expanding in size after a bottleneck. Significant negative values can be the result of either selection or population expansion.

The second approach is to examine mutations that could result in a new protein product. Among protein encoded regions of DNA, two types of mutation are possible – those that are effectively silent (synonymous) or those that cause the protein to replace one amino acid for another (nonsynonymous) (Kreitman, 2000). Potentially, nonsynonymous mutations can cause different protein products that have either beneficial or deleterious properties, although they can also create protein products that are essentially identical to the original form. Comparisons of the number of these mutations have been used to assess whether selection has occurred for a given locus and if so, under what type of selection. With a neutral locus, the expectation is that the rate of replacement substitutions (nonsynonymous mutations) approaches the synonymous mutation rate. Therefore, roughly equal numbers of nonsynonymous and synonymous mutations should occur. Given, however, the degenerate nature of the genetic code, there are potentially more nonsynonymous sites per locus than synonymous ones. For this reason, a less biased estimation is the ratio of nonsynonymous mutations per possible nonsynonymous site ( $d_N$ ) versus the number of synonymous mutations per possible synonymous site ( $d_S$ ). If selection was not acting on the locus of a protein-encoded

region, then we can expect there to be no preference between synonymous or nonsynonymous mutations. Fisher's Exact Test can be used to determine if the differences between the two mutation classes are significant (Gerber, Loggins, Kumar, & Dowling, 2001).

This method makes two assumptions. The first is that synonymous mutations are actually "silent." For human mtDNA, this does seem to be the case; there is no evidence of either codon bias or lineage-specific selection (Kivisild et al., 2006; Zeng, Comeron, Chen, & Kreitman, 1998) (but see Yang & Nielsen, 2008). The second is that nonsynonymous mutations represent haplotypes with protein products significantly different from those produced by the ancestral allele – either beneficial or deleterious. The argument goes – if positive selection acted on a locus, then there would be more nonsynonymous mutations than average, resulting in a  $d_N/d_S$  ratio greater than 1. If purifying selection were to occur, then the expectation would be fewer nonsynonymous mutations, and therefore a  $d_N/d_S$  ratio less than 1.

Ratios of the estimates for synonymous to nonsynonymous mutations ( $K_A/K_S$ ;  $d_N/d_S$ ,  $Mn/Ms$ ;  $\omega$ ) have consistently shown that there is a preponderance of synonymous changes throughout the mtDNA genome and phylogeny (Elson et al., 2004; Kivisild et al., 2006; Mishmar et al., 2003; Moilanen, 2003; Pereira et al., 2009; Soares et al., 2009). Many of the nonsynonymous changes are located near the tips of phylogenetic branches, suggesting that these mutations are slowly removed from the populations through purifying selection. Such a pattern could be the end product of purifying selection acting on the mtDNA over thousands of years, or it could be due to a recent relaxation of purifying selection (Gerber et al., 2001; Kivisild et al., 2006; Loogvali et al., 2009). To

gain a better understanding of the mitochondrial genome's mutational characteristics, it would be informative to review in-depth analyses of mtDNA data, with a particular focus on nonsynonymous mutations, since these are the ones most likely acted upon by natural selection.

A recent study of over 5,000 mtDNA genomes provides the most detailed information on human mtDNA properties to date (Pereira et al., 2009). The mtDNA genome can be characterized by protein encoding genes, control regions, tRNA encoding regions and rRNA encoding regions. Considering the 13 protein encoding genes first, substitutions occurred in 36% of the coding region nucleotide positions. The most frequent of these polymorphisms were scattered evenly across all 13 protein-encoding genes. Of this set, 24% occurred in the first codon position, 13% in the second position, and 63% in the third, with a total nonsynonymous to synonymous ratio of 1 to 1.97. The synonymous mutations were correlated with the number of possible synonymous sites, but for nonsynonymous mutations, five genes showed differences from the expected values. ATP6 and ATP8 had higher numbers of nonsynonymous mutations, while CO1, ND4 and ND5 had fewer. Also, one group of amino acids had greater numbers of changes relative to all others. These were mostly changes among neutral nonpolar amino acids (valine, alanine, methionine, and isoleucine) and one neutral polar amino acid (threonine). These kinds of substitutions appear to be tolerated more than other amino acid changes (Pereira et al., 2009).

Looking more closely at the nonsynonymous changes, Moilanen and Majamaa (2003) investigated the physiochemical differences in amino acids as the result of the nonsynonymous substitutions in human mtDNA. While CO1, ND4 and ND5 had the

lowest levels of nonsynonymous mutation in the Pereira et al. (2009) study, they were not in Moilanen and Majamaa (2003). On the other hand, their observation of the two genes with the most nonsynonymous mutations (ATP6 and ATP8) was congruent with the Pereira et al.'s findings. Although, when examining the location of conservative versus non-conservative changes in amino acids due to nonsynonymous substitutions in the mtDNA phylogeny, they noted differences between substitutions that occur internally (shared by many haplotypes) and those that occur at terminal positions (little sharing/private mutations) (Moilanen & Majamaa, 2003). The private (terminal) nonsynonymous mutations showed greater non-conservative changes than internal ones, but changes in size, charge, aromaticity and aliphaticity of the affected amino acids were not significantly different (Moilanen & Majamaa, 2003). High levels of physiochemical changes among amino acids appear to be common in human mtDNA. They concluded that "evaluation of the pathogenicity of an amino acid replacement should not rely solely on these structural considerations" (Moilanen and Majamaa, 2003, 1206).

Thus, haplotypes with nonsynonymous changes do not necessarily produce polypeptides that are drastically different from the ancestral version. Also, there are differences in the type of nonsynonymous mutation (conservative versus non-conservative) depending on the location in the mtDNA phylogeny where they occur. Most non-conservative changes occur at the terminal ends of a branch in only one haplotype and will likely be removed by purifying selection over time. Therefore, entire mtDNA haplogroups are probably not targets for positive selection.

Is there evidence of selection on different haplogroups? By identifying the nonsynonymous changes by haplogroup, it was determined that amino acid replacements

that help to distinguish haplogroups are often found in other portions of the phylogeny. Of these diagnostic haplogroup markers, only two were not conservative in their physiochemical changes. These results point to the neutrality of the nonsynonymous mutations that delineate major haplogroups (Moilanen & Majamaa, 2003). In fact, major haplogroups were found to have relatively similar amino acid sequences. Most of the variation caused by nonsynonymous changes occur within a haplogroup, and often are not shared between samples in that same haplogroup (i.e., the non-conservative substitutions were private mutations). Nevertheless, a significant difference in nucleotide diversity for the ND5 encoding region of haplogroup J was noted, suggesting that selection could potentially act on certain mtDNA lineages (Moilanen et al., 2003).

Of the non-protein coding regions of the mtDNA (control region, tRNA and rRNA), only sections encoding for tRNAs show a bias in the locations of substitutions (Pereira et al., 2009). It is not surprising then that many mtDNA pathologies are the result of mutations in tRNAs – particularly, the stem region (MITOMAP, 2009; Pereira et al., 2009; Tuppen, Blakely, Turnbull, & Taylor, 2010). In fact, Pereira et al. (2009) showed fewer substitutions in stem regions than in loop regions, suggesting evolutionary constraints in the former. The control and rRNA regions showed results consistent with the expected values.

To summarize, the mtDNA genome shows differential patterns in synonymous to nonsynonymous mutations in protein encoded regions and stem to loop mutations in the tRNA-encoded regions. This pattern suggests that selection is conservatively maintaining the polymorphisms in these regions of mtDNA. In other words, selection

will remove deleterious mutations that occur in the mtDNA, but that does not mean that positive selection is selecting for them in particular mtDNAs (Pereira et al., 2009).

Does the bias in nonsynonymous to synonymous mutations suggest that the neutral theory does not apply to human mtDNA? No, it does not. The observed distribution of substitutions among the codon positions of mtDNA proteins is expected based on this theory (Kimura, 1977). The neutral theory states that most mutations that arise are neutral (or nearly so) and those that are not neutral are deleterious (more often than not), which purifying selection will remove from the population (Kimura, 1983; Nei, 1987). Purifying selection on functionally important loci is expected according to this theory. Evidence for this effect was found with the comparison of numbers of mutations at first, second and third codon positions of protein encoded mtDNA genes. Any mutation at the third position is a synonymous change, thus the frequency of mutational occurrence is highest. Any mutation at the second position produces a nonsynonymous change. As predicted, the lowest occurrence of mutations is at the second position, with the first position at an intermediate level between the other two.

The neutral theory and the utility of mtDNA studies were recently questioned. A fundamental tenet of this theory is that, as the effective population size or the mutation rate increases, heterozygosity will increase as well. Bazin et al. (2006), however, claim that there is no correlation between mtDNA genetic diversity and population size. To support this claim, they compared average genetic diversities among taxa (Bazin et al., 2006). Estimates from nuclear loci did not differ from the expected pattern – invertebrates had higher diversity values compared to vertebrates. By contrast, mtDNA estimates were similar across taxa but highly variable between species within a group.

To determine what selective process caused this pattern in animal mtDNAs, Bazin et al. (2006) calculated neutrality indices (NI) for these species. Here, the ratio of nonsynonymous to synonymous mutations within a species was compared to this ratio between species (Bazin et al., 2006). NI values equal to 1 were considered to be evidence for neutral evolution, whereas values over 1 were viewed as evidence for purifying selection and values under 1 support for adaptive (positive) selection. Invertebrates were found to have NI values less than 1, indicating positive selection on their mtDNAs (Bazin et al., 2006).

Despite the final sentence of the abstract where Bazin et al. (2009) state that their paper challenges the neutral theory of molecular evolution, the authors use the neutral theory extensively as a null hypothesis and even provide evidence (from the nuclear loci) that the theory is correct (570). The question then is what makes the mtDNA diversity pattern different from nuclear loci?<sup>2</sup> The mtDNA genome has characteristics that make it distinct from nuclear loci. For example, the typical mtDNA haplotype diversity range for human populations is between 0.87 and 1.00, where 0.87 is considered severely reduced haplotype diversity (Derenko et al., 2003). It has a fast mutation rate, site heterogeneity is typical, and mutation saturation is common, resulting in generally very high within-group mtDNA diversities. Therefore, estimates like the NI ratios are inappropriate for distantly related taxa, where a bias of NI values less than 1 occurs (Wares, Barber, Ross-Ibarra, Sotka, & Toonen, 2006).

---

<sup>2</sup> It is important to note that Bazin et al., (2006) did not state that vertebrate mtDNAs are affected by positive selection; it is only invertebrate mtDNAs (Eyre-Walker, 2006). Therefore, these findings would not affect human population studies based on mtDNA data.



Furthermore, NI suffers from the same problem as the CI ratios mentioned above (A. L. Hughes, 2008). Being highly dependent on the denominator, NI is undefined in any instance where there are zero synonymous differences within or between species, or zero nonsynonymous differences between species. Regardless, NI values less than 1 would be expected for human mtDNAs according to the neutral theory, since this pattern can be easily explained as a fixation of deleterious mutations during a bottleneck (A. L. Hughes, 2008).

Despite the problems with the NI ratios, there are at least three other possibilities as to why the invertebrate mtDNAs showed evidence of adaptive selection. First, the choice (or length) of mtDNA genes used for comparisons could have affected the results (Meiklejohn, Montooth, & Rand, 2007). Secondly, nucleotide bias can affect the estimation of  $d_N/d_S$  between species (Albu, Min, Hickey, & Golding, 2008). Albu et al. (2008) showed that there is a nucleotide bias between invertebrate and vertebrate mtDNAs and that it could explain the observed differences between taxa. Finally, some invertebrate mtDNAs can be infected with the bacteria *Wolbachia*, which may have altered invertebrate species mtDNA diversities as results of selective sweeps (Hurst & Jiggins, 2005).

To reaffirm the relationships between mtDNA diversity and population size, Mulligan et al. (2006) analyzed data from placental mammal species (Mulligan, Kitchen, & Miyamoto, 2006). Comparisons between allozyme diversity and mtDNA synonymous diversity showed that both estimates were positively correlated, and the same was true for allozyme diversity and mtDNA total diversity. Thus, mtDNA diversity is related to population size in species with known or smaller expected populations (Mulligan et al.,

2006). Other investigators have also shown that human effective population sizes are correlated with mtDNA diversity (Atkinson, Gray, & Drummond, 2008). Furthermore, Eyre Walker (2006) noted, “it may be that humans have such small effective population sizes that adaptive evolution in the mitochondrial genome is very rare; the neutrality index in human mitochondrial DNA...gives no indication of adaptive evolution” (Eyre-Walker, 2006, 538).

#### 2.1.4 Climate and Positive Selection on mtDNA

There is little evidence that positive selection affects variation found in human mtDNA. As mentioned above, two studies found inconsistencies in a clocklike mutation rate for branches of haplogroup L2, although they did not agree on which branches natural selection had acted, nor what adaptation was provided by the mutations on each branch (Howell et al., 2004; Torroni, Rengo et al., 2001). Several studies have noted differences in nonsynonymous rates between humans and chimpanzees (Kreitman, 2000; Nachman et al., 1996), but these studies relied on McDonald-Kreitman (MK) tests, which are not appropriate for mtDNA analysis and may be affected by demography (Gerber et al., 2001; Parsch, Zhang, & Baines, 2009).

To date, the only significant effort to investigate positive selection on mtDNAs is the argument that the patterns of mtDNA variation are the result of adaptations to climate. Mishmar et al. (2003) was the first study to test this hypothesis. They noted the stark contrast between the diversity of mtDNAs in northeastern Africa compared to the rest of the world, where essentially only two haplogroups (macro-haplogroups M and N) were successfully established. Furthermore, haplogroup frequencies of A, C, D and G

differed remarkably between Central Asia and Siberia. To explain this pattern, they proposed that the difference between the two regions was the result of enrichment of these haplogroups due to positive selection. Specifically, since the mtDNA is responsible for the production of ATP (and consequently produce heat as a side product), they believe that in “arctic populations” mtDNAs are selected for possessing mutations that decouple the ATP production process. Under this hypothesis, the inefficiency of ATP production would increase the amount of available body heat needed for survival in arctic climates. To support the hypothesis, they cited a study that concluded indigenous Siberians have higher basal metabolic rates (Leonard et al., 2002). Others have continued this same line of questioning (Balloux et al., 2009; Ruiz-Pesini et al., 2004).

Two of these studies used the same general methodology, employing the number of differences between nonsynonymous and synonymous mutations to assess the amount and type of selection acting on each mtDNA haplogroup (Mishmar et al., 2003; Ruiz-Pesini et al., 2004). One study estimated  $K_A$  and  $K_S$  for each of the 13 mtDNA genes and the significance of these indices with Wilcoxon rank-sum tests (Mishmar et al., 2003). The other study compared the number of nonsynonymous mutations (normalized by the number of synonymous mutations) at internal (I) and external/terminal (T) segments of the mtDNA haplogroup phylogenies in the form of I/T ratios, with significance estimated by Fisher’s Exact Tests. They also calculated conservation indices (CI) by comparing the conservation of replacement amino acids across a number of species, with significance estimated using t-test p values (Ruiz-Pesini et al., 2004).

Mishmar et al. (2003) found three genes that had higher variability in “arctic populations.”<sup>3</sup> The three genes were ATP6, CO3 and ND6. Four nonsynonymous changes were identified in ATP6. One belonged to a diagnostic mutation for M8 (which includes haplogroups C and Z)<sup>4</sup>, while the other three mutations occurred in haplogroup N. Of the latter three, one belonged to haplogroup A and another to haplogroup N1b, while the third was a defining mutation for the entire macrohaplogroup N. Of these examples, C, Z and A were defined as “arctic,” whereas N1b was considered “temperate.” All of the members of macrohaplogroup N were classified as “temperate,” except for haplogroups A, X and Y. The final nonsynonymous polymorphism in macrohaplogroup N is found in every haplogroup in Europe (including all “temperate populations”). Therefore, the nonsynonymous mutations in ATP6 are not necessarily associated with haplogroups found in high frequencies in colder climates. They also noted that mutations in ATP6 were slightly elevated in “tropic populations” (Mishmar et al., 2003). Given that nonsynonymous mutations in ATP6 occur among all types of haplogroups regardless of climate, these findings suggest a relaxation in selective constraints (purifying selection) for that gene.

It should also be noted that when the Mishmar et al. data were reassessed using the more appropriate Fisher’s Exact Test to determine significance, only the ATP6 gene showed significant values (Elson et al., 2004). In addition, northern Eurasia was inhabited only relatively recently (tens of thousands of years ago), but the nonsynonymous mutations in ATP6 occurred gradually and were not episodic as would

---

<sup>3</sup> Note: Mishmar et al. (2003) readily equated the terms “populations” and “haplogroups.”

<sup>4</sup> Mishmar et al. (2003) did not consider M8 haplotypes in this analysis, but they are typically found throughout China and are generally considered “East Asian.”

be expected if they were selected for as an adaptation to the newly inhabited colder climates (Ingman & Gyllensten, 2007a).

Ruiz-Pesini et al. (2004) also found higher ratios of nonsynonymous mutations and higher conservation indices in “arctic” haplogroups. I/T ratios for A, C, D, and X (1.01, 1.10, 0.91 and 2.91, respectively) were greater than those for B or L (0.75 and 0.70, respectively). Furthermore, conservation indices for nonsynonymous mutations on internal branches were stated as being higher in arctic haplogroups A (53%) and C (73%) compared to temperate/tropical haplogroups B (31%) and L (36%). They concluded that this is strong evidence for adaptive selection on “arctic” mtDNAs.

According to the data of Ruiz-Pesini et al, however, haplogroups V/HV, J, and W also have high I/T ratios (2.89, 1.02 and 1.67, respectively), and haplogroups H, J, T and I/N1b also have high conservation indices for internal mutations (46%, 42%, 52%, and 50%, respectively). They stated that some “European” (i.e., temperate) haplogroups also have higher ratios of nonsynonymous to synonymous mutations than “African” (i.e., tropic) haplogroups, but this was attributed to “episodic periods of cold associated with the repeated continental glaciations” (Ruiz-Pesini et al., 2004, 224). Strangely enough, haplogroup U was not implicated even though it has been in Europe for the longest period of time (Krause, Briggs et al., 2010; Richards et al., 1996; Richards, Macaulay, Bandelt, & Sykes, 1998).

Ruiz-Pesini et al. further examined the conservation indices of each internal nonsynonymous mutation in “arctic” haplogroups in an effort to identify positively selected mutations. They found four: two in haplogroup A (one in the ND2 gene and one in the ATP6 gene) and two in haplogroup C (one in ND4 and one in CytB). It should be

noted that only one of the four mutations implicated in Mishmar et al. (2003) was amongst them. Of these four mutations, the one found in ATP6 had the lowest conservation index.

Objections have been raised against the conclusions of these studies (Elson et al., 2004; Elson, Turnbull, & Taylor, 2007; Ingman & Gyllensten, 2007a; Kivisild et al., 2006; Sun, Kong, & Zhang, 2007). One problem with the abovementioned studies is the classification of haplogroups into populations. Only in extremely small and isolated populations have researchers found that entire populations are represented by single haplogroups – for example, haplogroup B in Polynesia or haplogroup D in Commander Islands (Derbeneva, Sukernik et al., 2002; Friedlaender et al., 2007; Friedlaender et al., 2005). The second problem is classifying haplogroups into discrete climate categories accurately and consistently. The “arctic” category included A, C, D, G, X, Y and Z in Mishmar et al., but Ruiz-Pesini et al. only included A, C, D and X. While A, C and D are certainly found throughout Siberia, X is only found in Altaians, Evenks and Khants of Siberia, which live adjacent to the steppe of Central Asia, the taiga of Siberia, and the mountains between Mongolia and Russia (Derenko et al., 2003; Reidla et al., 2003). Haplogroup X is found mostly in the Middle East (its putative origin), which can hardly be considered an arctic environment. In addition, haplogroups C and D are found throughout Central and East Asia and southern Siberia, not just subarctic and arctic regions of Siberia.

Certainly, portions of northern Europe have experienced similar climatic conditions as Siberia. In addition, regional distributions of some haplogroups, like haplogroup U, also exist, such that U1, U2, U3, and U7 are located in different regions

than U4 and U5. Some are even located in very different climates (U6 in northern Africa and U4, U5 and U7 in eastern and northeastern Europe and western Siberia). Therefore, placing all U sub-haplogroups into the same climatic category is not justified.

Instead of placing haplogroups into three climatic categories, they can just as easily be separated into two temporal categories (newer and older) (Elson et al., 2004; Kivisild et al., 2006; Zeng et al., 1998). A higher number of nonsynonymous changes in the younger haplogroups are expected since purifying selection has not yet completely removed all deleterious mutations. Because this pattern is consistent regardless of the geographic location of any haplogroup, it seems the better and more parsimonious interpretation. Ruiz-Pesini et al. (2004) essentially tested this very property, although caution must be used when assessing ratio data. “Such a measure, being a ratio of ratios, compounds the statistically undesirable properties of ratio data. Ratios are sensitive to stochastic error, particularly in the denominator” (Hughes, 2008, 170). Ultimately, few differences exist in the number of nonsynonymous to synonymous mutations between regional sets of haplogroups for the same gene (Sun et al., 2007). Thus, if there are no significant differences between genes of haplogroups from East Asia, South Asia, Oceania, and Europe, then it seems unlikely that climate has much to do with mtDNA variation.

One important aspect not taken into consideration when comparing levels of nonsynonymous substitutions in mtDNA genes is the difference between core functional domains and the remainder of a gene. These comparisons showed significant differences among the various gene segments in the mtDNA (Ingman & Gyllensten, 2007a). Nonsynonymous substitutions occurred more often in segments outside of the core

functional domains. This trend was found in interspecies comparisons as well. Furthermore, genes that have disproportionately smaller core functional domains relative to the total gene size have more nonsynonymous mutations. These genes are the same as those implicated by Mishmar et al. (2003) – ATP6, CytB and ND3 – providing more evidence of a relaxation of purifying selection on ATP6.

Difficulties in using the numbers of nonsynonymous and synonymous mutations as a means of understanding selection within species have been documented (A. L. Hughes, 2008; Kryazhimskiy & Plotkin, 2008; Rocha et al., 2006). To avoid this issue altogether, one study attempted to test the correlations between mtDNA diversity (defined by the number of average pairwise differences in a population) and climate (as defined by the lowest yearly temperature in the location of the sampled population<sup>5</sup>) (Balloux et al., 2009). After accounting for associations between genetic diversity and geography (as defined by distance of the sampled population from Ethiopia), it was claimed that within population mtDNA diversity (as expressed with pairwise differences in HVS1 sequences and complete genomes) decreased with climate (Balloux et al., 2009).

Despite the assertions of this study, the data do not show a strong correlation between mtDNA and climate. The  $R^2$  values for the mtDNA tests were quite low (even though they were significantly different from zero), yielding values of approximately 0.2 for associations between mtDNA diversity with geography and mtDNA diversity with climate and geography. For comparative purposes, the genetic diversity at nuclear loci was assessed in relation to geography and climate. The  $R^2$  values for X-chromosome

---

<sup>5</sup> It is not clear that minimum temperature actually represents climate well, because annual average temperatures and the range of variability in these temperatures is not considered.



STRs, autosomal STRs and autosomal SNPs were all over 0.8 for associations with geography, but showed no correlation with climate. Given the low  $R^2$  values for the mtDNA tests, these findings actually show that geography and climate are poor predictors of mitochondrial diversity (specifically, the number of pairwise differences).

Balloux et al. (2009) did not believe this pattern was due to small population sizes at higher latitudes because they could not find similar associations with nuclear loci. In fact, the mtDNA does have a lower effective population size compared to autosomal loci. In addition, the populations from the coldest climates in northeastern Siberia and northwestern Americas used in this analysis came from small, isolated populations, a factor that by itself could account for finding relatively low genetic diversity estimates in colder climates.

One way to test this hypothesis is to compare the results to another locus that has an equally small effective population size – in this case, the Y-chromosome. The correlation tests using Y-chromosome microsatellite diversity and geography had an even lower  $R^2$  value ( $R^2 = 0.164$ ) than the results for the mtDNA. The correlation between Y-chromosome diversity and climate was also non-significant. Therefore, these results indicate that lower effective population sizes allow stochastic processes to affect these loci more than autosomal loci. They also demonstrate that geography is a poor predictor of mitochondrial or Y-chromosomal diversity, and that climate is an equally poor predictor of mtDNA diversity. In any case, correlation does not necessarily equate with causation.

If climate does not play a role in determining the level of mtDNA diversity in a population, then it remains to be explained why there is a difference between the results

of nuclear loci (which show no correlation) and mtDNA (which shows extremely weak correlation). The answer may be as simple as sampling bias. All nuclear loci data were obtained from the HGDP-CEPH human genome diversity panel database (<http://www.cephb.fr/en/hgdp/>) (Balloux et al., 2009). This panel includes 963 people from 51 populations. Of these 51 populations, only two (Russian, Yakut) reside in the subarctic zone (between 50 and 70 degrees latitude), and make up only 5% of the total sample. The mtDNA data, on the other hand, come from HVRBase and the Human Mitochondrial DNA Database. Out of 109 populations represented in this database, 14 are located in the subarctic region, and make up 17% of the mtDNA samples archived there.

The study also searched for SNPs that were correlated with temperature (Balloux et al., 2009). Out of the 34 polymorphic sites examined, only five were nonsynonymous. Of those five nonsynonymous mutations, two supposedly showed a significant correlation to minimum temperature – 8701 and 10398. Both of these SNPs occur at the root of macrohaplogroup N. Therefore, their distributions should be correlated as they both define the same macrohaplogroup. It is less obvious as to how they could have been associated with cold climates. The derived alleles occur more frequently in the Middle East and Europe, almost to the exclusion of the ancestral allele (Macaulay et al., 1999; Quintana-Murci et al., 2004; Richards et al., 1996; Richards et al., 2000; Richards et al., 1998; Torroni et al., 1998; Torroni, Bandelt et al., 2001; Torroni, Lott et al., 1994; Torroni et al., 2000). Given that all Middle Eastern and European populations have haplogroups with 8701 and 10398, it is not clear how these SNPs could be associated with adaptation to colder climates.

If there is a correlation between these two SNPs and minimum temperature, then we could expect the European populations living in locations with the lowest temperatures in Europe (which are in northeastern Europe) to have the highest frequency of these SNPs (high frequencies of macrohaplogroup N). Nevertheless, these groups actually have the lowest overall frequency of N-derived haplogroups and higher frequencies of M-derived haplogroups (Ingman & Gyllensten, 2007b; Lappalainen et al., 2008; Tambets et al., 2004). Consequently, this predication is not supported by the current data sets.

Most haplogroups in northern Asia belong to macrohaplogroup M, with only a portion of them carrying the two polymorphisms. One N-derived haplogroup (A) is found in higher frequencies in northeastern Asia and northwestern North America. These small, isolated populations, however, are known to be related and historically linked, and they have been severely affected by genetic drift (Crawford, 2007; Crawford et al., 1997; Derbeneva, Sukernik et al., 2002; O'Rourke, Hayes, & Carlyle, 2000; Schurr et al., 1999; Starikovskaya et al., 1998; Sukernik et al., 1981; Sukernik & Osipova, 1982; Sukernik, Osipova, Karafet, Vibe, & Kirpichnikov, 1986; Sukernik, Vibe, Karafet, Osipova, & Posukh, 1986; Volod'ko, Eltsov, Starikovskaya, & Sukernik, 2009; Volodko et al., 2008).

An attempt was made to determine whether a single population skewed the results by reassessing the correlations after removing one population at a time (Balloux et al., 2009), but the signal remained because multiple populations from northeastern Siberia were used in the study – including three Koryak populations that are essentially identical in mtDNA diversity (Schurr et al., 1999; Volodko et al., 2008). Balloux et al. noted the strong influence of these populations on the analysis, yet results describing the

correlations between mtDNA and minimum temperature when excluding the Beringian and American populations were not included (Balloux et al., 2009). Phylogenetic analysis of the two SNPs and the population genetic analysis of the groups living around the Bering Strait provide enough information to demonstrate that the signals displayed by these two polymorphisms are the result of population histories, not climatic adaptations.

If differences in the activity and/or function of mtDNA protein products exist between haplogroups or haplotypes because of nonsynonymous changes, then we could expect to see differences in the relative fitness caused by these respective mtDNAs. In this regard, the hypothesis used throughout the abovementioned studies cites differences in basal metabolic rates (BMR) among indigenous and non-indigenous populations as a result of mtDNA uncoupling to adapt to cold climates (Mishmar et al., 2003; Ruiz-Pesini et al., 2004). The study that is referenced with regard to these differences examined the BMRs of indigenous groups (Inuit, Evenk and Buryat) and non-indigenous controls living in the same locations (Leonard et al., 2002). Three different methods were used to assess BMR levels. Expected BMRs were predicted from standard recommendations and compared to observed rates (Consolazio, Johnson, & Pecora, 1963; Poehlman & Toth, 1995; Schofield, 1985). Indigenous men had significantly elevated BMRs in all three measures relative to predicted rates. Non-indigenous men had significantly elevated BMRs in two out of three measures, but in no instance were the BMRs of indigenous and non-indigenous men significantly different from each other. For the females, indigenous women had significantly elevated BMRs in all three instances. Non-indigenous women had significantly lowered BMRs in two out of three instances. In all cases, indigenous and non-indigenous women had significantly different BMRs.

These observations suggested that sex differences in metabolic rates might be present (in particular, it was non-indigenous women whose rates varied from other groups). However, given the evidence from men, it is difficult to reconcile a significant difference between indigenous and non-indigenous BMRs. Furthermore, Leonard et al. (2002) cited evidence that suggests the level of thyroid hormones is positively correlated with BMRs and seasonal fluctuation. It appears that other genes are involved in effecting basal metabolic rates. Even still, there may not be a clear adaptive difference between indigenous and non-indigenous genotypes. As no study has collected BMR and mtDNA data from the same individuals, it is difficult to assess this argument empirically.

To evaluate empirical differences between mtDNAs, one study used molecular kinetic analysis of the mitochondrial oxidative phosphorylation process to determine whether mtDNA haplotypes provide different phenotypic effects (Amo & Brand, 2007). This method employed cybrids that had different mtDNAs added to cells with the same nuclear genetic background. In the study, haplogroups A, C, D, L1, L2 and L3 were compared to each other. These mtDNAs represent three of the “arctic” and three “tropical” haplogroups of Mishmar et al. (2003) and Ruiz-Pesini et al (2004). If the climatic adaptation hypothesis were true, then the expected result would be a lower ATP coupling efficiency in the “arctic” relative to the “tropical” haplogroups. Amo and Brand (2007) found no significant difference between these mtDNAs. Only one haplogroup showed a slightly lower coupling efficiency – a “tropical” haplogroup – but it was not significantly lower. Therefore, this study showed no difference in the performance of arctic and tropical haplogroups. Along with the evaluation of previous studies, there

currently is no evidence of positive selection having selected human mtDNAs for adaptation to cold environments.

## **2.2 The Y-Chromosome**

The Y-chromosome is the second smallest nuclear chromosome and one of two sex chromosomes, with about 60 Mb (Mb = one million base pairs) of euchromatin sequence (Jobling & Tyler-Smith, 1995; Skaletsky et al., 2003). During meiosis, the X- and Y-chromosomes pair together, but the vast proportion of these two chromosomes is not homologous, thus little crossover occurs between them.

The relatively recent full sequencing of the Y-chromosome has allowed for a fuller understanding of the molecule and its gene products (Skaletsky et al., 2003). Three classes of sequence have been identified from the euchromatin making up the Y-chromosome – X-transposed, X-degenerate and ampliconic (Skaletsky et al., 2003). Given that many of the Y-chromosome genes are degenerate forms found on the X-chromosome, it has been suggested that the sex chromosomes evolved from autosomes, with the X-chromosome retaining the majority of its functionality and the Y-chromosome being the degenerate or degraded form (Graves et al., 1995; Lahn, Pearson, & Jegalian, 2001). The nature of the evolution of the Y-chromosome has specific implications for its function and for its use in genetic studies.

The X-transposed genes of the Y-chromosome are unique to humans. They resulted from a rare crossover event that occurred sometime after the split between humans and chimpanzees. Consequently, the Y-chromosome has 99% identical DNA sequences with the Xq21 location on the X-chromosome, which also contains only two

genes and a large number of interspersed repeat elements (Skaletsky et al., 2003). Although the Neanderthal genome has been published (Green et al., 2010), the Y-chromosome from this hominin has not been extensively described. Therefore, it is not currently known whether this chromosome translocation exists in Neanderthals as well.

The other two classes of sequence contain the majority of Y-chromosome genes. The X-degenerate class has 27 protein products that are homologous to those of X-linked genes. Of these 27 genes, 14 are pseudogenes, while 13 appear to produce functional products nearly identical to the proteins from the X-chromosome (Skaletsky et al., 2003). All but one of these functional genes are expressed throughout the body, the exception being the Sex determining Region Y (SRY) gene, which is expressed in the testes and is responsible for the developmental catalyst of male characteristics. The remaining class – ampliconic sequences – contains the highest density of genes, which belong to nine protein gene families (Skaletsky et al., 2003). These genes are found in multiple copies, but are only expressed in the spermatogenic cells of the testes.

Segments of the Y-chromosome can be classified in two categories. The tips of the Y-chromosome contain pseudoautosomal regions (PAR1 and PAR2), which bear sequences homologous with the X-chromosome. Necessary for proper segregation during meiosis, these small regions of the Y-chromosome (about 5% of the entire chromosome) can recombine with the X-chromosome. The remainder of the Y-chromosome does not recombine with the X. Fittingly, it was named the non-recombining region of the Y-chromosome (NRY) (Hammer & Zegura, 1996), although some prefer to call it the male-specific region (MSY) because ancient recombination events helped to create this segment (Rozen et al., 2003; Skaletsky et al., 2003).

The consequence of having a pair of non-identical sex chromosomes becomes readily apparent. The X-chromosome is free to recombine with other X-chromosomes in females, increasing the amount of new combinations of X-linked genes. As a result, the X-chromosome in many ways resembles other autosomes in size, function and ability to produce viable gene products and handle deleterious mutations. By contrast, the Y-chromosome is in essence stuck with the hand it was dealt. The dissimilarity in chromosomal structure reduces the chance of crossover events to an extraordinary rarity. It was argued that the inevitable risk of Muller's ratchet (Muller, 1918, 1964) would eventually reduce the Y-chromosome to a non-functional "genetic wasteland" (Charlesworth & Charlesworth, 2000; Skaletsky et al., 2003). Instead, variation on the Y-chromosome originated from (and is maintained by) an increased reliance on gene conversion, as evidenced by the many copies of identical or nearly identical genes throughout the ampliconic sections of the Y-chromosome (Rozen et al., 2003; Skaletsky et al., 2003). While the risk of obtaining deleterious mutations remains, their effects can be mitigated by the redundancy of gene products from the Y-chromosome (Marais, Campos, & Gordo, 2010) but also those from the X-chromosome and some autosomes.

The physical characteristics of the Y-chromosome make it especially useful for inferring population histories (Jobling & Tyler-Smith, 1995, 2003). It is essentially a haploid genetic system that is passed down from father to son without the obscuring effects of recombination. Being able to follow the line of paternal transmission and place the Y-chromosomes unambiguously into a phylogeny makes it particularly useful. In addition, the Y-chromosome is present only in males and only in one copy. Thus, it has a smaller effective population size as compared to autosomal DNA and is more susceptible



to genetic drift. Males also have a high variance of reproductive success, which further reduces the effective population size (Rozen, Marszalek, Alagappan, Skaletsky, & Page, 2009; Tyler-Smith, 2008).

Two classes of mutation allow for the creation of high-resolution NRY phylogenies. The first type is biallelic markers – sometimes called unique event polymorphisms (UEPs) in the NRY literature – which mostly consist of single nucleotide polymorphisms (SNPs) and small nucleotide insertions and deletions (indels). These types of mutations occur at a low rate. A recent estimate from next generation sequencing data is  $3.0 \times 10^{-8}$  mutations/nucleotide/generation (Xue et al., 2009). Given the length of the NRY (60 MB), it is unlikely that recurrent mutation will occur over the course of modern human history. Therefore, SNP and indel mutations on the Y-chromosome generally conform to the infinite-sites and infinite-allele assumptions, unlike the mtDNA where higher rates of recurrent mutation are found. Biallelic markers provided the fundamental basis for understanding human NRY diversity and its applications in evolutionary and population genetic studies (Hammer, 1995; Hammer et al., 1998; Hammer et al., 2001; Hammer et al., 1997; Hammer & Zegura, 1996; Jobling & Tyler-Smith, 2003; Shen et al., 2000; Underhill et al., 1997; Underhill et al., 1996; Underhill et al., 2001; Underhill et al., 2000).

The second mutation class found in the NRY is microsatellites composed of short tandem repeats (STRs) (Goldstein et al., 1996; Roewer et al., 1996; Ruiz Linares et al., 1996). These mutations are scored as the number of repeats of a given sequence (typically a 2-6 base pair sequence although some are more complex consisting of compound STRs). Mutations in STRs have been modeled as following a simple single-

step extension or deletion of a repeat unit. Given the architectural configuration of these mutations, repeat differences occur at a much higher rate than for biallelic markers. In combination, these mutation classes provide a high-resolution phylogeny for human Y-chromosomes (de Knijff, 2000; Hurles et al., 1999; Ramakrishnan & Mountain, 2004).

### 2.2.1 Y-STR Mutation Rates

Given that microsatellites tend to evolve much more quickly than biallelic markers, they can be extremely useful in elucidating relationships between closely related populations and individuals (Bowcock et al., 1994). Previous studies have shown that the amount of differences between any two microsatellite haplotypes is linear with time (Goldstein, Ruiz Linares, Cavalli-Sforza, & Feldman, 1995; Slatkin, 1995). While Goldstein et al. (1995) assumed a multinomial distribution and a strict stepwise mutation model, Slatkin (1995) derived his calculation using coalescent theory and employed a model that allowed mutations of multiple repeat differences.

Much like the mtDNA mutation rates, mutation rates for Y-STRs have been calculated using pedigree and phylogenetic methods. Pedigree rates were calculated by examining the number of changes in STR scores between deep-rooted pedigrees and father and son pairs (Heyer, Puymirat, Dieltjes, Bakker, & de Knijff, 1997; Kayser et al., 2000). The average for 15 Y-STR loci was  $2.8 \times 10^{-3}$ /locus/20 years (Kayser et al., 2000). Application of the pedigree rate, however, proved insufficient when considering phylogenies already constructed with biallelic markers. For this reason, an attempt was made to determine Y-STR mutation rates from phylogenetic networks using an internal

(archaeological) calibration point (Forster et al., 2000). This rate proved to be about 10 times slower than the pedigree rate ( $2.6 \times 10^{-4}$ /locus/25 years).

To address this discrepancy and the occasional problem of multi-step repeat changes, Zhivotovsky et al. (2004) estimated an effective evolutionary mutation rate. This estimate is the product of the mutation rate and the variance of mutational changes in repeat scores (Zhivotovsky et al., 2004). The effective mutation rate was calibrated from three archaeologically and historically known colonization events: the settlement of New Zealand and the Cook Islands and the migration(s) of Roma populations into Europe. In addition, Y-STR variation was compared to autosomal STR variation from global collection of over 50 populations (Zhivotovsky et al., 2004). Based on these comparisons, the effective mutation rate was estimated at  $6.9 \times 10^{-4}$ /locus /25 years.

Several possible explanations for the discrepancy between pedigree and phylogenetic mutation rates have been proposed. Much like the issues with mtDNA mutation rates, there was discussion over differences in rates between loci, recurrent mutations and/or parallel mutations among microsatellites (Zhivotovsky et al., 2004). It is also possible that, as for the mtDNA, purifying selection and/or variance in reproductive fitness eliminates some of the novel repeat score changes, although this has yet to be shown experimentally.

### 2.2.2 Selection on the Y-Chromosome

Neutrality of the Y-chromosome is generally assumed when considering population histories and dispersal routes of human populations. Yet, some have questioned the selective neutrality of the Y-chromosome and, therefore, its utility in

investigating male population histories. The lack of recombination (one of the advantageous features of the Y-chromosome that made it constructive for studies of population histories) makes the molecule prone to the accumulation of deleterious mutations. If, however, a mutation happens to be beneficial, then positive selection can drive it to fixation. This process, called genetic hitchhiking, would result in the fixation of the haplotype including the beneficial mutation and all associated deleterious mutations (Rice, 1987). After fixation, the positive selection would no longer be effective on the haplotype because there would be no standing variation in the population (Charlesworth & Charlesworth, 2000).

Sequence comparisons within and between species were used to assess the influence of selection on the Y-chromosome, and determine how it might have shaped NRY variation. Several studies have estimated the number of nonsynonymous and synonymous mutations between species. One examined partial sequences of DAZ genes on the Y-chromosome using a maximum likelihood approach and found weak positive selection acting on the gene (Bielawski & Yang, 2001). An elevated number of nonsynonymous mutations were also found between human and chimpanzee DNA sequences (Bielawski & Yang, 2001). A separate study also found more nonsynonymous mutations between humans and chimpanzees, but the number between humans and mouse was extremely low (Agulnik et al., 1998). In addition, comparisons of relative substitution rates between the autosomal DAZL1 gene (the result of a translocation event after the split of New and Old World monkeys) and DAZ genes on the Y-chromosome showed no functional constraint on the DAZ genes but purifying selection on the DAZL1

(Agulnik et al., 1998). Thus, the higher  $K_A/K_S$  ratios were due to a relaxation of functional constraints.

Similarly, another study compared  $K_A/K_S$  ratios of a few Y-chromosome genes across species and found elevated rates of nonsynonymous mutations as compared to the X-chromosome versions (Gerrard & Filatov, 2005). This pattern is also found when comparing longer DNA sequences of chimpanzee and human Y-chromosomes. The researchers concluded that the pattern could have been caused by positive selection, the relaxation of negative selection, or both (Kuroki et al., 2006).

Some of these studies would suggest that the human Y-chromosome was acted upon by positive selection. Certainly, positive selection must have played a role in the speciation of hominoids. With more Y-chromosome sequence available from chimpanzees, it is clear that profound differences exist between *Pan* and *Homo* (J. F. Hughes et al., 2010). While it is possible that positive selection was involved in the divergence of human Y-chromosomes during speciation, fixed substitutions between species indicate nothing about the variation within a species. Only by examining the variation within humans can we understand how selection might be biasing the results of phylogeographic studies complementing the intraspecies comparisons.

As a general feature, sequence analysis of human Y-chromosomes has shown that their standing variation is greatly reduced with respect to autosomal chromosomes (Hammer, 1995). Different studies have estimated the amount of sequence diversity among human Y-chromosomes and the amount of sequence divergence between species. They used Hudson-Kreitman-Agaudé (HKA) tests, which can be useful for identifying reductions in genetic diversity due to population bottlenecks or selective sweeps, and

found that neutrality could not be rejected (Hammer, 1995; Jobling & Tyler-Smith, 2000). To explain these findings, it was suggested that a selective sweep was responsible for the low amounts of variation on the Y-chromosome (Dorit, Akashi, & Gilbert, 1995; Whitfield, Sulston, & Goodfellow, 1995), but these earlier studies relied upon relatively few samples and short DNA sequences, hence the discovery of relatively few mutations.

Recent studies have reported hundreds of polymorphisms in the NRY, with more being found every day (Karafet et al., 2008). In fact, NRY sequencing studies have shown that there is an excess of low frequency mutations in the Y-chromosome, which gives negative Tajima's D values when tested for neutrality (Hammer et al., 2001; Shen et al., 2000; Thomson, Pritchard, Shen, Oefner, & Feldman, 2000). This pattern could be attributable to selection, differences in effective population size, differences in male and female dispersal rates, or population structure (Hammer, Blackmer, Garrigan, Nachman, & Wilder, 2003; Hammer et al., 2001).

Ultimately, it is believed that drift plays a more significant role than selection in shaping Y-chromosome diversity, mostly because of reduced effective population sizes (Hammer, 1995; Shen et al., 2000; Underhill et al., 1997). Ratios of diversity among human Y-chromosomes and the nucleotide diversity between species were found to be similar among different genetic systems, further supporting this point (Hammer, 1995; Shen et al., 2000). Microsatellites for the Y-chromosome confirmed these results. Although one study noted that the observed levels of microsatellite diversity could be due to natural selection or a recent increase in population size from a small ancestral population (Pritchard, Seielstad, Perez-Lezaun, & Feldman, 1999), others see no

evidence for a recent selective sweep in the Y-chromosome microsatellite data (Goldstein et al., 1996; Perez-Lezaun et al., 1997).

Only one study has examined the differences in nonsynonymous and synonymous mutations among Y-chromosomal haplogroups (Rozen et al., 2009). Sixteen single-copy X-degenerate genes and five single-copy pseudogenes were analyzed in 105 Y-chromosomes. These Y-chromosomes belonged to 47 different haplogroups. Rozen et al. (2009) found 126 SNPs, only 12 of which were nonsynonymous substitutions. One was found on the branch defining haplogroup F\*, which is present in the majority of non-African Y-chromosomes, but the substitution was believed to be of little functional significance since it was also found in 11 of 12 other mammal and bird Y-chromosomes (Rozen et al., 2009). They also found that the number of nonsynonymous mutations were significantly lower than synonymous mutations. Additionally, mutations occurring in intronic regions and pseudoautosomal regions were significantly more frequent than the occurrence of nonsynonymous mutations (Rozen et al., 2009). They correctly concluded that the pattern of Y-chromosome variation is characterized by the persistence of neutral mutations and the removal of potentially deleterious mutations through purifying selection. However, they erroneously concluded that the action of purifying selection makes this molecule unsuitable for population studies (see discussion above concerning purifying selection and the neutral theory).

The clearest evidence for positive selection should come from the correlation of NRY haplotypes (or haplogroups) with particular disease phenotypes. Because the functionality of the Y-chromosome is reduced due to gene loss, the majority of deleterious phenotypes appear as reductions in fertility. Deletions that can be identified

through karyotypic methods have long been a primary focus of the medical community to assess and diagnose infertility in men (Foresta, Moro, & Ferlin, 2001; Vogt, 2005).

Deletions of sections in the Y-chromosome can impart a phenotype exhibiting a reduction in sperm output in some instances or complete spermatogenic failure in others (Vogt, 2005), although many cases of male infertility cannot be associated with these deletions (McElreavey & Quintana-Murci, 2003). The three classic micro-deletions were termed AZoospermia Factors (AZF), each of which is characterized by a different segmental deletion (AZFa, AZFb, and AZFc). It is now known that multiple genes are located in these AZF regions, but it is not entirely clear which of the genes are relevant to an infertility phenotype (Quintana-Murci, Krausz, & McElreavey, 2001; Vogt, 2005).

One set of deletions called “gr/gr” are known to be rather varied in their structural forms, indicating that several different types of gr/gr deletions actually exist (Vogt, 2005). One of these deletions (a 1.6-Mb deletion) was found to have low penetrance and was transmitted consistently to the next generation (Repping et al., 2003). The bearers of this deletion were thought to have a higher risk of spermatogenic failure (Repping et al., 2003), although others cautioned against such an interpretation, noting the statistical assessment of the small sample sizes (Vogt, 2005). Regardless, the deletion occurs randomly in at least 14 different Y-haplogroups (Repping et al., 2003). Therefore, it does not appear that its effects would skew the demographic inferences concerning any one particular haplogroup. Swift selection against this deletion might be expected, but the reproductive fitness of some men with the deletion did not appear reduced. This is because these men apparently possess a duplication of the deleted DNA segment elsewhere on their Y-chromosome (Tyler-Smith & McVean, 2003). Thus, the



compensatory nature of copy number variation in Y-chromosomal genes must be considered in determining whether any deletion or SNP has deleterious effects on the fitness of the individual.

Deletions need not produce deleterious effects, however, as evidenced by the 1.8-Mb deletion of the Y-chromosome AZFc segment found in haplogroup N (Repping et al., 2004). While the AZFc is associated with a phenotype of reduced fertility or spermatogenic failure, the deletion was widely distributed throughout northern Eurasian populations. In this case, genetic drift certainly could have played a more significant role than selection in spreading this polymorphism (Repping et al., 2004). These findings further indicate that the actual relationships between Y-deletions and level of spermatogenic failure are not yet fully understood. Yet again, it seems the effects of some of these deletions may depend on the availability of compensatory gene products (Tyler-Smith & McVean, 2003).

If phenotypes could be correlated with particular Y-chromosome haplotypes (or haplogroups), then it might be possible to assess the effects that selection may have had on those lineages and, concomitantly, those portions of the Y-chromosome phylogeny. Some studies have attempted to associate genotypes and phenotypes of affected patients (Jobling & Tyler-Smith, 2000; Jobling et al., 1998; Krausz et al., 2001; Quintana-Murci, Krausz, & McElreavey, 2001). One example, involving a translocation of Y genes to the X chromosome, was found to be associated with a particular set of mutations (Jobling et al., 1998).<sup>6</sup> While one genotype was associated with the translocation more frequently

---

<sup>6</sup> I refrain from calling these “haplogroups” because the study only used three Y-chromosome markers to assess membership of two groups. These two groups are general categories and do not provide sufficient resolution to be placed into a YCC-designated haplogroup.

than the other, it was not clear whether this mutation represented a single Y-chromosome haplogroup, which is important as each haplogroup has its own geographic distributions and histories. Furthermore, the selective advantage of the non-susceptible genotypes were considered so small that it was likely outweighed by stochastic processes (i.e., genetic drift) (Jobling & Tyler-Smith, 2000).

Other studies have found associations between haplogroups and phenotypes, but no clear genetic cause for them (Krausz et al., 2001). Yet, many of these studies that attempt to correlate Y-chromosome genotypes and patient phenotypes have not taken other confounding variables into account such as geography, socioeconomic class or other social groupings, which can greater affect Y-chromosome haplogroup distributions in largely patriarchal societies (Jobling & Tyler-Smith, 2000).

### **2.3 Chapter Conclusions**

In this chapter, I have examined the reliability and usefulness of mtDNA and Y-chromosome lineages for answering questions of molecular anthropology. Both genetic systems are uniparentally inherited and do not recombine, making highly resolved haplotype construction possible. The mtDNA genome provides more mutations in a smaller package than the Y-chromosome, but both are excellent for providing a framework in which to examine population histories of both modern and ancient peoples.

The effects of natural selection on these systems have caused some to question whether they can be useful for population studies. In both cases, it is clear that purifying selection has acted to reduce the number of deleterious mutations in their respective phylogenies. This observation is in complete agreement with expectations of the neutral

theory. This pattern can also occur by fixation of deleterious mutations as the result of bottleneck events in non-recombining loci. This pattern has also been consistently found in human genetic studies (Harpending & Rogers, 2000). Therefore, these two molecules do act as systems of neutral markers.

Evidence of positive selection on these haploid systems is scarce, if present at all. When dealing with mtDNA and Y-chromosome data, it should be kept in mind that these molecules provide fundamental functions necessary to survive and reproduce. Thus, in some cases, we would expect positive selection to act on beneficial variants if they were present. Therefore, in a given circumstance, a nonsynonymous mutation might be beneficial and selected for, but only when the selective advantage prevails over random genetic drift in larger populations. Nevertheless, we must also remember that the current patterns of nucleotide variation in mtDNA and the Y-chromosome are consistent with neutral or nearly neutral markers.

The smaller effective population sizes associated with both genetic systems in comparison with autosomal loci appears to outweigh the selective advantage of would-be beneficial mutations, as current patterns of genetic variation show no sign of wholesale positive selection. Especially concerning selection at the haplogroup-level, there does not appear to be any selective advantage or disadvantage on most NRY and mtDNA haplogroups (although caution should be exercised with mtDNA haplogroups L2 and J). Given the lower effective population sizes of these molecules, the instances of positive selection will probably be few (and possibly difficult to detect), although they will likely be some of the most interesting cases because of their implications in understanding speciation events, geographic distributions or, more practically, disease phenotypes.

Understanding the effects that a selected mutation has on the survival and reproductive success of an organism will be critical for identifying these instances of positive selection.

## **Chapter 3: Materials and Methods**

The methodology employed in any investigation is fundamental to the success of the project in meeting its objectives. Clarity in the methods used and the theory behind each test is therefore vital for a proper assessment of the study. This chapter is composed of two parts. The first presents background information on the sample sets used in the analysis, the characterization of molecular genetic markers, and the statistical analyses applied to the genetic data. The statistics are further divided into three main sections. The first discusses the within-population diversity, the second emphasizes between-population variation, and the third presents the statistical assessment of haplogroups (phylogenetics).

### **3.1 Sample Collection**

The DNAs used in this dissertation research were collected during several field seasons conducted in the Altai Republic, Russia. Sample collection included interviews where all participants were asked questions regarding their family histories, including information on their language use as well as that of their parents and grandparents. Genealogies were obtained for each participant to verify their family histories and to exclude relatives from further the study. Self-identified ethnic membership was recorded and verified with information from their interviews.

The first group of samples with which I worked was collected in 2003 from eight villages in the Turochak district of the Altai Republic. This district is located in the northern section of the republic, and is characterized by milder climatic conditions than the more mountainous regions in the south. A total of 262 samples were brought to the

Laboratory of Molecular Anthropology at the University of Pennsylvania. These samples included 162 women and 100 men, and represented mostly indigenous Altaian ethnic groups (Chelkan, Kumandin, Tubalar, and Altai-kizhi); although a small number of other indigenous ethnic groups (Khakass, Kermiac, Bashkir) and Russians were also collected. Three of the eight villages are located in the northwestern portion of the Altai on or close to the Biya River, which joins with the Katun River to form the Ob (Figure 3.1). These

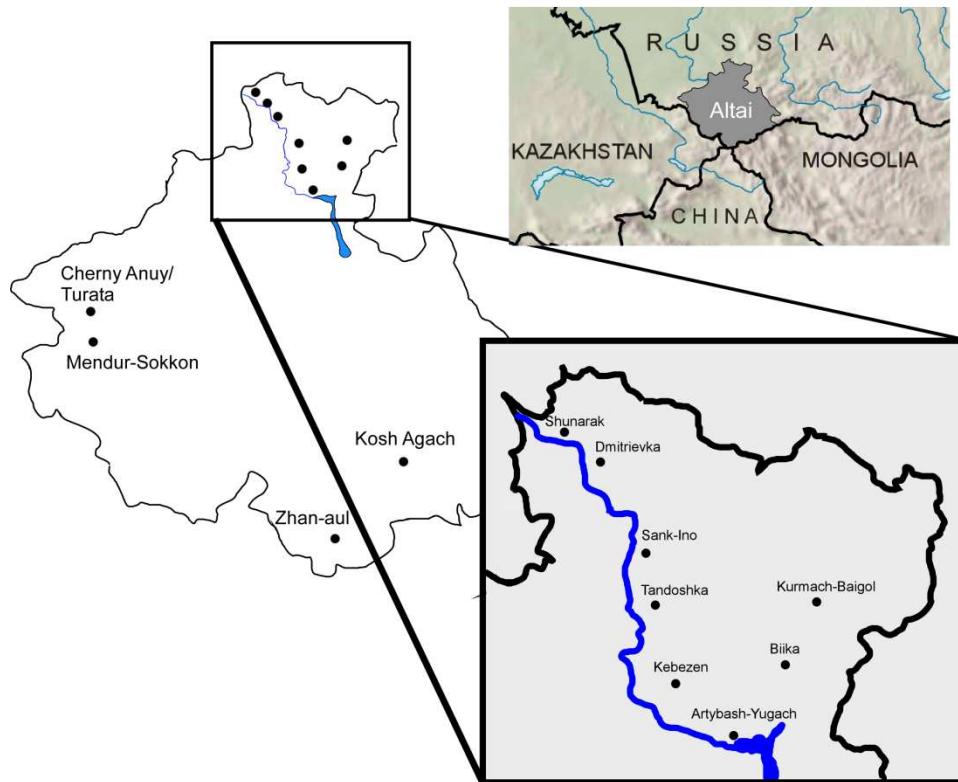


Figure 3.1 Locations of population sample collection in the Altai Republic

villages are Dmitrievka (13 samples), Sank-Ino (20) and Shunarak (16). People from these three villages who provided samples mostly self-identified themselves as Kumandin. Moving upriver (south) on the Biya, the next village is Tandoshka followed

by Kebezen. More samples were collected from these two villages than those from northwestern Altai (Tandoshka – 44; Kebezen – 35). These villages are mostly inhabited by Chelkan and Tubalar families. Samples were collected from two additional villages on the northern portion of Lake Teletskoye at Artybash and Yugach (25 samples total). Mostly Tubular people live at this location. The remaining two villages are located to the north of Lake Teletskoye and to the east of the Biya River. These are Biika and Kurmach-Baigol. The largest sample in the northern Altai came from Biika (79), while moderate numbers were obtained in Kurmach-Baigol (30). Chelkan were numerous in both locations, some Tubular, also.

The second group of samples was present in the Laboratory of Molecular Anthropology. These DNAs were extracted in Russia and brought to the lab for analysis in 2003. They include comparative populations from the southern Altai Republic. The majority of participants belonged to the Altai-kizhi ethnic group, although Altaian Kazakhs were also well represented. Both the Altai-kizhi and Altaian Kazakh samples were collected by our Russian collaborator, Dr. Ludmila Osipova, between 1991 and 2002 from four locations in southern Altai. Mendur-Sokkon and Cherny Anuy/Turata are located in the southwestern Altai Republic, while Kosh Agach and Zhan-aul are located in the southeastern Altai Republic. Altai-kizhi resided in Mendur-Sokkon, Kosh Agach and Cherny Anuy/Turata. The majority of Altaian Kazakhs comes from Zhan-aul, but they live in Cherny Anuy/Turata and Kosh Agach. Our work on mtDNA variation in Altaian Kazakhs was published elsewhere (Gokcumen et al., 2008). In total, there were 268 Altai-kizhi and 237 Altaian Kazakhs available for the comparative analysis.

The third group of samples came from published studies. Appendix 1 includes a list of the populations and references used in the comparative mtDNA and NRY analyses.

All aspects of this study, including sample collection, were vetted by the University of Pennsylvania Institutional Review Board #8 (Protocol #806740) and by the Institute of Cytology and Genetics in Novosibirsk, Russia. I also completed human subjects research training through the CITI courses in the Protection of Human Research Subjects.

### **3.2 DNA Extraction**

I was responsible for the DNA extraction of most of the samples from the northern Altai expedition. Approximately 10 mL of blood were collected from each participant. These were stored at 4 °C until the DNA was extracted. The DNA extraction procedure involved a standard organic phase protocol modified from earlier studies (Schurr et al., 1990; Torroni et al., 1992). First, whole blood samples were fractionated by gravity or through low spin centrifugation, resulting in the separation of the red blood cells from the upper plasma layer, with an interface of buffy coats. A total of 100 µl of the buffy coat interface of each participant's sample was used. This volume was added to an extraction cocktail of pH buffer (5 X STE), detergent (10% SDS), 20mg/ml Proteinase K and water. It was incubated at 55° C for 12-15 hrs to lyse the white cells and release the DNA. Cellular debris was then salted out through the addition of chilled potassium acetate and precipitated through centrifugation. The supernatant was then purified with phenol-chloroform, and the organic phase containing the DNA was precipitated in 95% ethanol. After centrifugation, the DNA pellet was then washed in 70% ethanol and



resuspended in 1X TE (10mM Tris-HCl; 1mM EDTA, pH 8.0). DNA concentrations were quantified using a Nanodrop spectrophotometer, although some samples were quantified using real-time PCR (see below).

### **3.3 MtDNA Characterization**

The mtDNA from each sample was characterized by identifying known diagnostic polymorphisms established from previous studies (Chen et al., 2000; Kivisild et al., 2002; Kong, Yao, Sun et al., 2003; Macaulay et al., 1999; Palanichamy et al., 2004; Richards et al., 1996; Richards et al., 2000; Richards et al., 1998; Schurr et al., 1990; Schurr et al., 1999; Tanaka et al., 2004; Torroni et al., 1996; Torroni, Lott et al., 1994; Torroni, Schurr et al., 1993; Torroni et al., 1992; Torroni, Sukernik et al., 1993; van Oven & Kayser, 2009). The nomenclature used throughout the dissertation follows the mtDNA tree Build 8 (21 Mar 2010) at PhyloTree.org (van Oven & Kayser, 2009).

SNPs were first characterized for each sample using restriction fragment length polymorphisms (RFLPs). RFLPs rely on the presence or absence of short stretches of specific nucleotide sequences recognized by restriction endonucleases. The endonucleases cut the DNA when the restriction site specific to that endonuclease is present. When SNPs occur in these recognition sites, either they will destroy the restriction sites, or it may create new ones in sequences that are one base pair different from a recognition sequence (semi-site). Thus, the presence or absence of a site is indicative of a particular SNP. Table 3.1 provides a list of all the mtDNA PCR-RFLP tests used for this dissertation.

The use of RFLPs allowed us to test for polymorphisms throughout the entire mtDNA genome. Starting first with the polymorphisms at the base of the mtDNA phylogeny, a hierarchical pattern was used to delineate the branch membership for each sample. The combined pattern of two restriction sites (10394 DdeI and 10397 AluI) allowed all the samples to be placed into one of three categories (+/+, +/-, or -/-). Once

Table 3.1 PCR-RFLP and deletion tests for mtDNA SNPs

Hg	SNP/Indel	RFLP Test	Primers (5'-3')	Size	Tm (C°)	Reference
A	663	+HaeIII 663	534-553 / 725-706	191	51	[1]
B	9-bp deletion	N/A	8188-8207 / 8366-8345	178/169	49	Modified from [1]
C	13263	-HincII 13259	13001-13020 / 13403-13384	402	45	Modified from [1]
C	13263	+AluI 13262	13001-13020 / 13403-13384	402	45	Modified from [1]
D	5178A	-AluI 5176	5151-5170 / 5481-5464	330	55	[1]
D5	10397	+BsrI 10396	10279-10296 / 10569-10550	290	51	This study
E	7598	-HhaI 7598	7367-7384 / 7628-7610	261	51	Modified from [2]
F1	12406	-HpaI 12406	12385-12405 / 12576-12595	191	55	Modified from [2, 3]
F1	12406	-HincII 12406	12385-12405 / 12576-12595	191	55	Modified from [2, 3]
F2	7828	+HhaI 7828	6890-6909 / 7131-7115	241	47	This study
G	4883	+HaeII 4830	4651-4670/ 4952-4934	301	46	Modified from [2]
H	7028	-AluI 7025	6890-6909 / 7131-7115	241	47	[4]
H2	4769	+AluI 4769	4651-4670 / 4952-4934	303	49	This study
H8	13101	+HpaII 13101	13001-13020 / 13403-13384	402	45	This study
HV	14766	-MseII 14766	14407-14424 / 14810-14791	403	51	Modified from [5]
V	4580	-NlaI 4577	4500-4519 / 4678-4659	178	51	Modified from [4]
I, X	1719	-DdeI 1715	1615-1643 / 1899-1879	284	54	Modified from [4]
J	13708	-BstOI 13704	13537-13556 / 13851-13832	314	51	Modified from [4]
K	9055	-HaeII 9052	8925-8953 / 9100-9081	175	53	Modified from [4]
M/N	10398	+DdeI 10394	10279-10296 / 10569-10550	290	51	Modified from [1]
M	10400	+AluI 10398	10279-10296 / 10569-10550	290	51	Modified from [1]
N9	5417	+Tsp509I 5417	5151-5170 / 5481-5464	330	55	This study
R	12705	+MboII 12705	12599-12618 / 12785-12766	186	51	This study
R2	14305	-AluI 14304	13940-13959/ 14385-14366	445	57	This study
R6	12285	-AluI 12285	12104-12124 / 12338-12309	234	59	This study
T	15607	+AluI 15606	15409-15428 / 15728-15709	319	51	Modified from [4]
U	12308	+HinfI 12308	12104-12124 / 12338-12309*	234	59	[4]
U1	13104	+MboI 13104	13001-13020 / 13403-13384	402	45	This study
U4	4646	+RsaI 4646	4500-4519 / 4678-4659	178	53	This study
U5	13617	-MboII 13617	13537-13556 / 13851-13832	314	51	This study
W	8994	-HaeIII 8994	8925-8953 / 9100-9081	175	53	Modified from [4]
X	14470	+AccI 14465	14407-14424 / 14810-14791	403	51	Modified from [5]
X2e	15310	+BsrDI 15310	15161-15180 / 15676-15658	515	53	This study
Y	7933	+MboI 7933	7871-7890 / 8020-8001	149	51	This study

Table 3.1 References: [1] Torroni et al. 1993; [2] Torroni et al. 1994; [3] Ballinger et al. 1992; [4] Torroni et al. 1996; [5] Macaulay et al. 1999. \*12338-12309 is a mismatch primer.

the base macrohaplogroup was identified, the samples were then further characterized using appropriate PCR-RFLP tests. In this manner, each sample could be unambiguously categorized into a predefined mtDNA lineage.

In addition to determining the haplogroup membership of each sample, I also used direct DNA resequencing to obtain a complete DNA sequence of the hypervariable regions (HVS) 1 and 2 of the control region. As their names suggest, these sections of mtDNA have higher mutation rates than the coding region (Anderson et al., 1981; Greenberg, Newbold, & Sugino, 1983; Stoneking, 2000), thereby providing a highly polymorphic sequence that is useful in differentiating samples within haplogroups (defined by SNPs). The mtDNA haplotypes provided by this sequence can then be used to construct the terminal branches of the mtDNA phylogeny, and present an unbiased basis for nucleotide comparisons between individuals.

The primer sequences and protocol information for this part of the analysis are provided in Tables 3.2 and 3.3. Briefly, each sample was amplified to include both HVS1 and 2 regions. Excess primers were destroyed using an Exonuclease I – Shrimp Alkaline Phosphatase mixture. Samples were then sequenced using the BigDye Terminator v3.1 kits (Applied Biosystems), which utilizes the Sanger method of dideoxynucleotide (ddNTP) sequencing but with the addition of fluorescent tags to each ddNTP.

After cycle sequencing, extraneous, unincorporated ddNTPs were removed from the samples with either Centri-Sep 8-Well Strips (Princeton Separations), Centri-Sep 96-Well Plates (Princeton Separations) or BigDye XTerminator Purification kits (Applied Biosystems). If samples were purified with the Centri-Sep products, then they were

rehydrated in 10 ml of Hi-Di® Formamide and sequenced using capillary electrophoresis. If the samples were purified with the BigDye XTerminator Purification kits, then they were placed directly in the capillary electrophoresis machine (ABI Genetic Analyzers). Most of the sequencing involved Centri-Sep purification, and purified sequences were sent to the Department of Genetics Core DNA Sequencing Facility at the University of Pennsylvania for capillary electrophoresis. The most recent follow up analysis used the BigDye XTerminator Purification Kits, and sequencing on the 3130xl Genetic Analyzer in the Laboratory of Molecular Anthropology. All sequences were aligned and edited using Sequencher v4.9 (GeneCodes).

Table 3.2 Control region amplification primers

<b>Primer Name</b>	<b>Primers (5'-3')</b>	<b>Range</b>	<b>Size</b>	<b>Tm (C°)</b>
15838FOR	15838-15857	15838 – 725	1455	53
725REV	725-706			

Table 3.3 Control region sequencing primers

<b>Primer Name</b>	<b>Primers (5'-3')</b>
15977FOR	15977-15996
16552REV	16522-16531
1FOR	1-19
429REV	429-409

Whole mtDNA genomes of a subset of samples were sequenced in their entirety. These samples were selected because of a need to further clarify their position in the mtDNA phylogeny (sub-haplogroup level) or to determine the amount of variation within a branch of a particular mtDNA haplogroup. The sequencing protocols followed those mentioned above for HVS1 and 2, but the primers and conditions used were from Palanichamy et al. (2004).

### **3.4 NRY Characterization**

Of the 767 samples from the Altai Republic, 416 were collected from men. All of the NRY data generated from these samples are novel. Sex identification was available for individuals from all locations but Zhan-aul and a subset of those from Cherny Anuy and Kosh Agach. For these 218 samples, I devised a series of tests that utilized two male-specific loci (M89 and M9) and one mtDNA locus (12705) to determine the presence of the Y chromosome for each sample. It can be difficult to characterize the Y-chromosome in older or degraded samples because it is the second smallest nuclear chromosome and has only one copy per male. The mtDNA test was used to ensure that a negative result from the Y-chromosome tests was not due to the complete lack of DNA, as some of these samples had not been collected recently and, therefore, could have degraded over time. These tests involved three of the TaqMan® assays as described below.

The concentrations and quality of Y chromosome DNA were evaluated from a subset of samples using the Quantifiler® Y Human Male DNA Quantification Kit (Applied Biosystems). This test uses real-time PCR technology to determine the amount of Y-chromosome DNA in a sample, and to assess each sample for the presence of PCR inhibitors. This was accomplished using an internal PCR control (artificial DNA sequence) and two TaqMan® minor groove binder (MGB) probes labeled with two tags that fluoresce at different wavelengths. When the DNA or PCR control was amplified, the respective tags fluoresced, tracking DNA amplification. The use of serial dilutions of a standard with a known DNA concentration allowed for the rapid assessment of relative

amounts of the unknown DNA samples compared to the known standard, thus providing absolute DNA concentrations for the samples in question.

Much like the mtDNA, the NRY is characterized by using both slower evolving (SNPs) and faster evolving mutations (STRs) to provide accurate placement of each sample into the Y-chromosome phylogeny.<sup>7</sup> Haplogroup designations follow the accepted nomenclature providing by the Y-Chromosome Consortium (Karafet et al., 2008; Y Chromosome Consortium, 2002). Samples were screened for SNPs in a hierarchical manner. A subset of samples was tested more broadly to confirm haplogroup assignments.

SNPs were characterized by one of three methods: TaqMan® assay, direct sequencing, or PCR-RFLP analysis. TaqMan® SNP Genotyping Assays operate in the same manner as described above for the Quantifiler® DNA Quantification kit. A small fragment of DNA containing the locus of interest is amplified. The presence of the derived allele is signaled by the detector of the corresponding probe. This method is extremely sensitive, making it ideal for the characterization of Y chromosome markers. All TaqMan® assays were run on an ABI Prism® 7900HT Real-Time PCR System. Each reaction was carried out using 5 µl total volume and the standard protocol supplied by Applied Biosystems, except instead of running the reaction for 40 cycles, I ran each reaction for 50 cycles. SDS v2.4 was used to run all reactions, and all results were called by hand. AutoCaller v1.0 was used to verify the results of several independent runs for a subset of markers. The list of NRY TaqMan® assays used in the dissertation can be found in Table 3.4. In total, 70 TaqMan® assays were used in this research.

---

<sup>7</sup> In this case, the Y-SNPs are analogous to the PCR-RFLP results for the mtDNA, while the Y-STRs are comparable to the mtDNA control region sequences.

Direct sequencing was used to assess the presence of some NRY markers because either an appropriate set of primers and/or probes could not be designed as a TaqMan® assay, or there was a low probability of a particular marker being present in the sample set. Direct sequencing provides more information than the TaqMan® assays, and primers cost less than TaqMan assays for smaller batches of samples. Therefore, for markers that are not found in many populations (potentially private mutations) or are unlikely to be

Table 3.4 NRY TaqMan assays

	<b>Haplogroup</b>	<b>Marker</b>
1	C	M130
2	C3	M217
3	C3	PK2
4	C3a	M93
5	C3c	M86
6	D	M174
7	D1	M15
8	D2	M55
9	D3a	P47
10	E	M96
11	E1b1b1	M35
12	E1b1b1	M81
13	E1b1b1c	M123
14	F	M89
15	G	M201
16	G1	M285
17	G2a	P15
18	H	M69
19	I	M170
20	I1	M253
21	I2	P215
22	I2a	P37.2
23	I2b	M223
24	J	M304
25	J1	M267
26	J2	M172
27	J2a	M410
28	J2b	M12
29	J2b	M102
30	K	M9
31	T	M70
32	K1	M147
33	K2	P60
34	K4	P261
35	S	M230

	<b>Haplogroup</b>	<b>Marker</b>
36	L	M20
37	M	P256
38	M1	M186
39	NO	M214
40	N1	LLY22g
41	N1a	M128
42	N1b1	P63
43	N1c	M46 (Tat)
44	N1c1	M178
45	O	M175
46	O1a	M119
47	O2	P31
48	O3	M122
49	O3a3c	M134
50	O3a3c1	M117
51	O3a3c1a	M162
52	P	M45
53	Q	M242
54	Q1a1	M120
55	Q1a2	M25
56	Q1a3a	M3
57	Q1a6	M323
58	Q1a3	M346
59	R	M207
60	R1	M173
61	R1a1a	M56
62	R1a1b	M157
63	R1a1c	M204
64	R1b1	P25
65	R1b1a	M18
66	R1b1b	P297
67	R1b1c	M335
68	R1b1b1	M73
69	R1b1b2	M269
70	R2	M124

present (for example, they have a restricted distribution), this approach was best. Like the direct sequencing carried out in the mtDNA studies listed above, all of the NRY characterization used BigDye XTerminator® Purification Kits and a 3130xl Genetic Analyzer. All sequences were aligned and edited using Sequencher v4.9 (GeneCodes). All of the primer information for NRY markers used in the dissertation is listed in Table 3.5.

Table 3.5 NRY marker sequencing reactions

Hg	Marker	SNP	Primers (5'-3')	Size	T <sub>m</sub> (C°)	Reference
C3b	P39	G95A	AAAATTTGCCAAGCATGGTG	184	59	Designed from [6]
			CAAGGGGCAGATTGATTGAT			
N1b	P43	G268A	TTTGGAGGGACATTATTCTC	519	53	[5]
			GAAGCAATACTCTGAAAAGT			
N1c	TAT-C M46	A69G	GACTCTGAGTGTAGACTTGTGA	112	60	[3]
			GAAGGTGCCGTA AAAAGTGTGAA			
C3d	M407	A149G	GTTATACCCTGCTCTAAAGTGCTTC	418	57	[4]
			GTAGAGATGGGGCTTCACCGTGTTAC			
Q1a3	M346	C33G	CCCCGTTTTTTCCTCTCTGCC	419	58	[4]
			AATCTGCCTTCCAACAAACC			
R1a1	M17	68G del	CTGGTCATAACACTGAAAATC	333/334	51	[2]
			TGAACCTACAAATGTGAAACT			
Q1a	MEH2	G880T	ATTCATAATATTTGATTTCAGAACAG	938	52	[1]
			TACCATGAAAATTCATAATCCACA			
C3e	P53.1	T112C	GATGTCACCTTCCGTCTA	476	51	[1]
			ACATGGTCATCTGTAGCTCC			
C3f	P62	443C insert	AACCCCTGCCACAAATACAT	623/624	58	[1]
			TCTGGAACCCCTGGAGAGATC			
N1c	P105	G580A	TTATTCCACCCAGCACTGTTA	1090	56	[1]
			AGGCACAAATGGTAAGGTCTT			
R1b	M343	C402A	TTAACCTCCTCCAGCTCTGCA	424	59	[1]
			ACCCCCACATATCTCCAGG			
Q1a4	P48	428T insert	TGAAGGACAGTAAGTACACA	637/638	47	[1]
			TAAGTCCATTGATCTACAGA			
Q1a5	P89	G258T	ACATTACAGGACCTTGAT	857	47	[1]
			TGCCTAACAACTACTCC			
R1a1d	P98	C504T	TGGAGGGTAAGTGAGTAG	954	47	[1]
			TTTTAATGGAACACCGTAG			
O3a3c2	P101	G360A	TGAGGTTGATGTTTACTAAGATC	506	51	[1]
			CCTGCTAAATCAGTTTCCACAC			
R1a1e	PK5	C186T	TTCCAAACACATGCTTCTGC	393	57	[1]
			TAAAAAGGAGGAGGGACTGC			

Table 3.5 References: [1] Karafet et al. 2008; [2] Underhill et al. 1997; [3] Zerjal et al. 1997; [4] Sengupta et al. 2006; [5] Karafet et al. 2002; [6] Zegura et al. 2004



Only a single RFLP test was used to characterize an NRY marker. This test involved marker M175, which is a five base pair deletion that is diagnostic for haplogroup O (Karafet et al., 2008). While a five base pair deletion can be detected by electrophoresis in a standard 3% Agarose gel, the *Eco*I digestion of the amplicon containing this deletion helps to differentiate the derived allele from the ancestral (Cox, 2006).

In addition to the SNP genotyping of each Y chromosome, I also generated haplotypes of 17 Y-STRs using the AmpF/STR® YFiler® PCR Amplification Kit (Applied Biosystems). Standard protocols were followed for amplification and capillary electrophoresis runs on the 3130xl Genetic Analyzer, and the results were edited by visual inspection using GeneMapper® ID v3.2 software. The DYS389I locus was subtracted from the DYS389II locus and labeled “DYS389b,” as DYS389I was amplified a second time with DYS389II (Hurles et al., 1999).

Overall, a Y-chromosome lineage is defined as the combination of derived SNP information and Y-STR profile for each sample (Appendix 2).

### **3.5 Statistical Analysis**

#### **3.5.1 Population Genetic Statistics – Within Population Estimates**

The genetic variation in populations can be summarized using several statistics. These statistics are generally descriptive, providing quantitative estimates of basic features for each population. The statistical values provide basic characteristics that allow each population to be assessed relative to other populations of known diversity or

size. These estimates are therefore useful in supplying a general understanding of the amount of variation in population, but not necessarily the type of variation.

The first of these summary statistics is the gene diversity or average heterozygosity. This statistic is defined as the probability that any two alleles (either haplogroup or haplotype) randomly chosen from a population are different (Nei, 1978, 1987). The estimation of this statistic was implemented using Arlequin v3.11 (Excoffier, Laval, & Schneider, 2005). I calculated the gene diversities for both mtDNA and NRY data using haplogroup designations and haplotypes. Those calculated from haplogroup frequencies are referred to as “haplogroup diversities,” while those from the haplotype data are “haplotype diversities.”

Haplogroup diversity is estimated from haplogroup frequencies, with each haplogroup essentially representing a different allele. For this reason, the definition of a haplogroup is important when describing the amount of variation in a population. A haplogroup is essentially a monophyletic clade in a phylogeny that shares a number of unique polymorphisms relative to other clades of equal depth (Richards et al., 1998).

The haplogroups for human mtDNA studies were defined by RFLPs (SNPs and indels). It has become clear, however, that not all of these SNPs are at the same level (or depth) in the phylogeny. For example, the SNP defining haplogroup U (12038) is quite old and encompasses a number of other haplogroups (U1-U7, U8/K and U9) (van Oven & Kayser, 2009). Throughout these analyses, all haplogroup designations are related to branches that are approximately the same depth in the phylogeny. The NRY haplogroups are similarly resolved for this analysis.

Haplotype diversity is calculated like haplogroup diversity. However, this gene diversity estimate is the probability that any two haplotypes (HVS1 DNA sequences for mtDNA or Y-STRs for NRY) randomly chosen from a single population are different. Evaluation at the nucleotide level is made between comparisons of the same stretches of sequence or repeats. For this reason, each unique haplotype serves as a unique allele. Moreover, haplotype diversities are not biased in the same way because it is not necessary to group samples. In fact, no *a priori* categories are used for these estimates.

Differences at the haplotype-level were examined with additional statistics. For the mtDNA sequences, nucleotide diversity, average pairwise differences, and mismatch distributions were estimated using Arlequin v3.11 (Excoffier et al., 2005). Nucleotide diversity is the probability that any two nucleotides randomly chosen from a population are different (Nei & Li, 1979; Nei & Tajima, 1981). It is can be more informative than gene diversity estimates for relationships between haplotypes since the statistic takes into account the amount of difference between sequences instead of simply whether two alleles are different. Similarly, the average number of pairwise differences takes into account the DNA sequence and is defined as the average number of differences between all pairs of haplotypes in a population (Tajima, 1983).

These estimates have become standards for describing the amount of genetic diversity within a population (Nei, 1987). This is mostly because they are not influenced by sample size in the same manner as counting the number of segregating sites between sequences or counting the number of alleles in a sample – both measures of which can be affected by deleterious mutations (Tajima, 1983). Nevertheless, it should be noted that large stochastic variances can be associated with average pairwise difference estimates.

Average pairwise distributions are calculated from the observed number of differences between pairs of haplotypes, which is called a mismatch distribution (Excoffier et al., 2005). Using a single stepwise expansion model, the shape and raggedness of the mismatch distribution curve can provide insight into a population's demography (Rogers & Harpending, 1992). This approach was expanded to include rate heterogeneity, making it more suitable to mtDNA sequence analysis (Schneider & Excoffier, 1999). Simulation studies indicate that the population size parameters are too conservative when estimated with rate heterogeneity. Therefore, the magnitude of expansion cannot be determined with this method (Schneider & Excoffier, 1999), although the parameter for time of expansion is still unbiased (Excoffier et al., 2005; Slatkin, 1995).

For Y-chromosome microsatellite data, two statistics were employed – the number of different alleles and the sum of squared differences. The number of different alleles is the equivalent of the number of unique haplotypes for the mtDNA data. The other statistic is specific to microsatellites. The sum of squared differences is dependent on the similarities in repeat length at each locus, and can therefore be used to estimate distance between haplotypes (Slatkin, 1995). The statistic “counts the sum of the squared number of repeat differences between two haplotypes” (Excoffier et al. 2005:104). The fewer the number of repeat differences per locus between two haplotypes, the more similar two STR haplotypes are.

The population parameter,  $\theta$ , theta ( $\theta = 2N\mu$ , where  $N$  is the inbreeding effective population size and  $\mu$  is the neutral mutation rate), was estimated using mtDNA HVS1 sequences. Four different calculations were made using Arlequin v3.11 (Excoffier et al.,

2005). The four estimates for mtDNA haplotypes are based on expected homozygosity ( $\theta_{(H)}$ ), number of segregating sites ( $\theta_{(S)}$ ), expected number of alleles ( $\theta_{(k)}$ ), and the average number of pairwise differences ( $\theta_{(\pi)}$ ) (Ewens, 1972; Excoffier et al., 2005; Tajima, 1983; Watterson, 1975; Zouros, 1979). This analysis provides information on the relative strength of mutation versus genetic drift in a population (Templeton, 2006). For the Y-STR haplotypes, only one  $\theta$  estimate was calculated, and it was based on expected heterozygosity using a pure stepwise mutation model (Excoffier et al., 2005; Ohta & Kimura, 1973). The other three estimates were not used for Y-STRs because either they are based on sequence data ( $\theta_{(S)}$  and  $\theta_{(\pi)}$ ) or were not appropriate due to violations of fundamental assumptions associated with the statistics (infinite-allele equilibrium,  $\theta_{(k)}$ ).

Neutrality indices were also calculated with Arlequin v3.11 using some of the aforementioned  $\theta$  estimates. In Tajima's test of neutrality, the D statistic is calculated using the difference between  $\theta$  estimates derived from the number of segregating sites and the average number of pairwise differences (Excoffier et al., 2005; Tajima, 1989a, 1989b, 1996). Similarly, Fu's  $F_S$  test of neutrality considers the probability of observing an equal or fewer set of alleles as the population in question for a random neutral population compared to the  $\theta$  estimate obtained from the average number of pairwise differences (Excoffier et al., 2005; Fu, 1997). We consider any estimate with a P-value at 0.02 as significant for Fu's  $F_S$  test of neutrality, following the recommendations in Excoffier et al. (2005). Both of these neutrality tests essentially assess the number of rare alleles in a population. The excess of rare alleles can be due to selection, but it is also characteristic of population expansion. Conversely, the presence of several alleles at high frequencies indicates balancing or diversifying selection or even population substructure.

### 3.5.2 Population Genetic Statistics – Between Population Estimates

Several statistics were used to assess between population variation. First, haplogroup frequency data were used to assess the overall genetic similarities between populations. Populations were represented as allele frequencies, where each allele corresponds to a particular haplogroup. Principal components analysis (PCA) was then employed in SPSS v.11 to determine which variables might underlie the genetic diversity among populations.

Comparisons at the haplogroup level are useful for obtaining a relatively general view of genetic differences among populations. Because the haplogroup information is retained in the haplotype data, haplotypic data can be more useful for determining the fine-scale relationships between people or populations. Therefore, two primary statistics were used to evaluate the genetic distance between populations at the haplotypic level.

The first estimates population  $F_{ST}$  values. For the HVS1 sequences, these estimates were calculated using the Tamura-Nei substitution model. This model was designed specifically for human mtDNA, taking into account the excess of transitions over transversions, differences between transition rates of purines and pyrimidines, unequal nucleotide frequencies and variation of substitutions among different sites (Excoffier et al., 2005; Tamura & Nei, 1993).

The genetic distances for the Y-STR haplotypes were estimated as  $R_{ST}$  and  $\delta\mu^2$  values. Because we are not using the sequence variation as means of addressing closely related haplotypes as with the mtDNA, but rather the repeat variance of STRs, the sum of squared differences was used to estimate the genetic distances between haplotypes among populations (Slatkin, 1995). The  $R_{ST}$  values were calculated with Arlequin v3.11, and the

$\delta\mu^2$  values were calculated with Arlequin v3.5.1.2 (Excoffier et al., 2005; Excoffier & Lischer, 2010).

In all cases where pairwise population genetic distances were obtained, the resulting matrices were visualized using multi-dimensional scaling (MDS) plots. These were generated using SPSS v11 (SPSS Inc., 2001). The significance of differences among populations was assessed using p-values generated from 100,000 permutations of haplotypes between populations (Excoffier et al., 2005). The null hypothesis is that no difference exists between two populations. Generally, a p-value of 0.05 is considered appropriate.

With multiple comparisons, however, there is a question as to whether differences between populations are significant merely by chance (Type I errors). Therefore, Bonferroni corrections were also implemented. For these corrections, the significance level is 0.05 multiplied by the number of comparisons being made, although this step may also increase Type II errors.

Throughout the discussions of pairwise genetic distances, I use both non-corrected (0.05) and corrected (Bonferroni) levels of significance to assess the population relationships. Values at the extremes are noted. Any p-values less than the Bonferroni corrected p-value are significantly different, while any p-value greater than 0.5 is not.

Analysis of Molecular Variance (AMOVA) was used to investigate genetic structure of populations. The relative proportions of genetic variance within populations, among groups and among populations within groups were determined using a hierarchical analysis of variance (Excoffier et al., 2005; Excoffier, Smouse, & Quattro, 1992). For this analysis, groups must be created by the user. In each case, I define the

groups being used in the AMOVA, the genetic component estimates attained, and the p-values generated for those components. Generally, most data sets were investigated using categories defined by geography, linguistics and ethnicity/cultural groupings. However, variations on these themes are also used as noted in the following chapters.

Spatial analysis of molecular variance (SAMOVA) was also used to investigate geographic structuring of the populations without *a priori* categories. The SAMOVA program utilizes geographic information in the form of points of longitude and latitude, creating a matrix of all distances between all populations (Dupanloup, Schneider, & Excoffier, 2002). Correlations between genetic distance matrices (generated from either  $F_{ST}$  or  $R_{ST}$  values) and the geographic matrix indicate where genetic barriers (or barriers to gene flow) may exist. The correct number of groups in the analyzed data set can be obtained by running multiple tests, with different group numbers being used each time. The test with the maximum  $F_{CT}$  value is considered the best fit to the data (Dupanloup et al., 2002). In this manner, the relationship between genetic data and geography can be investigated.

### 3.5.3 Phylogenetics – mtDNA Data

The use of phylogenetic methods can provide useful information about the populations in question. However, they do have limitations. Phylogenetics can be seen as an offshoot or a natural extension of the early taxonomic, then cladistic, methodologies that used phenotypic characteristics to categorize species. These phenograms initially represented relationships between species, although it could be argued that they



represented similarities in adaptation since morphological traits were often used to create the categories.

With phylogenetics, questions of genetic relationship are often resolved (although not always), depending on the type of DNA mutations and loci used. It should be noted here that the phylogenetic trees are gene trees, which can be different from a population or species tree. Therefore, it should not be assumed that gene trees show the exact relationships between all populations or are necessarily representative for an entire species.

At the population level, reduced median (RM) -median joining (MJ) networks were generated (Bandelt, Forster, & Rohl, 1999). These networks allow for direct comparison between haplotypes, and provide information about haplotype-sharing among populations, including differences between ethnic or linguistic groups and their geographic distributions. Median networks are also useful in that they show all likely relationships between all haplotypes in a particular data set. As a result, reticulations may occur in these networks, as they show the possible ways that the current haplotypes could be related to one another. In this manner, the topology is not forced to choose one version over another, unlike any phylogenetic tree using maximum parsimony, maximum likelihood or Bayesian analysis. Of course, there is only one gene history that is represented by only one tree. Nevertheless, using networks that allow multiple cycles shows all possible trees while also pointing out the haplotypes in the network whose relationships are not entirely known.

For each network, a combined RM-MJ network was created. First, a reduced median network was generated. The resulting rmf file was then used as the input file for

a median-joining network. This combined approach removes many of the superfluous cycles (reticulations). A weighting scheme was used for networks generated from mtDNA haplotypes, where each nucleotide position was given a relative weight corresponding to the amount of recurrent mutation witnessed at that position among a worldwide sample of sequences. The conventions used for choosing fast-evolving sites versus slow-evolving sites follows established recommendations (Bandelt et al., 2002).

Others have noted that some sites in the HVS1 are more mutable than other sites (Hasegawa et al., 1993; Meyer et al., 1999; Sigurgardottir, Helgason, Gulcher, Stefansson, & Donnelly, 2000; Soares et al., 2009; Stoneking, 2000; Wakeley, 1993; Yang, 1996). These sites are found in a large variety of haplotypes with very different backgrounds. For the network analysis, a two-class system was ideal for generating the initial networks. Once the networks were generated, reticulations were examined. If a reticulation was caused by a fast-evolving site, then that particular site was further down-weighted. Transversions, which occur less frequently than transitions, were given higher weights because they are less likely to recur. Also for the mtDNA networks, SNP information from the more slowly evolving coding region can be incorporated to provide a framework on which the HVS1 haplotypes can be placed. These SNPs, like transversions, occur less frequently, and therefore, were weighted much higher than the HVS1 transitions. All of the networks were generated with Network v4.5.1.6 (Bandelt et al., 1999) and visualized using Network Publisher v1.2.0.0 (Fluxus Technology Ltd).

Analysis of mtDNA haplogroup phylogenies requires several tests to check for clocklike evolution and selection. Likelihood ratio tests (LRTs) were used to ensure that a haplogroup was evolving in a clocklike manner. First, sets of three maximum

likelihood (ML) trees were created in PAML (Yang, 1997) using (1) complete mtDNA genome sequences, (2) coding region sequences only, and (3) control region sequences only. The ML trees were generated both with and without outgroup sequences. Either an L0a or an L3 sample were used as an outgroup for all analyses. In all cases, the trees generated without outgroups produced better trees, according to the LRTs (data not shown). A nonhuman primate outgroup was not used because there is uncertainty in mtDNA sequence evolution between primate species and the potential problems of saturation (Howell et al., 2004). Therefore, all remaining analyses did not use outgroups in ML tree construction.

TREE-PUZZLE was used to assess ML tree branch lengths, which are compared between two models of sequence evolution, one involving a molecular clock, the other without (Schmidt, Strimmer, Vingron, & von Haeseler, 2002). For each mtDNA haplogroup, the overall haplogroup tree was tested, as was each major haplogroup branch. Three separate tests were conducted, the first using the entire mtDNA genome, a second with only the coding region sequence, and a third with only the control region sequence. This approach was used to investigate the roles of the coding and control regions on the clocklike evolution of the entire molecule.

Site variability was considered in all tests given the nature of the mtDNA genome, because there is evidence that the molecule does not have a uniform substitution rate (Yang, 1996). Both the HKY and Tamura-Nei DNA substitution models were tested, but did not produce significantly different results (LRT data not shown). Therefore, all analyses used the simpler HKY substitution model. To reduce the impact of type I errors,

I followed the Howell et al. (2004) method in which a Bonferroni correction is used and a p-value of 0.0050 was retained for all tests.

For these analyses, all sequences obtained from GenBank were identified using Sequencher v4.9 (GeneCodes) and Mega v4 (Tamura, Dudley, Nei, & Kumar, 2007). Sequences were aligned in Mega v4 by hand. Pairwise differences were calculated in Mega v4 to identify duplicate sequences. Once duplicate sequences were removed, the file was exported as an aligned fasta file. The aligned fasta file was then converted to phylip format with SeqVerter v2.0.4.3 (GeneStudio).

I used mtDNA GeneSyn v1 (Pereira et al., 2009) to examine the types of mutations in each haplogroup. Despite the problems associated with interpreting the ratios of nonsynonymous and synonymous mutations, I used this measure to assess each of the gene products of the mtDNA. These estimates were calculated in Mega v4 using only the DNA sequences for that particular region of the mtDNA. Thus, 13 separate alignment files were created and analyzed in this manner for each haplogroup. The nonsynonymous / synonymous rates were calculated between all sequence pairs and averaged. Fisher's Exact Tests were used to determine if the values were significantly different (Gerber et al., 2001). Z-codon scores (dN and dS tests) were also calculated in Mega v4 to assess the effects of positive selection, purifying selection and neutral evolution. These were calculated using the Nei-Gojobori method based on the number of pairwise differences (Tamura et al., 2007).

### 3.5.4 MtDNA Coalescence Estimates

Recently, several studies have independently re-examined the mtDNA to determine the most accurate mutation rate considering issues of time-dependency and differences between phylogeny and pedigree-based rates (Henn et al., 2009; Soares et al., 2009). Henn et al. (2009) calculated average pairwise differences ( $\pi$ ) and average distance to a founding haplotype ( $\rho$ ) and used within-human divergences points to calibrate their effective mutation rate, which was generally consistent with pedigree rate estimates up until divergence points about 5,000 years ago. The effective rate was also mostly consistent with the phylogenetic-based rate after 20,000 years, where the time between 5,000 and 20,000 years experienced a decay in mutation rate (Henn et al., 2009).

Considering the potential problems associated with mtDNA (purifying selection, time dependency of the mutation rate and saturation of mutations with site-specific variability), Soares et al. (2009) also produced a method to determine the coalescence estimates of mtDNA lineages that takes these issues into account, as mentioned in Chapter 2. After assessing the amount of recurrent mutation, mutational saturation and relative mutation rates among different portions of the mtDNA genome, the neutral mutation rate was calculated from the synonymous mutation rate (Soares et al., 2009). An interspecies calibration point was used to calculate the mutation rate. A 7-myra split between *Pan* and *Homo* was used, assuming that the population divergence (~6-6.5 myra) actually occurred after the molecular divergence. The mutation rate for the entire mtDNA genome was  $1.665 \times 10^{-8}$  ( $\pm 1.479 \times 10^{-9}$ ) mutations/nucleotide/year (Soares et al., 2009). The examination of ML tree branch lengths showed that rate variation between these two species did not influence the mutation rate estimate. This mutation rate

provided the expected divergence times for other hominoid species (9.4 mya human-gorilla; 2-2.5 mya chimp-bonobo; 500 kya modern human – Neanderthal) (Soares et al., 2009).

To correct for purifying selection, age estimates were calculated in which the proportion of synonymous mutations are correlated with the total variation in a given phylogenetic branch to determine the amount of deleterious mutations that have not yet been removed from the phylogeny. Several archaeological calibration points were used to test the internal consistency of the mutation rate (Soares et al., 2009). A highly significant correlation ( $R^2 = 0.9452$ ) was found between the time-dependent corrected mutation rate of the complete genome and the linear synonymous mutation rate.

I used the Soares et al. (2009) method for calculating the coalescent dates for each haplogroup and sub-haplogroup. To estimate the rho statistics and accompanied sigma values, I generated networks as described above, but specific for the haplogroup in question. Thus, six median joining-reduced median networks were generated for each haplogroup. The first used all mutations, except 16519 (which is the most hypervariable site in the mtDNA genome) and insertions into the poly-cytosine tract around 16189C (which are due to polymerase slippage) (Bandelt et al., 2002; Soares et al., 2009). The second network used only synonymous mutations from the coding regions. The third network used the entire control region, except 16519 and insertions around 16189C. The fourth network used the DNA sequence range of 16090-16365 in HVS1, as employed by Forster et al. (1996). The fifth used the DNA sequence range of 16051-16400. The final network used a partial HVS2 sequence, ranging from 73-263.

### 3.5.5 NRY Phylogenetics

Lineages for the Y-chromosome were created by combining microsatellite and SNP data. The issue of selection cannot be addressed with the data presented here because continuous stretches of DNA sequence were not used in this analysis. Furthermore, the SNPs characterized in the study were chosen because they are known to be polymorphic sites, and thus, ascertainment bias is present in the SNP sample set. However, there is reason to believe that the Y-chromosome acts as a neutral marker (see previous chapter). Therefore, in the phylogenetic analysis of NRY haplogroups, tests of selection or clocklike rates are not undertaken. The objective for analyzing the Y-chromosome lineages is specifically to determine the chronological and geographic origins of haplogroups. Thus, the methods used are strictly for inferring coalescence estimates and diversity among populations.

For any coalescent-based analysis, an accurate mutation rate is critical for obtaining times to the most recent common ancestor. There is evidence that Y-STRs evolve at different rates mostly due to the number of base pairs and the composition of the repeat sequences (Kayser et al., 2001). All STRs used in this dissertation have the same relative mutation rates, and thus, a uniform rate is appropriate for the analysis of rho statistics (Shi et al., 2010). In Bayesian analysis, there is an option to use different mutation rates. In my analysis, the results revealed no significant differences between estimates when the same mutation rate is used compared to different mutation rates for each locus (data not shown).

As mentioned in the previous chapter, there are currently two mutation rates available that were estimated from different sources. The first mutation rate comes from

pedigree studies and provides the faster of the two rates ( $2.3 \times 10^{-3}$  mutations/locus/generation). The second is called the evolutionary rate and was calculated using prehistoric and historical events (confirmed with archaeological evidence) ( $6.8 \times 10^{-4}$  mutations/locus/25 years). Given the antiquity of Y-chromosome haplogroups inferred through coalescence dating of Y-SNP accumulations, the evolutionary rate is the more appropriate mutation rate for analyzing entire haplogroups. This rate will always provide an estimate about three times greater than the pedigree rate. Current Y-chromosome studies have tended to use the evolutionary rate more often, but there is still not a total consensus on this issue.

For NRY haplotype networks, a variable number of loci were used to construct Y-STR haplotypes. While I produced a 17 Y-STR loci profile for my samples, many of the published data used for comparisons were at a lower resolution, with as few as five Y-STRs being used. Depending on the data set and haplogroup, the number of loci used varies. The goal throughout this analysis was to retain the maximum coverage with as many loci as possible. These networks were weighted approximately using the inverse of the variance in repeat number calculated for each microsatellite locus. Y-SNPs can be used to provide the basic framework for the Y-STR haplotypes. All of the networks were generated with Network v4.5.1.6 (Bandelt et al., 1999) and Network Publisher v1.2.0.0 (Fluxus Technology Ltd).

Several Y-STR loci were not used in this analysis. DYS385 actually represents two separate loci amplified together. The convention for listing these loci is to place the smaller repeat number first, followed by the larger – for example, 12-13 or 12-14. However, a problem arises when using the YFiler amplification kit (or any other



commercially available multiple), since there is no way to determine which repeat allele is DYS385a and which is DYS385b (Gusmao et al., 2006). Construction of accurate networks requires that each repeat is assigned to the appropriate locus – a fact often overlooked in network and phylogenetic tree reconstructions. In addition, duplicated alleles cannot be appropriately included in networks – even as a unique event polymorphism – for the same reason. Only one locus has duplications in this dataset (DYS19), and it only occurs in one haplogroup (C3c).

With geographic information, Y-STR networks help to evaluate the mutations from a phylogeographic point of view. Rho statistics can be calculated from the networks to provide coalescence estimates for each branch of the phylogeny and for the entire haplogroup. Despite potential problems with coalescence dating, rho values can provide another means of assessing the amount of diversity in a population or haplogroup. Similar to the rho statistics, the variance of STR repeats can be used to assess the amount of diversity in a population or haplogroup (Kayser et al., 2001). Intrapopulation variance was therefore calculated for each haplogroup and haplotype clusters if the clusters were readily identifiable.

In addition, coalescence estimates were also calculated using a Bayesian analysis. Batwing was used to generate phylogenetic trees with the input of prior distributions for various parameters (Wilson, Balding, & Weale, 2003). The program explores distribution space to determine which parameter values best fit the given model. Three models are available in Batwing. The first is constant population size, the second exponential population growth, and the third is a model where the population initially has a constant population size, but begins growing exponentially at time  $\beta$ . The third model

is commonly used in human population analyses of the Y chromosome and is used exclusively in this dissertation.

For all tests carried out, each run consisted of a 5,000-cycle burn-in followed by 25,000 cycles. Each cycle consists of 10 iterations, such that 250,000 iterations total were searched. All prior distributions followed previous studies (Xue et al., 2006). It is critical in Bayesian analysis to ensure that posterior distributions of each parameter converge. Convergence is accomplished when the space for any given parameter is fully explored, with the results being nearly the same with each independent run. I ran the analysis 10x the initial number of cycles to ensure that the results are consistent with the shorter runs per previous recommendations (Xue et al., 2008).

## **Chapter 4: Mitochondrial DNA and Population Histories**

To explore the maternal genetic ancestry of Altaian populations, I characterized coding region SNPs and control region sequences from 262 inhabitants of the northern Altai region. Of these individuals, 216 belonged to one of three northern Altaian ethnic groups (Chelkan – 91; Kumandin – 52; Tubalar – 71), while two people had mixed northern Altaian ancestry. The remaining 46 samples included Altai-kizhi and Russians, as well as several people self-identified as belonging to various ethnic groups (Kermiac, Teleut, and Telenghit). In addition, I analyzed 505 individuals characterized previously by others in the Laboratory of Molecular Anthropology at Penn. These samples were from locations in the southern Altai and represent Altaian Kazakhs (Gokcumen et al. 2008) and southern Altaians (mostly Altai-kizhi).

Assignment of samples into discrete mtDNA haplogroups through PCR-RFLP and DNA sequencing revealed the presence of 25 mtDNA haplogroups in the Altaian data set (Table 4.1). MtDNA nomenclature followed the current standards outlined in PhyloTree.org (mtDNA Tree Build 9). For the following analyses, samples were categorized by region and by ethnic group as self-reported at the time of sample collection. Individuals verified through genealogies to be related through at least the past two to three generations were removed from the analysis as mentioned above. The northern Altaian group comprised Chelkan, Kumandin and Tubalar, while the southern Altaians consisted of only Altai-kizhi. I used published data for the remaining Altaian ethnic groups that were not represented in our sample collection (northern Altaian: Shor; southern Altaian: Teleut and Telenghit) (Derenko et al., 2003; Derenko, Malyarchuk,

Grzybowski et al., 2007; Starikovskaya et al., 2005). MtDNA data from the Altaian Kazakhs were the focus of a separate study (Gokcumen et al., 2008). Because these Kazakhs were relatively recent immigrants to the Altai, their mtDNA results were

Table 4.1 MtDNA haplogroup frequencies for Altaian populations

Hg	Chelkan	Kumandin	Tubalar1	Tubalar2	Shor	Altai-kizhi1	Altai-kizhi2	Telenghit	Teleut
#	91	52	71	72	28	276	48	55	33
<b>C</b>	15.1	41.5	35.6	20.8	17.9	31.4	25.0	14.6	24.2
<b>Z</b>			2.7		3.6	4.3	4.2		3.0
<b>M8</b>						3.6	4.2		
<b>D4</b>	13.9	15.1	24.7	15.3	25.0	13.0	6.3	18.2	24.2
<b>D5</b>	8.6	3.8	4.1	5.6	3.6	0.7			3.0
<b>G</b>	3.2					4.0	4.2	3.6	
<b>M7</b>								1.8	
<b>M9</b>						1.4			
<b>M10</b>	1.1				3.6	0.4	2.1		
<b>M11</b>							2.1	1.8	3.0
<b>M*</b>								1.8	
<b>A</b>		1.9		11.1	3.6	2.9	4.7	7.3	
<b>I</b>					3.6	1.4	2.1	1.8	
<b>N1a</b>								1.8	
<b>N1b</b>						0.4			
<b>W</b>	1.1								
<b>X</b>		3.8		1.4		2.2	2.1		3.0
<b>N9a</b>	19.4	1.9	2.7	6.9				1.8	
<b>B</b>	3.2	3.8	2.7	4.2	3.6	1.4	6.3	14.6	6.1
<b>F1</b>	10.8	3.8		1.4	14.3	8.3	4.2	1.8	3.0
<b>F2</b>	15.1		2.7		3.6	2.5	2.1		
<b>H</b>	1.1		2.7	1.4	3.6	2.5	8.3	9.1	9.1
<b>H2</b>						3.3	2.1		
<b>H8</b>		5.7	2.7	4.2	3.6	1.4			
<b>HV</b>								1.8	
<b>V</b>									6.1
<b>J</b>					3.6	4.0	6.3	1.8	
<b>T</b>		1.9				0.4		3.6	6.1
<b>U2</b>				2.8		0.7		1.8	3.0
<b>U3</b>							2.1		
<b>U4</b>	4.3	3.8	15.1	18.1	3.6	0.7	2.1	1.8	3.0
<b>U5</b>	2.2	9.4	4.1	5.6		3.3	2.1	1.8	
<b>U8</b>								1.8	
<b>K</b>					3.6	3.3	6.3		3.0
<b>R9</b>	1.1	3.8		1.4		2.2		5.5	
<b>R11</b>							2.1		

considered separately from the indigenous Altaians and will be revisited in the chapter on Altaian Kazakhs NRY variation.

#### **4.1 Northern Versus Southern Altaian Genetic Variation – Haplogroup Level**

The majority of mtDNA haplogroups identified in Altaians was of East Eurasian origin (~78%). Northern Altaians comprise Chelkan (89.5%), Tubalar (77.8%), and Kumandin (70.4%), and for the southern Altaians, Altai-kizhi (75.2%). All three non-African macrohaplogroups had East Eurasian representatives in the Altai (**M** – C, D, G, M7, M8, M9, M10, M11, Z; **N** – A, N9a; **R** – B, F). This finding was expected given previously published reports of genetic variation in southern Siberian populations (Derenko et al., 2003; Starikovskaya et al., 2005). Haplogroups C, D, U, and F are typically found throughout this region of Siberia, but the distribution of these types varies among the populations. Among our indigenous Altaian samples, the most frequent haplogroup was haplogroup C (28.9 %), followed by D (15.9 %), F1 (5.3 %) and F2 (5.3 %). All other haplogroups were present at frequencies below five percent.

Despite the prevalence of eastern Eurasian haplogroups among Altaians, the profiles differed between the northern and southern Altai regions. Haplogroups C, D, F, and N9a constituted the majority of East Eurasian mtDNAs in northern Altaian populations (27.4%, 23.7%, 13.0%, and 8.8 %, respectively). Haplogroups C, D and M8 were the most numerous in southern Altaians (30.1 %, 13.8 % and 5.4 %, respectively). While haplogroups C and D were common in both regions, southern Altaians had larger numbers of haplogroup C, whereas northern Altaians had a slightly higher frequency of haplogroup D.

Even greater differences between northern and southern Altaians became obvious when considering the relative contributions of West Eurasian haplogroups to people in each region. Southern Altaians had a larger number and greater variety of West Eurasian haplogroups. These included haplogroups from both major West Eurasian macrohaplogroups (**N** – I, W, X and **R** – HV, H, V, U, K, J, T). Only two haplogroups were found at frequencies higher than 5%. For the northern Altaians, U4 was the most prevalent West Eurasian haplogroup (7.9%), while haplogroup H was most prevalent in southern Altaians (5.9%). Haplogroups U4, U5, H (including H8), T, and X were found in both regions, but only U4, U5 and H8 were more frequent in northern Altaians. Several West Eurasian haplogroups were found exclusively among the southern Altaians, including K, J, V, U2e and H2, but W was only found in northern Altaians.

#### **4.2 Northern Altaian Genetic Variation**

One of the primary objectives of this dissertation was characterizing the genetic structure among the northern Altaian ethnic groups. Comparisons of the haplogroup profiles making up the maternal genetic diversity of each of these ethnic groups provided evidence for different population histories (Table 4.1). Haplogroups C and D played a significant role in origins of the northern Altaian ethnic groups, as they have for many of the populations in Siberia (Derenko et al., 2003; Schurr et al., 1999; Starikovskaya et al., 2005; Volodko et al., 2008). Haplogroup C was most frequent in Kumandin and Tubalar, while it was a close second for Chelkan. Haplogroup D was found in all three of these groups at high frequencies, with a subset of this haplogroup in each ethnic group belonging to D5.

Certainly, at the level of the mtDNA haplogroups, differences among northern Altaians were also present. The Chelkan were distinct from the other northern Altaians in having a prevalence of haplogroups F1, F2 and N9a mtDNAs. Even in comparison with populations outside of the Altai, the Chelkan were unique in their mtDNA haplogroup profile, as they possessed high frequencies of N9a and F2a and the lowest frequency of haplogroup C (~15%) reported anywhere in (non-European) southern or northern Siberia (Derenko et al., 2003; Schurr et al., 1999; Starikovskaya et al., 2005; Volodko et al., 2008). This ethnic group in particular may have experienced considerable genetic drift.

Besides the dominant C and D haplogroups, the Tubalar were characterized by an abundance of U4 – the most of any of the Altaian groups. The Kumandin, on the other hand, had the greatest frequency of U5, and were the only northern Altaian group to have A, H8, T and X, although these latter four haplogroups were present at low frequencies. The Chelkan and Tubalar also possessed haplogroups that are not shared among the other northern Altaian ethnic groups. The Chelkan possessed G and W, whereas the Tubalar had the only instance of haplogroup Z in the northern Altai. In general, however, these haplogroups are relatively scarce for this region.

#### **4.3 Haplotype-Sharing among Altaians**

MtDNA haplogroups are useful for understanding the broader patterns and relationships between populations. However, some of these haplogroups are ancient. Simply stating that two populations share the same haplogroup could create the perception that the populations are more similar than they actually are. It is possible that

the haplotypes from the two populations are distinctive and come from different branches that separated when the haplogroup arose. Therefore, the actual relationships between groups can be hidden when relying only on haplogroup information.

Haplotype analysis provides a means of understanding the relationships between populations at a more refined level. The haplotypes discussed here are defined by both PCR-RFLP and DNA sequence data to provide the greatest amount of detail that is feasible with the number of samples involved (comparable to the NRY lineages that combine Y-SNP and Y-STR data) (Table 4.2). Complete mitochondrial genomes would be the ultimate level of detail, but cost considerations made this kind of analysis prohibitively expensive at the time these samples were being characterized. (Targeted analysis of complete genomes was possible for some samples, and these cases are noted below.) Assessment of haplotype sharing by simple counting provides a method for identifying which groups share mtDNAs and therefore share common maternal ancestry. Throughout this next section, haplotypes from the four ethnic groups available in our sample collection will be compared, after which statistical analysis that quantifies these differences will be presented.

The high frequencies of haplogroups C and D in Altaians required that they get special attention. Haplogroup C mtDNAs were more prevalent in southern than northern Altaians, so it was not surprising that there were double the number of C haplotypes in the Altai-kizhi. Sixteen of the twenty-five haplotypes were not found in the northern Altaians. Only two of the 20 haplotypes were found in all four Altaian groups. The first of these haplotypes possessed only the root HVS1 motif for haplogroup C (16223-16298-16327). All of these likely belong to the C4 branch, but they could be from different



clusters. The second was a C5b haplotype (16093-16223-16288-16291-16298-16327). Of the remaining haplotypes, two others were shared between southern and northern Altaians, but five were exclusive to northern Altaians.

Haplotype sharing of mtDNAs belonging to haplogroup D among Altaian groups presented a similar perspective as seen with haplogroup C. Of the twelve haplotypes

Table 4.2 Haplotype-sharing among Altaian populations

Hg	HVS1 (16024-16400)	Chelkan	Kumandin	Tubalar	Altai-kizhi
A4	223-290-319-362				4
A4	192-223-290-319-362		1		1
A4a1	223-249-290-319-362				3
B4b1	086-136-189-217	3	2	2	4
C	223-298-327	1	8	10	5
C	223-242-298-327				1
C	093-223-298-327				1
C	223-298-311-327		1	3	
C	223-298-327-329				1
C	223-298-327-329h(A/G)				1
C1a	223-298-325-327-356				2
C4a1	129-150-223-298-327				1
C4a1b'd	129-223-298-327				2
C4a1c	093-129-223-298-327	3			19
C4a1c	093-129-223-298-311-327				2
C4a1c	093-129-223-298-312-327				1
C4a2	171-223-298-327-344-357				6
C4a2	171-223-278-298-327-344-357				4
C4a2	171-223-278-298-344-357				2
C4a2	171-223-278-298-344h(T/C)-357				1
C4a2	171-223-298-327-344	3		1	
C4b3?	223-291-298-327			5	22
C	223-288-291-298-327			1	
C5	148-223-288-298-327				3
C5a	223-261-288-298				5
C5b	093-223-288-291-298-327	7	11	3	2
C5b	093-223-288-291-298-311-327		1		
C5b	093-223-288-291-298-327-362			1	
C5c	093-223-288-298-327-390				6
D	223-362	3		3	8
D	218-223-362				1

Hg	HVS1 (16024-16400)	Chelkan	Kumandin	Tubalar	Altai-kizhi
D	140-223-274-311-362				8
D	082-147A-223-362	3		2	1
D1	223-325-362				1
D4	223-319-362				5
D4b1b2b1	093-172-173-215-223-319-362		3	6	3
D4j	223-291-362				3
D4m	042-214-223-362	4	5	4	3
D4o	223-290-362				3
D4o1	176-183-223-274-290-319-342-362	3		2	
D4o1	129-176-183-223-274-290-319-342-362			1	
D5a2a	092-126-164-189-223-266-362			3	2
D5c2	129-188.1C-193.1C-362-390	8	2		
F1a1	162-172-304	3	1		
F1b	189-232A-249-304-311	7	1		4
F1b	189-232A-249-294-304-311				2
F1b	189-270-304				1
F1d'e	189-304				16
F2	092A-291-304	14		2	7
G1a1	223-325-362				8
G2a	223-227-278-362	2			
G2a	223-227 h(G/A)-278-362	1			
G2a	223-278-304-362				1
G2a	223-278-287-304-362				2
H	CRS			1	
H	311				2
H	092-245-362				1
H	092h(A/C)-245-362				1
H	169-184				1
H	261-311				1
H1b	189-356			1	
H1b	080-189-356	1			1
H2a1	354				8
H2a1	311-354				1
H8	288-362		3	2	4
I	129-223-391				4
J	069-126-241				3
J	069-126-241-301				1
J	069-126-189-260				1
J	069-126-145-261-290				4
J1b	069-126-145-172-222-261				1
J1b	069-126-145-172-222-261-304G				1
K1a1a	093-224-311				7
K1a1a	093-189-224-311				2

Hg	HVS1 (16024-16400)	Chelkan	Kumandin	Tubalar	Altai-kizhi
M8a	184-223-298-319				6
M8a	184 h(T/C)-223-298-319				1
M8a	134-184-223-298-319				2
M8a1	172-184-189-223-298-319				1
M9a1a	223-234-316-362				3
M9a1a	223-234-316-344-362				1
M10	129-186-223-311-362	1			
M10a1	093-193-223-311-357-381				1
N1b	145-176G-223-258-291-390				1
N9a	223-248-257A-261-311	16	1	2	
R9b1	093-207-304-362-399				6
R9b1	145-192-243-304-309-390	1	2		
T1a	126-163-186-189-294		1		
T2b	111-126-294-296-304-311-327				1
U2e	051-129C-189-214-258-362				2
U4	356	1		6	
U4	129-356			1	
U4b1b	311-356	3	2	4	2
U5a1	192-241-256-270-287-304-325-399	2	5	2	3
U5a1	192-256-270-319-320-399			1	
U5b2	192-249-311				5
U5b2	093-192-249-311				1
W6	192-223-292-325	1			
X2e	189-223-278		2		6
Z	185-223-260-298				8
Z1a	129-185-223-224-260-298			2	4
	<b>Total</b>	91	52	71	276

belonging to haplogroup D4, only one was shared among all Altaian groups (16042-16214-16223-16362). It belonged to D4m2. Six of the haplotypes were exclusive to the Altai-kizhi, and only two were confined to the northern Altaians. However, sharing of haplogroup D mtDNAs was common among the northern Altaian ethnic groups. There was only one seemingly significant pattern among the northern groups: the Kumandin tended to lack some D4 haplotypes, while Chelkan and Tubalar had a greater variety of them. They also exhibited differences in D5 haplotypes. There were only two distinct

haplotypes, each belonging to a different branch of the D5 phylogeny. Prevalent among the Chelkan was a D5c haplotype that was also found in two Kumandin. Conversely, the Altai-kizhi and Tubalars shared a D5a haplotype.

In addition to the C and D mtDNAs, all Altaian groups shared B, U4 and U5 haplotypes. There was no variation in haplogroup Bs – only a single haplotype was observed (16086-16136-16189-16217). All Altaians also shared a specific U4 haplotype (16311-16356). Two other U4 haplotypes were found in northern Altaians, but most of those mtDNAs belonged to Tubalar individuals. Similarly, all Altaian groups shared one U5a haplotype (16192-16241-16256-16270-16287-16304-16325-16399). An additional haplotype was identified in a single Tubalar. Of these four groups, the Altai-kizhi was the only one to have U5b haplotypes.

The Kumandin and Tubalar mitochondrial gene pools did not have many unique haplotypes. Kumandin had very few haplotypes that would distinguish them from other Altaians. For example, the A and X Kumandin haplotypes matched those found in Altai-kizhi. The R9b Kumandin haplotype was shared with Chelkan. In addition, the Kumandin H8 haplotype was shared with Tubalar and Altai-kizhi.

The Chelkan were the most distinctive of the Altaian groups. Most of the Chelkan mtDNAs fell into F1, F2 and N9a, although there was little genetic diversity within each of these haplogroups. F2 and N9a consisted of only one haplotype each (16092A-16291-16304 for F2; 16223-16248-16257A-16261-16311 for N9a), and F1 comprised only two haplotypes (16162-16173-16304 for F1a1; 16189-16232A-16249-16304-16311 for F1b).

The majority of the above-mentioned haplotypes belonged to the Chelkan. One Kumandin shared each of the two F1 haplotypes with the Chelkan. Sixteen of the nineteen N9a haplotypes belonged to the Chelkan, while the remainder were present in the Tubalar and Kumandin. Fourteen of the twenty-two F2a haplotypes belonged to Chelkan, with six occurring in Altai-kizhi and two in Tubalar. One of the two Tubalar individuals who had this haplotype lived in the same village where eight Chelkan were found to have F2a mtDNAs. Among the four haplotypes just discussed, this was the only instance of haplotype-sharing between ethnic groups in the same village.

To summarize, the characteristics that set the northern Altaian groups from each other were not due to ethnic group-specific haplotypes. Instead, it was the relative frequencies of these haplotypes in each group. Thus, Kumandin and Tubalar had high frequencies of C and D haplotypes, but Kumandin had higher frequencies of U5 and H8, and Tubalar had U4. Differentiation between southern and northern Altaians was also due to a combination of differences in relative frequencies of haplotypes and differences among a few group-specific haplotypes. For instance, northern Altaians have N9 and D5c while southern Altaians do not, and southern Altaians have M8, M9, K, J, and U2 which northern Altaians lack. This pattern of genetic diversity is consistent with a common maternal origin of northern Altaians, with differences among ethnic groups caused by isolation and/or endogamy in small-sized populations. The differences between southern and northern Altaians suggest either a different maternal origin of southern Altaians and/or greater gene flow from other populations as compared to the northern Altaians.

#### **4.4 Genetic Structure in Altaian Populations**

An important issue for understanding the population histories of people residing in the Altai is the manner in which genetic diversity in Altaian villages is structured. This issue is complicated for northern Altaians because many of their villages are inhabited by multiple ethnic groups (including non-indigenous Altaians). If each village was composed of only a single ethnic group, then a population could be identified easily by village membership.

Ultimately, the question posed here is “what is the population?” Or more appropriately, should the unit of analysis be a village, an ethnic group, or both? The simplest approach is to use the village as the unit of analysis because membership is clear-cut. Little ambiguity would exist as to how each sample should be classified.

There is, however, one aspect of the village (as a unit of analysis) that does need clarification. Are the northern Altaian villages genetically stratified? Inhabitants of each village come from at least two different ethnic groups, if not more. Therefore, does ethnic group membership create genetic architecture within the village such that, in reality, there are two or more populations per village? No village exists in isolation from all others; marriages do occur between people from different locations.

Why then should “the village” be analyzed? The local population (or deme) surely includes multiple villages. Thus far, the data has been described along the lines of ethnic group membership (Chelkan versus Altai-kizhi, etc.). It can be argued that this approach is appropriate, since most marriages occur between members of the same ethnic group and occur between villages.

Therefore, to determine the best unit of analysis for understanding how northern Altaian populations may be structured, I investigated the genetic diversity of these samples using the different categories just mentioned. First, I explored the structure in the data set using each village as a separate population. Next, I explored the genetic distances as related to geography. Finally, I examined the genetic diversity among ethnic groups partitioned by villages.

Geography is often the best predictor of genetic variation (Betti, Balloux, Amos, Hanihara, & Manica, 2009; Hunley, Healy, & Long, 2009; J. Z. Li et al., 2008; Manica, Prugnolle, & Balloux, 2005; Novembre & Stephens, 2008; Rosser et al., 2000; Serre & Paabo, 2004). Given the correlation between geography and genetic relatedness, this is the logical hypothesis to test first. If geography played the greatest role in structuring the genetic variation in the Altai, then we would expect genetic diversity to be subdivided into northern and southern regions. We would also expect the northern villages to have smaller genetic distances among each other as compared to the southern villages, as the distances between northern villages would be much smaller than the distances between northern and southern villages.

To test this hypothesis, pairwise  $F_{ST}$  values were calculated using each village as a separate population and plotted using multidimensional scaling (MDS). Based on the MDS plot of  $F_{ST}$  values, there was no clear separation into geographic clusters (Figure 4.1). The villages from the northern Altai were scattered across the entire plot. The southern Altai locations also did not group together. Instead, Kosh Agach was an outlier, situated at the bottom right-hand portion of the MDS plot. Cherny Anuy was closest to Tandoshka on the left of the plot. Finally, Mendur-Sokkon was located in the bottom-

center of the plot along with the northern village of Dmitrievka. Thus, there was no clustering of northern and southern villages.

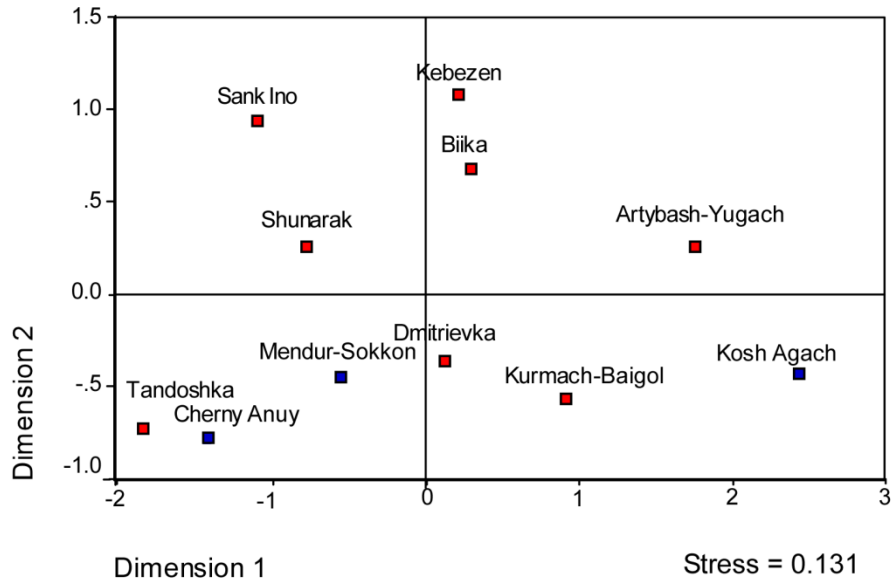


Figure 4.1 MDS plot of  $F_{ST}$  values for Altaian villages. Northern villages are shown in red; Southern Altaian villages are in blue.

Because the pairwise  $F_{ST}$  calculations involved multiple comparisons, a Bonferroni correction was utilized to determine the significance of the p-values. This correction attempts to reduce Type 1 errors, but can be overly conservative, thereby increasing Type 2 errors. Comparisons between populations using both a 0.05 significance level and the significance level with a Bonferroni correction provide a relative means of assessing which groups are different from each other. Using a significance level of 0.004 (as calculated with a Bonferroni correction), Tandosha was significantly different from Artybash-Yugach, Biika and Kebezen. Dmitrievka and Shunarak were similar to all other northern villages (p-values > 0.05). Overall, the extent of differences in the mtDNA diversity among these villages was relatively low.



Analysis of molecular variance (AMOVA) was carried out to determine the amount of variation present between the northern and southern groups. The northern group consisted of Artybash-Yugach, Biika, Kebezen, Kurmach-Baigol, Tandoshka, Dmitrievka, Sank-Ino and Shunarak. The southern group comprised Mendur-Sokkon, Cherny Anuy/Turata and Kosh Agach. No significant differences were found between these two groups (Table 4.3). About 97% of the genetic variation was accounted for by the “within population” category. The remainder was found between populations within the same group. Thus, there was greater variation among the northern villages and among the southern villages than between the two geographical groupings.

Table 4.3 AMOVA of northern versus southern Altaian villages

<b>Groups</b>	<b>Percentage of Variation</b>	<b>P-value</b>
<b>Geography</b>		
<i>Among group</i>	-0.40	0.459
<i>Among population within group</i>	3.95	0
<i>Within population</i>	96.45	0

Table 4.3 The “Northern” category comprises Artybash-Yugach, Biika, Dmitrievka, Kebezen, Kurmach-Baigol, Sank-Ino, Shunarak, and Tandoshka. The “Southern” category includes Mendur-Sokkon, Cherny Anuy and Kosh Agach.

To explore the role of geography further, I used the SAMOVA program, which performs AMOVA without using *a priori* categories. Essentially, the program identifies genetic boundaries among the compared populations. The user fixes the number of groups ( $k$  from 2 – 20), and the program splits the populations into the specified number of groups. The scenario with the smallest  $F_{CT}$  value is determined. If the greatest divergence occurs between the northern and southern villages, then the expectation is that, when  $k$  equals two, membership in one group should include all of the northern

villages, while membership of the other group should consist of the southern villages. These should also have the highest  $F_{CT}$  value.

When the SAMOVA program was run for two groups, the first village removed was Kosh Agach, with all other villages falling into the other group. Nevertheless, the p-value was not significant (p-value = 0.081). This result was not altogether unexpected, as Kosh Agach was the farthest removed from all the other locations sampled. As  $k$  was increased incrementally from three to eight, Dmitrievka (p-value = 0.091), Sank-Ino (p-value = 0.098), Kurmach-Baigol (p-value = 0.065), Kebezen (p-value = 0.105), Biika (p-value = 0.108), and Artybash-Yugach (p-value = 0.004) were removed from the main group. Using eight groups provided the first instance where the “among group” variation was greater than the “between population within group” component and had a significant p-value. When the number of groups was increased to nine, Mendur-Sokkon was pulled from the main group, although the  $F_{CT}$  value was extremely low and the p-value high (p-value = 0.301). With 10 groups, Cherny Anuy was separated from Shunarak and Tandoshka (p-value = 0.234).

At no point did a clear division between northern and southern villages with statistically significant p-values emerge. It was only with eight groups that the “among group” component was statistically significant. This set included one group composed of two northern and two southern villages. Thus, geography alone did not sufficiently explain the genetic structure of the populations in our data set. These results were unexpected at face value because geographic distance is generally a good predictor of genetic diversity among populations. They also point to a complex picture of the genetic diversity in Altaian groups. It is possible that, at some locations in the northern Altai,

populations interacted with people from the southern Altai more than others (recent genetic exchange). Another possibility is that some of the northern Altaian ethnic groups have greater affinities with southern Altaian ethnic groups due to a shared common origin.

Thus, to expand this analysis further, each ethnic group from every village was examined.  $F_{ST}$  values were again calculated and used to create an MDS plot (Figure 4.2). First, it should be noted that the stress value for the two-dimensional MDS plot was relatively high (0.298), meaning that some of the relationships between the points may be skewed. A three-dimensional version was also created, but the stress value only decreased to 0.211. In these plots, the Artybash Kumandins were an outlier. This population consisted of only two individuals who shared the same R9b1 haplotype. Aside from the Artybash Kumandins, the R9b1 haplotype was only found in one Chelkan from Kebezen.

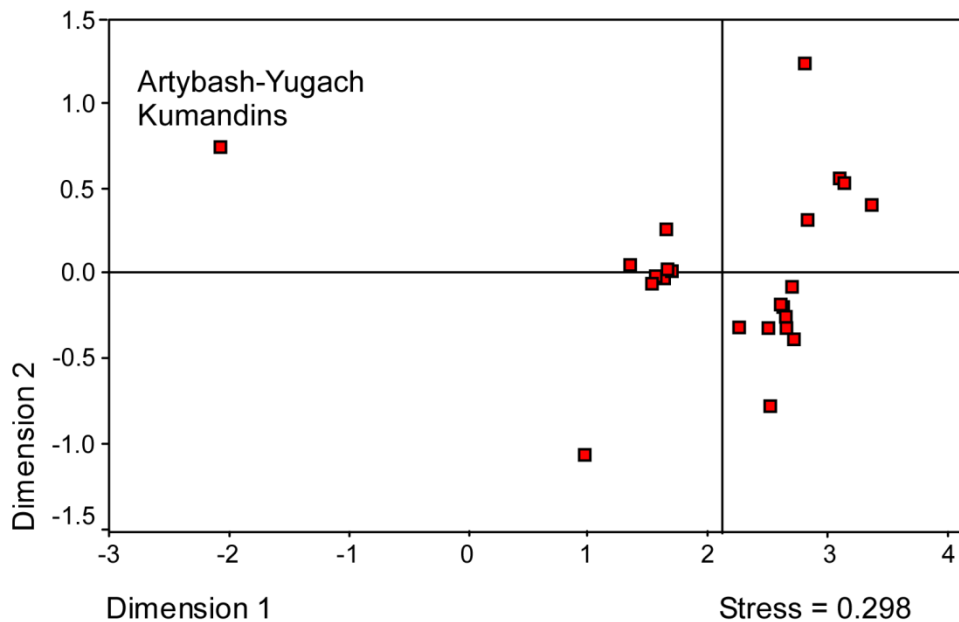


Figure 4.2 MDS plot of  $F_{ST}$  values for each ethnic group by village

There was some ambiguity with the results of this analysis because there were not the same numbers of participants from each village. Also, for several villages, only a few individuals comprised some of the ethnic groups. These small sample sizes can negatively affect the analysis by skewing relationships and amplifying dissimilarities. To reduce the effects that the smaller populations have on the MDS plot, the following groups were removed from the analysis: Chelkan from Artybash-Yugach, Sank-Ino Shunarak, and Tandoshka; Kumandin from Artybash-Yugach, Kurmach-Baigol and Kumandin; and Tubular from Dmitrievka, Kurmach-Baigol and Sank-Ino.

The new MDS plot had a stress value of 0.159 (Figure 4.3). In this figure, the Chelkan were designated by squares, Kumandin by filled circles, Tubalar by open circles and Altai-kizhi by an "X." Each village was represented by different colors. The three northern Altaian ethnic groups showed different patterns. The Chelkan had a wide distribution, stretching over the right half of the plot. The Biika (light green) Chelkan showed similarities with Kosh Agach (black) southern Altaians and Artybash (gold) Tubalar, while the Kurmach-Baigol (red) Chelkan were near the center of the graph. The Artybash Tubalar was the only population separated from the main cluster of points in the center of the MDS plot where the other Tubalar were located. The Kumandin populations had a larger distribution than the Tubalar. Some were located in the central cluster, but with several outlier populations located at the top of the plot. The Mendur-Sokkon (dark purple) Altaians were found in the central cluster as well, very close to several Kumandin and Tubalar populations. In this context, it will be interesting to see if Kumandins and Tubalars are in fact more closely related to the southern Altaians than the

Chelkan, as this plot suggests. Haplogroup analysis has already suggested that the Chelkan are unique, which is also highlighted here.

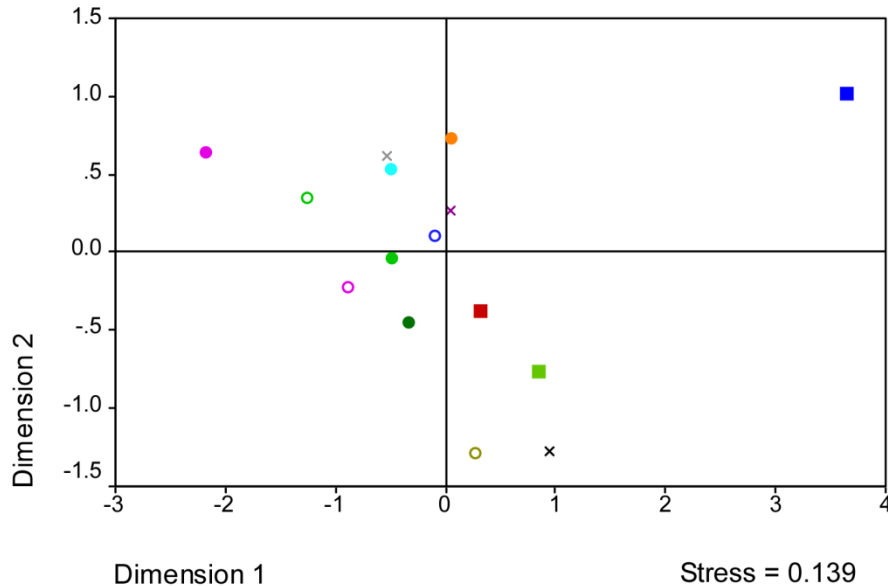


Figure 4.3 MDS Plot of  $F_{ST}$  values - ethnic groups and villages. Ethnic groups are designated by symbols: Chelkan (square), Kumandin (filled circle), Tubalar (open circle), and Altai-kizhi (x). Villages are designated by color: Artybash (gold), Biika (light green), Dmitrievka (dark green), Kebezen (dark blue), Kurmach Baigol (red), Sank-Ino (light blue), Shunarak (orange), Tandoshka (light purple), Mendur-Sokkon (dark purple), Cherny Anuy (Gray), Kosh Agach (black).

An AMOVA was run using all samples defined by village residence to determine if there was greater variance between villages (among group) or among the ethnic groups in a village (among population within group). For example, the three ethnic populations of Chelkan, Tubalar and Kumandin from Artybash-Yugach were defined as three separate populations and then grouped together as “Artybash-Yugach Village.” The results indicated greater variance among different ethnic groups within the same village than among villages (Table 4.4). This analysis was also performed after removing populations with very few members (for example, there are only two Shunarak Chelkan),

and the same results were obtained. This finding confirmed the previous experiments in that geography did not play a significant role in structuring of the mtDNA diversity in northern Altaian villages. Rather, the genetic variation residing in each village seems structured along clan or ethnic lines.

Table 4.4 AMOVA of northern Altaian villages

<b>Groups</b>	<b>Percentage of Variation</b>	<b>P-value</b>
<b>Geography</b>		
<i>Among group</i>	-1.43	0.656
<i>Among population within group</i>	8.93	0.000
<i>Within population</i>	92.50	0.000

Table 4.4 Each village was defined as a group, with each ethnic group within a village defined as a population: Artybash-Yugach (Chelkan, Kumandin, Tubalar), Biika (Chelkan, Kumandin, Tubalar), Dmitrievka (Tubalar, Kumandin), Kebezen (Chelkan, Tubalar), Kurmach Baigol (Chelkan, Kumandin, Tubalar), Sank-Ino (Chelkan, Kumandin, Tubalar), Shunarak (Chelkan, Kumandin) and Tandoshka (Chelkan, Kumandin, Tubalar).

To confirm the results of the previous AMOVA, the next AMOVA grouped the northern Altaian ethnic groups together to determine whether categorization by ethnic group provides a better framework for understanding how the genetic variation is partitioned. The seven Chelkan populations (from Artybash-Yugach, Biika, Kebezen, Kurmach-Baigol, Sank-Ino, Shunarak, and Tandoshka) were grouped into a “Chelkan” group, and populations for the other two ethnic populations were similarly grouped. In total, there were three groups (Chelkan, Kumandin, and Tubalar) comprising 21 populations (seven populations for Chelkan, seven for Kumandin and seven for Tubalar).

Table 4.5 AMOVA of northern Altaian ethnic group

<b>Groups</b>	<b>Percentage of Variation</b>	<b>P-value</b>
<b>Geography</b>		
<i>Among group</i>	2.03	0.027
<i>Among population within group</i>	6.11	0.000
<i>Within population</i>	91.86	0.000

Table 4.5 Each ethnic group represented a category, with each ethnic group from a village as a population: Chelkan (Artybash-Yugach, Biika, Kebezen, Kurmach-Baigol, Sank-Ino, Shunarak, Tandoshka), Kumandin (Artybash-Yugach, Biika, Dmitrievka, Kurmach-Baigol, Sank-Ino, Shunarak, Tandoshka) and Tubalar (Artybash-Yugach, Biika, Dmitrievka, Kebezen, Kurmach-Baigol, Sank-Ino, Tandoshka).

The “among group” variation accounted for 2.0% of the variation and the “among population within group” for 6.1% (Table 4.5). While the “among population within group” category still had a higher percentage than the “among group” variation, there was a significant increase in “among group” variation when compared to results of the previous AMOVA. Furthermore, when the populations with small sample sizes were removed (similar to the preceding MDS plot and AMOVA analyses), the pattern remained the same (Table 4.6). In fact, the “among group” percentage rose to 3.4% and the “among population within group” percentage dropped to 4.1%, thereby indicating that the small sample size of from some of the villages affected the “among population within group” variation. Thus, the AMOVA provided clear evidence that membership in northern Altaian ethnic groups account for a significant amount of variation and were, therefore, more important than geographically-based categories.

Table 4.6 AMOVA of northern Altaian ethnic groups (without small populations)

Groups	Percentage of Variation	P-value
<b>Ethnic Group</b>		
<i>Among group</i>	3.61	0.005
<i>Among population within group</i>	4.45	0.002
<i>Within population</i>	91.95	0.000

Table 4.6 Each ethnic group represented a category, with each ethnic group from a village as a population: Chelkan (Biika, Kebezen, Kurmach-Baigol), Kumandin (Biika, Dmitrievka, Sank-Ino, Shunarak, Tandoshka) and Tubalar (Artybash-Yugach, Biika, Kebezen, Tandoshka).

The higher percentage of variation in the “among population within group” category indicated differences within ethnic groups. To determine which ethnic groups had high levels of variation, AMOVAs were run for each ethnic group independently. In doing this analysis, it was determined that the 4.5% of genetic variation in the “among population within group” came from the Chelkan and Kumandin. The differences between Chelkan populations among villages accounted for about 8% of the genetic variation. This was also the case for Kumandin populations. By contrast, the “among population between group” category for Tubalar populations accounted for 2% of the variation, although this value was not significant (p-value = 0.170).

SAMOVA was run separately for Chelkan and Kumandin populations to determine whether the structure within ethnic groups was due to geography. The SAMOVA results were not statistically significant. Thus, the variation uncovered between populations of the same ethnic group appeared to be caused by some kind of clan structure in the Chelkan and Kumandin.

Once evidence supporting the use of ethnic self-identification rather than geographic location for analysis was established,  $F_{ST}$  values were calculated for our



Altaian populations to determine genetic distances between ethnic groups (Table 4.7). The Chelkan were approximately equidistant from the other Altaian ethnic groups in our data set. They shared  $F_{ST}$  values around 0.4 with the other three ethnic groups. The Chelkan were also statistically distinctive from all others. The Kumandin and Tubalar were much more similar to each other than either was to the Chelkan. Unexpectedly, the  $F_{ST}$  value between the Tubalar and Altai-kizhi was smaller than the one between the Tubalar and Kumandin, although the p-value was significant (p-value > 0.05). Using a Bonferroni correction (0.0125), Altai-kizhi and Tubalar populations were not statistically significant. This result could explain the results from the SAMOVA of Altaian villages mentioned above. In the only instance where the  $F_{CT}$  value was significant, two northern and two southern villages were grouped together. These northern villages contained a substantial number of Tubalar. In addition, the Tubalar were found in nearly every village. Thus, the similarities between Tubalar and Altai-kizhi removed any signature of variation structure along geographic categories.

Table 4.7  $F_{ST}$  values between Altaian ethnic groups

	Chelkan	Kumandin	Tubalar	Altai-kizhi
Chelkan	*	0.000	0.000	0.000
Kumandin	0.046	*	0.097	0.007
Tubalar	0.041	0.011	*	0.028
Altai-kizhi	0.037	0.017	0.008	*

Table 4.7  $F_{ST}$  values are displayed in the lower matrix. P-values are located in the upper matrix.

Based on these analyses, northern Altaian populations have different relationships with the southern Altai-kizhi. The Chelkan are quite different from any other Altaian

population. Tubalars and Kumandin cluster together, but Tubalars appear to have a closer genetic relationship to the Altai-kizhi than the other northern Altaian populations.

#### **4.5 Within Population Variation**

Summary statistics were calculated for each of the four populations analyzed here (Table 4.8). Gene diversities were calculated using haplogroup and haplotype data as described in Nei (1987). The former was labeled “Haplogroup Diversity” and the other “Haplotype Diversity”. Gene diversity is defined as the probability that two randomly chosen haplotypes from a population are different. Two other statistics are useful in providing additional information about the haplotypes present in each population. These statistics focus on characteristics or mutations of the DNA sequence instead of simply the haplotype as a whole. Nucleotide diversity is essentially the same thing as gene diversity except that it measures the probability that any two nucleotide sites are different between randomly chosen haplotypes.

The second statistic, the average number of pairwise differences, also examines differences between haplotypes in a population. In this case, the statistic calculates the average number of nucleotide differences between two haplotypes in single population. Haplotypes from each population were compared using these statistics to assess the relative amounts of diversity within each population. Mismatch distributions were generated and a raggedness index was calculated for each of the mismatch distributions. These were used in tandem with the average number of pairwise differences to determine if (and when) a population went through a recent population expansion.

Table 4.8 Summary statistics for Altaian ethnic groups

Group	Northern Altaian			Southern Altaian
	Chelkan	Kumandin	Tubalar1	Altai-kizhi1
Population				
Number	91	52	71	276
Haplogroups	14	13	11	25
Haplogroup Diversity	0.886 ± 0.012	0.806 ± 0.047	0.800 ± 0.030	0.866 ± 0.015
Haplotypes	22	18	26	75
Haplotype Diversity	0.923 ± 0.013	0.914 ± 0.021	0.953 ± 0.010	0.976 ± 0.003
Nucleotide Diversity	0.020 ± 0.011	0.022 ± 0.011	0.019 ± 0.010	0.018 ± 0.009
Pairwise Differences	7.68 ± 3.61	8.22 ± 3.87	7.03 ± 3.34	6.84 ± 3.23
Raggedness Index	0.032	0.022	0.010	0.011
P-Value	0.000	0.149	0.635	0.388
Tajima D	1.201	-0.644	-0.701	-1.180
Tajima D P-Value	0.000	0.000	0.000	0.000
Fu's F <sub>S</sub>	3.417	-0.497	-3.877	-24.416
Fu's F <sub>S</sub> P-Value	0.002	0.000	0.000	0.000

Two tests of neutrality were implemented with these data – Tajima's D and Fu's F<sub>S</sub>. Tajima's D is a test statistic that uses two calculations of the population mutation parameter ( $\theta$ ). One  $\theta$  statistic is based on the observed number of segregating sites, and the other is based on the average number of pairwise differences. Negative values reflect deviation from neutrality, but two scenarios can cause this – selection or population expansion. For mtDNA studies, the latter interpretation is generally accepted (Jobling, Hurles, & Tyler-Smith, 2004). The negative values are also typically correlated with populations that have greater sizes. Fu's F<sub>S</sub> provides the same general assessment as Tajima's D. It is estimated using a comparison of  $\theta$  values calculated from the average number of pairwise differences, with  $\theta$  values estimated from the observed number of alleles (haplotypes).

The general trend for the intra-population statistics indicated that differences exhibited by the northern and southern Altaian populations are due to differences in the demographic histories of these two groups. The Kumandin and Tubalar had lower gene diversity estimates (both for haplogroup and haplotype data) and higher average number of pairwise differences. The Chelkan unexpectedly had the highest haplogroup diversity, but their haplotype diversity was low, as predicted. All of the neutrality index values were significant. Kumandin, Tubalar and Altai-kizhi populations had negative Tajima's  $D$  and Fu's  $F_S$ , indicating that they had undergone population expansions. The Chelkan had positive values, indicating an overall decrease in population size, which would, in turn, increase the effects of genetic drift. As noted above, the high frequencies of similar haplotypes in the Chelkan is precisely the genetic pattern that would result from genetic drift.

Mismatch distributions for the southern Altaians, Kumandins and Tubalar showed a unimodal distribution and raggedness indices below 0.03 (Figure 4.4). Visually, the Tubalar and Kumandin did not have mismatch distributions as smooth as the Altai-kizhi, yet their raggedness indices were not significant. On the other hand, the Chelkan possessed a multimodal mismatch distribution. Their raggedness index was 0.032, with a p-value of 0.000, indicating that a model of either population expansion or a stationary population was not a good fit to the data. Therefore, the Chelkan most likely have undergone some form of population decline or bottleneck/founder event without a subsequent population expansion.

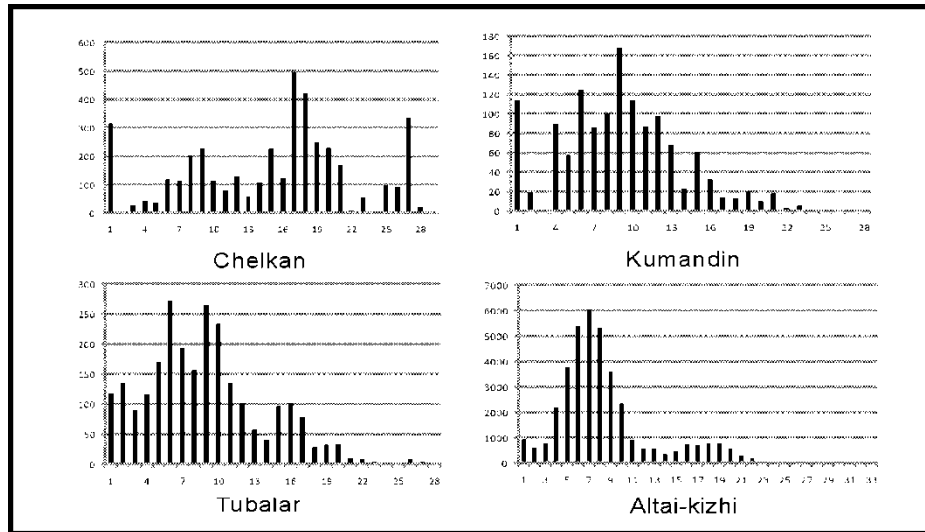


Figure 4.4 Mismatch distributions of Altaian ethnic groups

#### 4.6 Altaian Local Context

Now that the mtDNA variation of the four indigenous Altaians ethnic groups has been examined, their placement in the context of other Altaians can be explored. Five studies have used the mtDNAs from people living in the Altai, although they relied on southern Altaians as a representative sample or used a combination of various Altaian ethnic groups as a single entity. The first study collected sixteen samples from three villages across the southern Altai, but did not identify the ethnic composition of those participants (Shields et al., 1993). The second study broadly examined genetic diversity among southern Siberian populations (Derenko et al., 2003). The Altaians in this study included Altai-kizhi, Telenghits, Maimalars, Tubalars and Chelkan, totaling 110 individuals. The first three of these ethnic groups were classified as southern Altaians, while the last two were designated as northern Altaians. These same authors published a subsequent study with more Altaians, divided along ethnic boundaries (Derenko,

Malyarchuk, Grzybowski et al., 2007). In this study, there were 90 Altai-kizhi, 82 Shor, 71 Telenghits and 53 Teleuts, with the Shor being the sole northern Altaian group. The fourth study explored the mtDNAs of southern and eastern Siberian ethnic groups (Starikovskaya et al., 2005). They included Tubalar and Tubalar-Chelkan from ten villages in the Turochak and Choiski Districts of the northern Altai, including a few locations from which our samples were collected. Finally, the fifth paper utilized a subset of the same samples we have at Penn and collected at Mendur-Sokkon, although inconsistencies exist between this data set and our own (Phillips-Krawczak, Devor, Zlojutro, Moffat-Wilson, & Crawford, 2006).

It is not clear what portion of the Mendur-Sokkon samples are present in both our data set and that of Phillips-Krawczak et al., (2006). Those samples totaled 98 individuals – less than half of our total from that location. Comparisons of the published Mendur-Sokkon samples with our own data showed a number of inconsistencies. The first haplotype in their paper was not found in our samples. The next problem was that two haplogroup A haplotypes that were listed separately, but actually are the same (16223-16249-16290-16319-16362). One of these likely has a typo in the published text. We have three such individuals with that haplotype, plus several with minor variants. Next, a haplogroup Z haplotype was assigned to haplogroup F. This is not a minor difference in nomenclature as these two haplogroups are not even remotely similar, with F belonging to macrohaplogroup R and Z belonging to macrohaplogroup M. Furthermore, Phillips-Krawczak et al. tested for haplogroup F, making it uncertain as to why this haplogroup Z sample came up positive for F markers. We found no such issue with the multiple haplogroup Z samples in our own data set. In addition, several

haplotypes that we have identified as belonging to haplogroup R9b are listed as haplogroup U. Again, the haplogroup U SNP (12308) was tested for in the Phillips-Krawczak et al. study, making this inconsistency in nomenclature inexplicable. Finally, one haplogroup C haplotype has a mutation at 16194, a variant that we did not find in our expanded data set.

Methodological problems also exist. Only a portion of HVS1 was sequenced (from 16151 to 16383) in the Phillips-Krawczak et al. study. The entire HVS1 actually extends from nucleotide positions 16024 to 16383, but the current standard increases this range to include mutations up to and including position 16400. In addition, SNP typing was conducted at a low resolution. Some of the results of this SNP typing cannot be easily explained and conflict with HVS1 sequencing results. Given that our data set was characterized at higher resolutions for both SNP typing and sequencing, and given that we have a larger sample collection from this location without the inconsistencies noted above, we will not use the published data from Phillips-Krawczak et al. study in our analysis.

#### **4.7 Altaian Diversity - Haplogroup Level**

Northern Altaians shared a similar mtDNA haplogroup profile, but distinctions between the ethnic groups were discernable (Table 4.1). Chelkan, Kumandin and Tubalar populations had haplogroups C, D, N9a, H, U4 and U5. Prevalent among these haplogroups, C and D played a significant role. For the two Tubalar populations, these haplogroups were found in nearly a 1:1 ratio, although the haplogroups were slightly more frequent than in Tubalar2 (from Starikovskaya et al., 2005). Additionally, the

Tubalar had higher frequencies of U4 than any other group in southern Siberia. Both Tubalar groups showed the same haplogroup patterns, with the exception of several minor differences – for instance, the presence of F2 and Z in our Tubalars, versus A, F1, U2 and X in Tubalar2. The Kumandin had slightly higher frequencies of U5 compared to other southern Siberians and had markedly greater number of haplogroup C mtDNAs compared to haplogroup D. Chelkan had the opposite pattern (more D than C mtDNAs) and were distinct in their high frequency of F1, F2 and N9a.

While there is still a question as to whether the Shor should be categorized as a member of the northern Altaian cultural group, their mtDNA haplogroup profile indicated that it had some significant differences from the rest of the northern Altaians.

Haplogroups C and D constituted only 24% of their mtDNAs – far lower than any other southern Siberian population. The Chelkan, Telenghit and Khakass were the only groups that came close to this figure (38%, 38% and 35%, respectively). Additionally, 14% of the Shor mtDNAs belonged to haplogroup F1. Finally, N9a and U5 were absent, although they were found in only low frequencies in other northern Altaians. Therefore, if the Shor shared a common maternal ancestry with the northern Altaians, then genetic drift must surely have played a large part in their current levels of genetic diversity, much like the Chelkan.

Like the northern Altaians, the southern Altaian ethnic groups had a similar overall profile, but were different enough to be distinctive from each other. Haplogroups C, D, F1, M11 and U4 were common in southern Altaians. Also, much like the northern Altaians and other southern Siberians, haplogroups C and D were the most prevalent lineages. Frequencies of C and D were roughly equal in Telenghit (~14%) and Teleut



(~24%), but among the Altai-kizhi, C was far more common than D. Furthermore, the Altai-kizhi had greater numbers of haplogroups J, K and X, while Telenghit and Teleut contained more H, T and U2. Telenghits also had greater frequencies of haplogroups A and B, compared to the other southern Altaians.

#### **4.8 Altaian Populations – Haplotype Level**

Pairwise  $F_{ST}$  values were calculated using the HVS1 sequences of each of the Altaian groups and analyzed with the Tamura-Nei substitution model. As before, a complex interaction between the northern and southern Altaian populations was evident. The Shor were an outlier in the MDS plot, and the Chelkan were distinctive from all other Altaian populations (Figure 4.5). These were the only two unique populations among the Altaians. They showed no affinity to any other Altaian population when a Bonferroni correction was used to adjust the significance level for the  $F_{ST}$  genetic distances (0.004) or even when a standard threshold was used (0.05). Not surprisingly, the Chelkan shared the smallest  $F_{ST}$  values with Tubalar2 and the combined Altaian data set of Derenko et al. (2003), because these Altaian data sets contained some Chelkan individuals. The Shor were closest to Chelkan, but even these populations had a very large  $F_{ST}$  value (0.068).

The remaining samples formed two general clusters. The left cluster was made up of mostly southern Altaians, including Altai-kizhi1, Altai-kizhi2 and Teleut, as well as the northern Altaian Tubalar1. The right cluster contained Tubalar2 and Telenghit, with the combined Altaian samples from Derenko et al. (2003) positioned in the middle, closer to the center of the plot. The populations in each of these clusters shared extremely small

$F_{ST}$  values, with non-significant p-values, indicating their genetic similarity.

Nevertheless, the Teleut shared a small  $F_{ST}$  value with the Telenghit. Although it shared smaller genetic distances with Altai-kizhi populations, Teleut did not fall in the same cluster as the Telenghit.

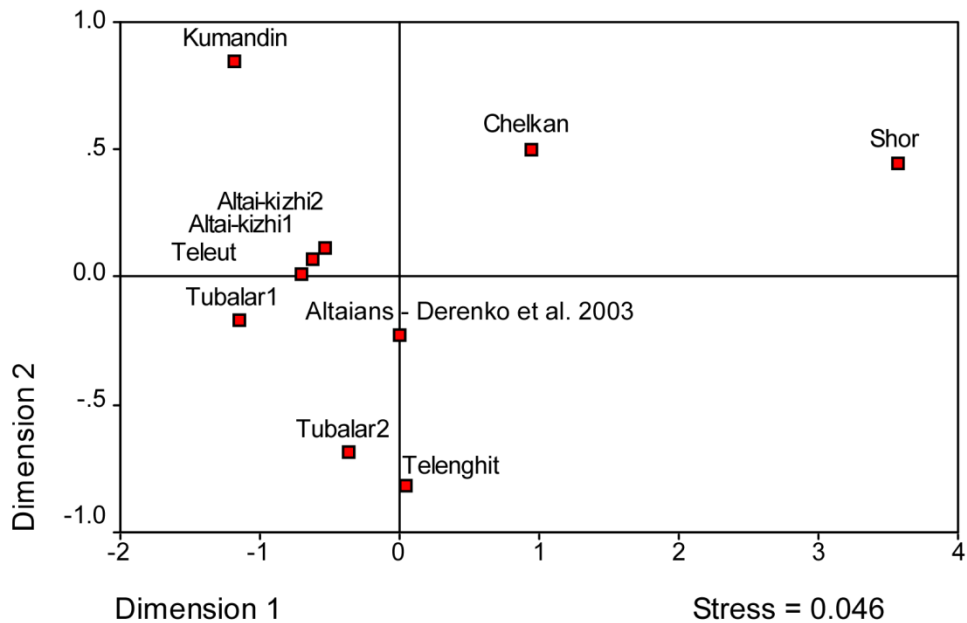


Figure 4.5 MDS plot of Altaian ethnic group  $F_{ST}$  values

The mtDNA variation in northern Altaians suggested complex relationships amongst them. The four northern Altaian ethnic groups were scattered across the entire MDS plot, with both Tubalar populations clustering among southern Altaians. Even though both Tubalar populations clustered with southern Altaians, they did not exhibit genetic similarities between each other. Tubalar2 actually clustered with Telenghit. This finding was not expected, especially given that one northern Altaian village (Artybash) was sampled in both Starikovskaya et al. (2005) and our datasets (although, it is impossible to know how much the sample collection overlapped between studies).

Furthermore, the Kumandin were separated in the MDS plot from the left hand cluster, even though they shared small  $F_{ST}$  values with all of the populations in it. Despite this pattern, the smallest genetic distance for Kumandin was with the Tubalar1.

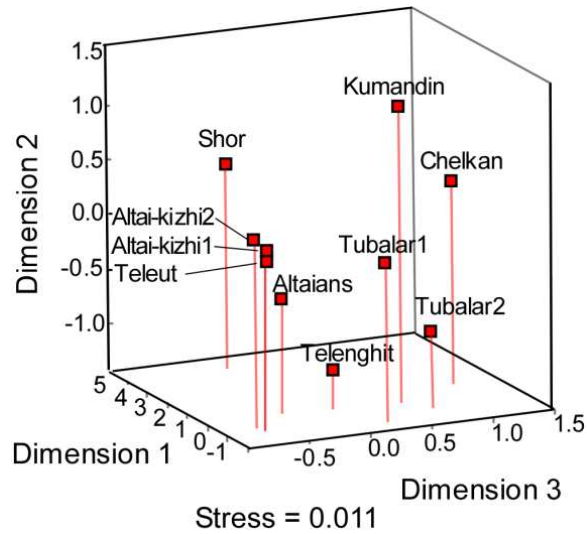


Figure 4.6 MDS plot of Altaian ethnic group  $F_{ST}$  values in three dimensions

There appeared to be little evidence of a distinct split between the northern and southern Altaians based on these genetic distances. However, one could argue for this separation when utilizing a three-dimensional MDS plot (Figure 4.6). The stress value for the plot was smaller (and therefore, more accurately reflects the relationships between the genetic distances), but the  $F_{ST}$  significance values remained the same. Even at a less conservative significance level (0.05), Kumandin and Tubalar were not different from the Teleut. Based on this analysis, a strict division between northern and southern Altaians does not exist at the maternal genetic level.

#### **4.9 Southern Siberians – Haplogroup Level**

To understand the genetic structure within Altaian populations, I included non-Altaian populations that also resided in southern Siberia in the analysis. Tuvinians are one of the largest indigenous populations in southern Siberia. Fittingly, three populations represent this ethnic group in the analysis (Derenko, Malyarchuk, Grzybowski et al., 2007; Pakendorf et al., 2006; Starikovskaya et al., 2005). Also added were three additional ethnic groups that supposedly originate from Tuva, namely Tofalars, Todzhans, and Soyots (Potapov, 1964e; Sergeyev, 1964). The Tofalar were represented by two populations (Derenko, Malyarchuk, Grzybowski et al., 2007; Starikovskaya et al., 2005), and the Soyots and Todzhan by one each (Derenko et al., 2003; Derenko, Malyarchuk, Grzybowski et al., 2007). Two additional Siberian populations were also included – the Khakass (southern Siberia) and Siberian Tatars (western Siberia) (Derenko, Malyarchuk, Grzybowski et al., 2007; Naumova et al., 2008). Finally, a number of northwestern and central Siberian ethnic groups were added because they make up a significant portion of the current populations living to the north and east of Altaians and, in some cases, may have intermarried with other smaller ethnic groups (like the Ket).

Haplogroup C made up about half of the Tuvinian population mtDNAs. The only ethnic group to have more haplogroup C haplotypes than Tuvinians was the Tuvinian-derived Tofalar. Greater than 60% of mtDNAs from both Tofalar populations in this analysis belonged to haplogroup C. Most of the Tuvinian haplogroups were East Eurasian in origin, with higher frequencies of haplogroups D, F1 and G. Todzhan had a similar profile as Tuvinians, but with greater G and less F1. Among the Buryat – a

Mongolic-speaking people – C, D and G made up the majority of mtDNAs, with a greater frequency of haplogroup D. The Soyot – who are known to have originated in Tuva but now live alongside the Buryat – had the highest frequency of haplogroup D among southern Siberians (47%) and also had high levels of A and C. Thus, many of the haplogroups found in these populations are the same, yet occur in different combinations and at different frequencies between ethnic groups.

Principal component analysis (PCA) of haplogroup frequencies in Siberian populations revealed that the Shor and Ket were outliers from other Siberian ethnic groups (Figure 4.7). These groups were set apart in the first component, which accounts for about 72% of the variation. A closer look at their haplogroup profiles showed high frequencies of C, F1 and H. Presumably, this similarity was due to a common origin among the historic Yeniseian populations that helped to create these modern ethnic groups.

The Chelkan and both Khanty populations were located between the Shor/Ket group and the rest of the populations. The Khanty are Ugric-speakers who (along with the Mansi) were believed to have moved north to their current locations. Thus, some of their mtDNAs may be representative of populations that lived in the forest belts of southern Siberia just to the north of the steppe during prehistorical periods.

Among the remaining southern Siberians, two large clusters were present. The left cluster contained Telenghit, Tubalar2, Buryat, Soyot, and Mansi. Both Tofalar populations and the Todzhan were located at the same position on the x-axis, but were separated along the y-axis. The right cluster included Tuvinians, Altai-kizhi, Teleut, Tubalar1, Kumandin and all of the central Siberian populations. It is not exactly clear

what caused this pattern. The majority of haplogroups found in populations of the left-hand cluster tended to be A, B, C and D. The populations of the right-hand cluster have greater overall numbers of C and D, but not as much A or B.

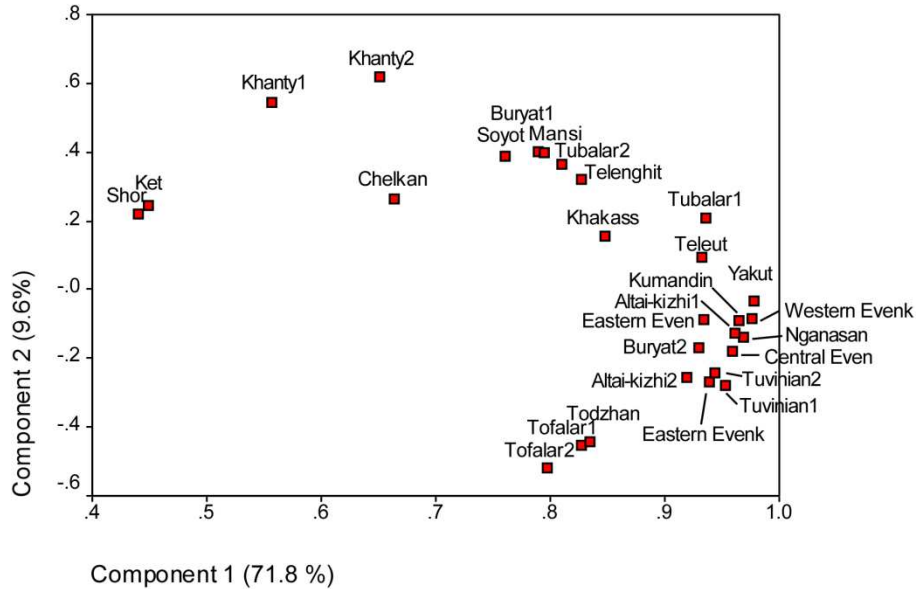


Figure 4.7 Principal component analysis of Siberian populations (1<sup>st</sup> and 2<sup>nd</sup> components)

Northern Altaians were separated along both axes. Tubalar1 were located near Teleut and Khakass, while the Kumandin were located near Even, Evenk and Nganasan populations. The second component, which accounted for about 10% of the variation, differentiated southern and central Siberian populations in the right-hand cluster. The third component displaced the Shor/Ket cluster further from other Siberians, but also distinguished the Buryat and Soyot from the other groups (Figure 4.8). Given these patterns, a closer examination of the haplogroup frequencies and HVS1 sequences is necessary to understand how these populations are related to one another and how they may have interacted, as well as if there are multiple origins for these groups.

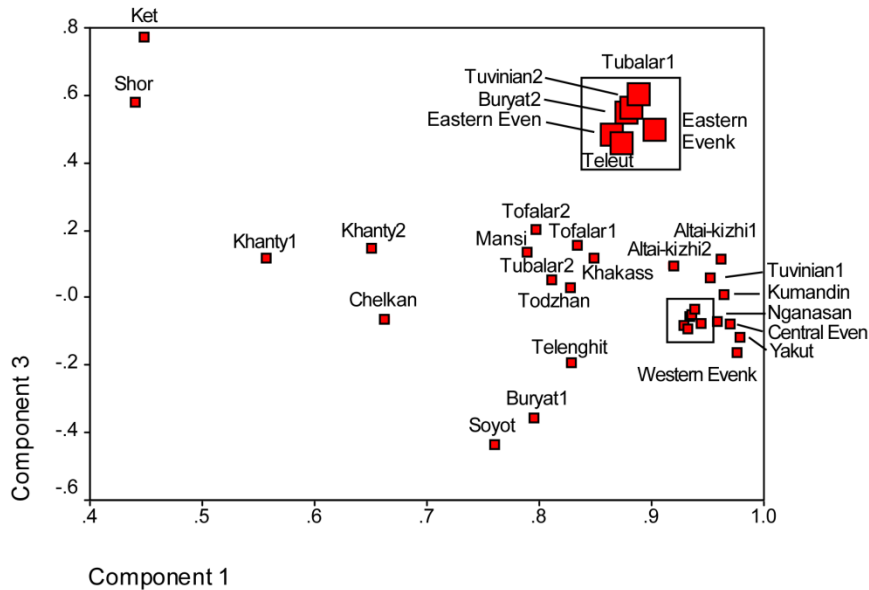


Figure 4.8 Principal components analysis of Siberian populations (1<sup>st</sup> and 3<sup>rd</sup> components)

#### 4.10 Southern Siberians – Haplotype Level

$F_{ST}$  values were calculated and an MDS plot generated to determine the genetic distances between southern Siberian populations (data not shown). From this analysis, it was clear that the three Tuvinian populations were extremely similar. In addition, the two Altai-kizhi populations already discussed showed great similarities. Therefore, only one of each population was retained for further comparative analysis (Altai-kizhi1 and Tuvinian1). Because the Altaians from Derenko et al. (2003) were an amalgam of different ethnic groups, this population was also removed from further analysis.

Altai-kizhi1, Tubalar1 and Teleut fell near the center of the southern Siberian populations in the MDS plot generated from  $F_{ST}$  values (Figure 4.9). Tuvinian, Todzhan and Western Even formed a small cluster to the left of the Teleut and Altai-kizhi, with the

Kumandin positioned in between. Again, the Tubalar, Teleut and Altai-kizhi were not statistically different from one another. This was also true of the Western Even. All of the populations in this central cluster had non-significant  $F_{ST}$  values (p-values < 0.05), showing the close maternal relationship between southern Altaian and Tuvinian populations. This pattern is not drastically different from that seen in the PCA above.

The Chelkan were found to the right of these central clusters close to Khakass and Siberian Tatar populations. While the Chelkan were clustered near these other populations, they were significantly different from all other ethnic groups. The p-value for the Khakass and Chelkan  $F_{ST}$  value was 0.0008, which is just barely passing as a significant value (Bonferroni corrected significance level is 0.001). Clearly, the most similar of all the southern Siberian populations to the Chelkan were the Khakass. However, for the Khakass, several other ethnic groups were more similar (Teleut, Telenghit, Altai-Kizhi1, Tubalar 2 and Siberian Tatar), as the Khakass shared smaller non-significant  $F_{ST}$  values with them. The Siberian Tatar were statistically similar only to the Khakass.

The Uralic-speaking northwestern Siberians clustered together, but they were also close to the Telenghit and Tubalar2. The remaining central Siberians and Nganasan were located to the left of the two central clusters, showing little affinities with Altaians.

Several outliers flanked the central clusters just described. The Shor remained the most distinctive population of all southern Siberian populations and were separated from the rest of the populations in the upper right hand corner of the MDS plot along with the Ket. Of the three Tuvinian-derived ethnic groups, two were outliers. The Soyot were located at the bottom of the plot with the Buryat, and the two Tofalar populations were



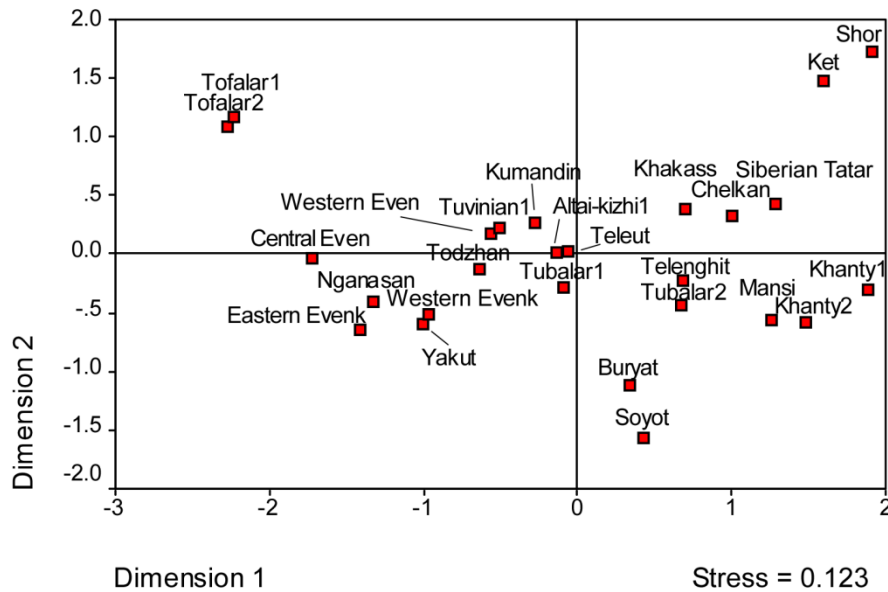


Figure 4.9 MDS plot of  $F_{ST}$  values for Siberian populations

located in the lower left corner. The Soyot clustered with both Tubalar populations, Kumandin, Teleut, Telenghit, and Todzhan populations, when using a Bonferroni correction. The Soyot was the only population associated with the Buryat.

Relationships among ethnic groups in southern Siberia were generally close when considering the mtDNA genetic diversity of those populations. The large pool of southern Altaian and Tuvinian groups identified in these plots revealed a shared maternal history. The fact that some groups were outliers, like the Tofalar, indicated that the size of the numerically smaller ethnic groups has played an important role in determining the characteristics of their current genetic diversity. Among the southern Siberians, the Altaians were not unique. In fact, they all appear to share many of the same maternal ancestors. In addition, the division between the northern and southern Altaians did not persist.

Yet, differences between the northern Altaians were evident. Tubalars and Kumandin had greater genetic affinities with southern Altaians than did Chelkan and Shor. Chelkan and Shor were also unique, and both were equally different from Tubalars, Kumandins, or any of the southern Altaian ethnic groups for that matter. Examination of the three-dimensional MDS plot of southern Siberian  $F_{ST}$  values provided further evidence for the Chelkan being an outlier, similar to the two Tofalar populations (data not shown). The Tubalar, however, seemed firmly entrenched in the main cluster of southern Altaian populations.

Genetic distances showed the general lack of genetic boundaries between northern and southern Altaians and the relative similarity of southern Altaians and Tuvinians. These results suggested that the unit of analysis for genetic studies is not at the village or even ethnic group level, but possibly, at a level that incorporates multiple ethnic groups.

To explore the genetic structure within southern Siberia, analysis of the genetic variation was performed using three AMOVAs. The first investigated whether there is genetic structure among the regional populations in southern Siberia (Table 4.9). Three groups were used – (1) Altai region: Chelkan, Tubalar, Kumandin, Shor, Altai-kizhi, Teleut, Telenghit; (2) Tuvan region: Tuvinians, Todzhans, Tofalars; (3) Baikal region: Buryat and Soyot. The AMOVA results indicated that the genetic variation among groups (the three regions) accounted for 2.80% of the total. Another 3.65% of the genetic variation was accounted for among the populations within each of these three groups.

The second AMOVA used each of the ethnic groups listed above as separate groups, such that there were twelve total groups (Table 4.9). This meant that there were multiple populations were present for the Altai-kizhi, Tuvinians, and Tofalars. The

AMOVA results showed that 4.72% of the genetic variation was accounted for by all twelve ethnic groups and that only 1.12% of the variation was due to difference within ethnic groups (Altai-kizhi, Tuvinians, and Tofalars).

The third AMOVA involved two groups defined by the languages used by southern Siberians. The Buryat and Soyot were grouped into one category, as “Mongolic-speakers”, and the rest were included as “Turkic-speakers” (Table 4.9). Here, the “among group” component accounted for only 0.74% of the total variance, but the p-value was not significant ( $p = 0.221$ ). Of course, this result may not come as a great surprise, as some populations in southern Siberia may have recently adopted Turkic language (like the Khakass), giving up or forgetting their ancestral Samoyedic, Ugric or Yeniseian languages (Menges, 1968; Potapov, 1964c). These results suggested that neither geographic nor current language groupings can explain the genetic structure observed in the southern Siberian populations better than groups defined by cultural affiliations and ethnicity. SAMOVA was used to further explore the relationships between genetic similarity and geographic proximity, but at no point were the groupings statistically significant.

Previous publications used tau ( $\tau$ ) values to estimate the expansion times of a given population. Tau is estimated from the mismatch distributions of each population created when every sample in a population is compared to each other and the number of pairwise differences is recorded. A histogram is often constructed to visualize these differences (see Figure 4.4). Smooth, bell-shaped unimodal curves are suggestive of population expansions. The degree of smoothness for any given curve is characterized by a raggedness index ( $r$ ). Tau is calculated using the highest point of the histogram, with

Table 4.9 AMOVA of southern Siberians

<b>Groups</b>	<b>Percentage of Variation</b>	<b>P-value</b>
<b>Geography</b>		
<i>Among group</i>	2.80	0.002
<i>Among population within group</i>	3.65	0.000
<i>Within population</i>	93.55	0.000
<b>Language</b>		
<i>Among group</i>	0.74	0.221
<i>Among population within group</i>	5.21	0.000
<i>Within population</i>	94.05	0.000
<b>Ethnicity</b>		
<i>Among group</i>	4.72	0.000
<i>Among population within group</i>	1.12	0.000
<i>Within population</i>	94.16	0.000

Table 4.9 Group Membership: Geography (Altai, Tuvan, Baikal Regions); Language (Turkic, Mongolic); Ethnicity (Chelkan, Kumandin, Tubalar, Altai-kizhi, Teleut, Telenghit, Shor, Khakass, Tuvinian, Todzhan, Tofalar, Soyot, Buryat).

each peak representing a wave of population expansion. The timing of this expansion is dependent on a model where a population with the size of  $N_0$  suddenly expanded to a size of  $N_1$ . Using this procedure, the timing of an expansion can be accurately estimated, although the confidence intervals for the initial and final population sizes cannot. Thus, the time of the expansion can be estimated, but no information is obtained on the magnitude of that expansion (Schneider & Excoffier, 1999). Studies that calculate the expansion times often give tau estimates of 30 to 65 kya (Jobling et al., 2004). In the case of southern Siberians, the estimates fell between 22 and 49 kya (Derenko et al. 2003).

Such estimates will not be calculated for this dissertation because they do not seem biologically meaningful. I know of no population that has existed in isolation for 22,000 years, let alone 49,000 years. If these estimates are taken at face value, then it

implies population expansions (either demographic or spatial) some time before the LGM in southern Siberia. I would argue that these estimates conflate the different gene histories for each of the mtDNA haplogroups present in a population into this single population expansion parameter. Further discussion on this issue is addressed in the next chapter.

#### **4.11 Altaians from a Global Perspective**

Placement of the southern Siberian populations into a broader geographical context further highlights the differences between the Altaian groups (Figure 4.10). Central Asian (green) populations tended to cluster closer together more so than southern (red) or central (blue) Siberians. Central and southern Siberian populations showed similarities mostly because of the southern Altaians, Tuvinians and Yakut populations. Yakut are believed to have originated in southern Siberia, moving eastward relatively recently (Khar'kov et al., 2008; Pakendorf et al., 2006; Puzyrev et al., 2003; Tokarev & Gurvich, 1964; Zlojutro et al., 2009). Northwestern Siberians (violet) tended to fall in the large Central Asian cluster. The only exception was the Nganasan, who had greater affinities with some Evenk, Even and Tuvinian populations, presumably due to a high occurrence of the same haplogroup C haplotypes.

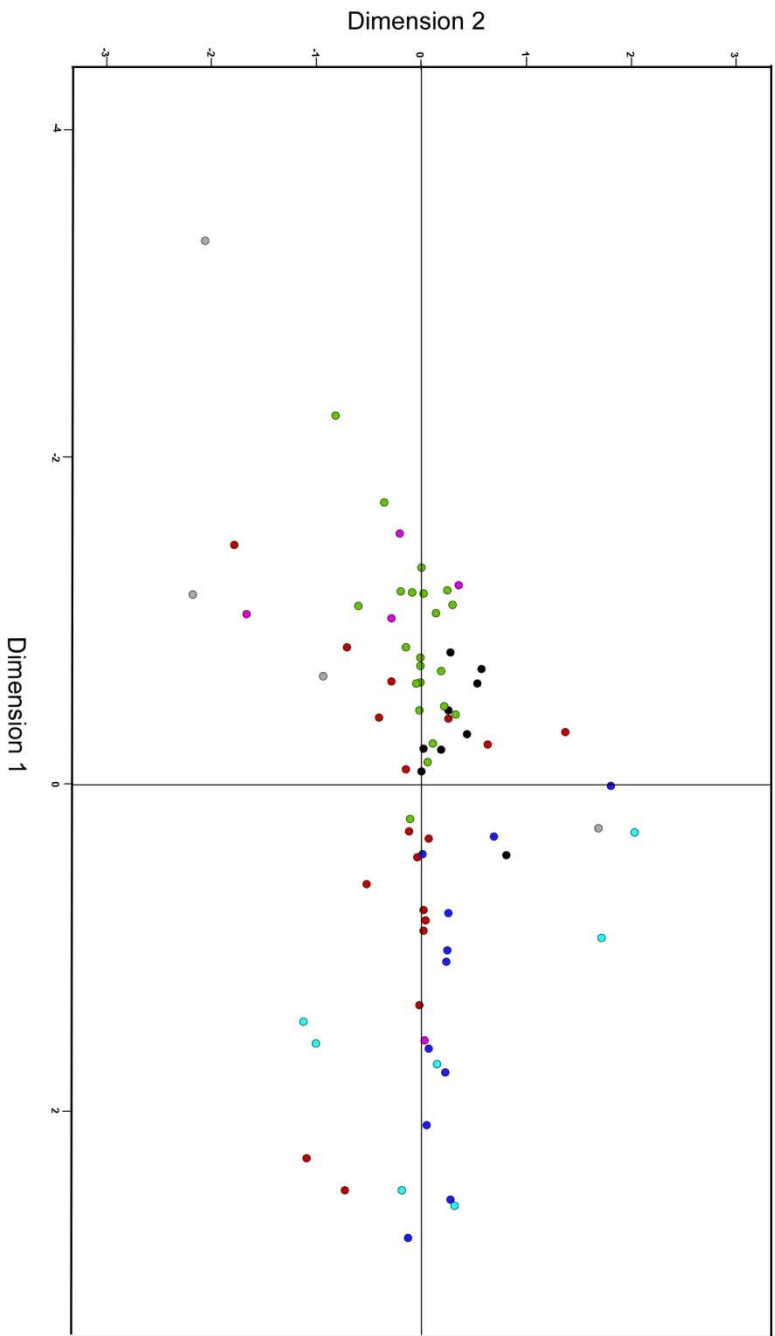
Some southern Siberians like the Chelkan, Shor, Khakass and Siberian Tatar, fell inside the cluster of Central Asian populations. Mongolian populations (black) also were located near the Central Asians, but they retained a tighter cluster than the southern Siberians. Many of the northeastern Siberians (Itel'men, Yukaghir and Koryak) were at the periphery of the MDS plot. The southeastern Siberians did not cluster tightly together

either. The Negidal and Orok were clear outliers, the Udegey fell near a Koryak population, and the Ulchi landed near the Siberian Tatars, nearby the Khakass, Mansi and Khanty. The small sizes of many of the eastern Siberian populations certainly had influence on their locations in the MDS plot.

Associations between Altaian ethnic groups and populations outside of southern Siberia indicated that differences in Altaian maternal gene pools exist. Most southern Siberians shared small genetic distances with Kyrgyz from Comas et al. (2004). Despite one theory that asserts that modern-day Kyrgyz are the direct descendents of the historical Yenisei Kyrgyz who once lived in southern Siberia, evidence from other Kyrgyz populations suggested otherwise. The lowland and highland Kyrgyz from Comas et al. (1998) did not exhibit the same affinities to southern Siberians.

Altaians shared relatively small genetic distances with other Central Asians as well. Telenghit, Teleut and Tubalar<sup>2</sup> had low  $F_{ST}$  values in common with Highland Kyrgyz, Kyrgyz, Kara-Kalpaks, some Kazakhs and Mongolians. Altai-kizhi from Derenko et al. (2007) also showed a close affinity with Mongolians. The genetic distances that the Telenghit shared with Central Asians were generally smaller than other Altaian groups.

Outside of southern Siberia, Tuvinians were closely related to central Siberian populations – Evens and Evenks. The populations from the Baikal region – Buryat and Soyot – shared smaller genetic distances with Mongolian and Uyghur (both in northwestern China), Kalmyk and Khamnigan.



- Green – Central Asia
- Red – Southern Siberia
- Dark Blue – Central Siberia
- Light Blue – Northeastern Siberia
- Violet – Northwestern Siberia
- Gray – Southeastern Siberia
- Black – Mongolia / Northern China

Figure 4.10 MDS plot of  $F_{ST}$  values for Siberian and Central Asian populations

#### 4.12 Chapter Conclusions

Based on the observations from this chapter, the maternal genetic diversity of southern Siberia can be described as a result of the interactions between populations in neighboring geographic regions. Southern Siberia served as a crossroads in antiquity for cultures inhabiting the steppe, the forest belts and the mountainous terrain of the region. The complex relationships among historically attested populations in southern Siberia make it difficult to associate any one modern ethnic group as descendants of any one archaeological culture. However, major distinctions remain. A line can be drawn whereby the Chelkan, Shor, Khakass, Ket, and Khanty are placed on one side and the Altai-kizhi, Tuvian and Tuvian derived populations are placed on the other. This division has already been identified based on cultural/ethnographic evidence and even cranial morphology. In this case, the mtDNA variation corroborates this divide, although this divide is permeable.

At the local level, one major objective was to characterize the mtDNA variation in northern Altaian ethnic groups and compare it with that of southern Altaians. The mtDNA data showed evidence of differences among northern and southern Altaians. This split reinforces the notion among anthropologists that these two groups grew from different origins, with the northern Altaian populations having been influenced largely by the historical Yeniseian and Southern Samoyedic groups that were known to have inhabited the region before Russian colonization. The southern Altaians likely evolved from the historical steppe populations that resided along the Altai Mountains, which likely felt a greater influence from Mongolian expansions and its political hegemony in the several centuries before Russian annexation.



Nevertheless, the northern Altaian story is not the same for all ethnic groups. While the Chelkan certainly show the greatest differences among the Altaian populations, and the Kumandins (like the Chelkan) appear to have greater affinities with Samoyedic and Yeniseian populations, the Tubalar clearly have a great affinity with southern Altaians. Tubalar are believed to have intermarried with southern Altaians, which would help to explain this genetic pattern (Potapov, 1962). Tubalar are also believed to have originated from Tuva, moving into the Altai only recently, which would explain their greater affinities with Tuvinians, as well (Potapov, 1964a). The most likely scenario is one in which a small group arrived from Tuva and lived amongst the historical Samoyedic and Yeniseian populations of the region. Either during or after the process of incorporating these people, the Tubalars begin interacting with their neighbors immediately to south to resemble their current genetic composition.

In the early historical period, the populations living along the Yenisei and in the northern Altai likely spoke non-Turkic languages. Whether these were Yeniseian, Samoyedic or Ugric speakers, it cannot be deciphered at this point, but they did not speak Turkic as they do today (Menges, 1968; Vajda, 2001), nor should it be assumed that they were monolingual. It is not clear when exactly these groups adopted Turkic languages, but they were certainly being spoken by groups such as the Yeniseian Kyrgyz in the 8<sup>th</sup> century CE. The current day Altaian ethnic groups formed out of these tribes living in the Altai region probably a couple hundred years ago (Potapov, 1962). A scenario could be proposed in which tribes intermarried and their cultures gradual melded to form the current northern Altaian ethnic groups. A similar situation was observed with the Khakass, who only recently became a recognized ethnic group (Potapov, 1964c). The

Khakass formed out of five different tribal elements with varying cultural and language usage (some were Yeniseian others were Southern Samoyedic) less than a century ago to become one single ethnic group as they are recognized as being today.

The northern Altai likely went through a similar process in the recent past. Whether each of the current ethnic groups was present then as they are now, or whether they originated from a single source population, is not known. However, the mtDNA evidence suggests at least some portion of their maternal ancestry came from common descent of the historical indigenous non-Turkic-speaking groups. In either case, the ethnic groups present in the northern Altai now are distinctive. While they do share some common ancestry with each other, it seems likely that differences between these groups have persisted in part because of clan structures but also because they tended to live in small, endogamous communities, such that even today the Chelkan are unique from the rest of the northern Altaians.

All evidence points to northern Altaian ethnic groups being affected more by genetic drift and in having greater interactions with their neighbors to the north, while the southern Altaians all seemed to possess greater population sizes due to their economic and subsistence strategies. This pattern allowed them to retain the genetic diversity that persisted in their populations. The closer affinities of southern Altaians to Central Asian groups lends support for the historical interaction between these different regions, possibly from the nomadic pastoralist tribes long known to inhabit the steppe. The possibility of teasing apart how (and when) these historical interactions took place is subject of the next chapter on the phylogeography of mtDNA haplogroups, where differences between Paleolithic, Neolithic and Metal Age migrations are explored.

Regardless, the variation found among these groups suggests a close association among southern Siberian populations as compared to eastern Siberian groups.

Ultimately, the northern Altaian groups share more cultural, linguistic, ethnographic and genetic affinities with Ugric, Samoyedic and Yeniseian groups, and thus seem to represent those populations that originated from the hunter-gatherers mentioned even in ancient Chinese and Turkic records. The southern Altaians appear to have originated from the cultures of the steppe. As a result, the Altai has consistently been a frontier and a location of great genetic admixture, serving as a genetic boundary between two regions, each having their own cultures, lifestyles, modes of subsistence and ways of life.

## Chapter 5: Phylogeography of mtDNA Haplogroups

Based on the data from the previous chapter, several mtDNA haplogroups stand out as being the most important in Altaian populations. East Eurasian haplogroups C and D are certainly significant given the high frequencies of these types in all Siberian populations. Haplogroups U4 and U5 also make up the most significant portion of the northern Altaians' West Eurasian haplotypes. The goal of this chapter is to apply phylogenetic analyses to these important haplogroups and thereby gain a better understanding of the origin and distribution of each. In using a phylogeographic approach, it is possible to understand when these mtDNA lineages arose and possibly, when they appeared in a given region. In doing so, we can gain a diachronic perspective that will allow us to better resolve the histories of Altaian populations.

Complete genomes were retrieved from GenBank, and selective pressures were assessed for each of the four haplogroups just mentioned (C, D, U4 and U5). These methods included calculations of synonymous and nonsynonymous mutations,  $K_A/K_S$  ratios, and z-codon tests ( $d_N/d_S$ ). Maximum likelihood trees were constructed for each haplogroup and branch lengths were assessed using two models, one based on a molecular clock and another allowing for non-clocklike behavior.

After assessing each haplogroup for evidence of natural selection, I calculated coalescence dates for each of the haplogroups, and where appropriate, several of their branches. This was accomplished by using the methodology of Soares et al. (2009), where a series of six reduced median-median joining networks were created and rho estimates obtained for each MRCA of interest. Algorithms generated by Soares et al. were used to convert the rho estimates into TMRCAs. This step allowed me to reanalyze

each haplogroup and determine both the age at which it arose and the time at which it most likely became prevalent in Altaians.

## **5.1 Basal mtDNA Haplogroup Phylogenies**

The first step in understanding mitochondrial phylogeography is to examine the current phylogenies for the four haplogroups in question. This section will introduce these four haplogroups and note the mutations that comprise their major branches (or sub-haplogroups). The type of mutation is also noted (synonymous, nonsynonymous, control region, tRNA, rRNA, etc.) to help ascertain whether any of these mutations have the potential to be targeted by positive selection (Figure 5.1).

Siberian populations largely consist of haplogroup C mtDNAs. For this reason, comprehension of the phylogeny for haplogroup C is essential for understanding the population histories of Siberian peoples, and therefore, Altaians. Initially, this haplogroup was defined as any mtDNA possessing the loss of the 13259 Hinc II RFLP, which is associated with a root HVS1 motif (16223-16298-16362) (Torroni, Schurr et al., 1993). It is one of five haplogroups that comprise nearly all indigenous Americans mtDNAs (Schurr et al., 1990; Torroni, Schurr et al., 1993; Torroni et al., 1992; Wallace et al., 1985). These haplogroups originated in Asia, making Siberians the most likely source for the Native American groups (Shields et al., 1993; Torroni, Schurr et al., 1993; Torroni et al., 1992; Torroni, Sukernik et al., 1993). The sub-haplogroups of haplogroup C were initially defined using HVS1 motifs, which only later were expanded to include full mitochondrial genomes (Achilli et al., 2008; Kivisild et al., 1999; Kong, Yao, Sun et

al., 2003; Starikovskaya et al., 2005; Tamm et al., 2007; Tanaka et al., 2004; Volodko et al., 2008).

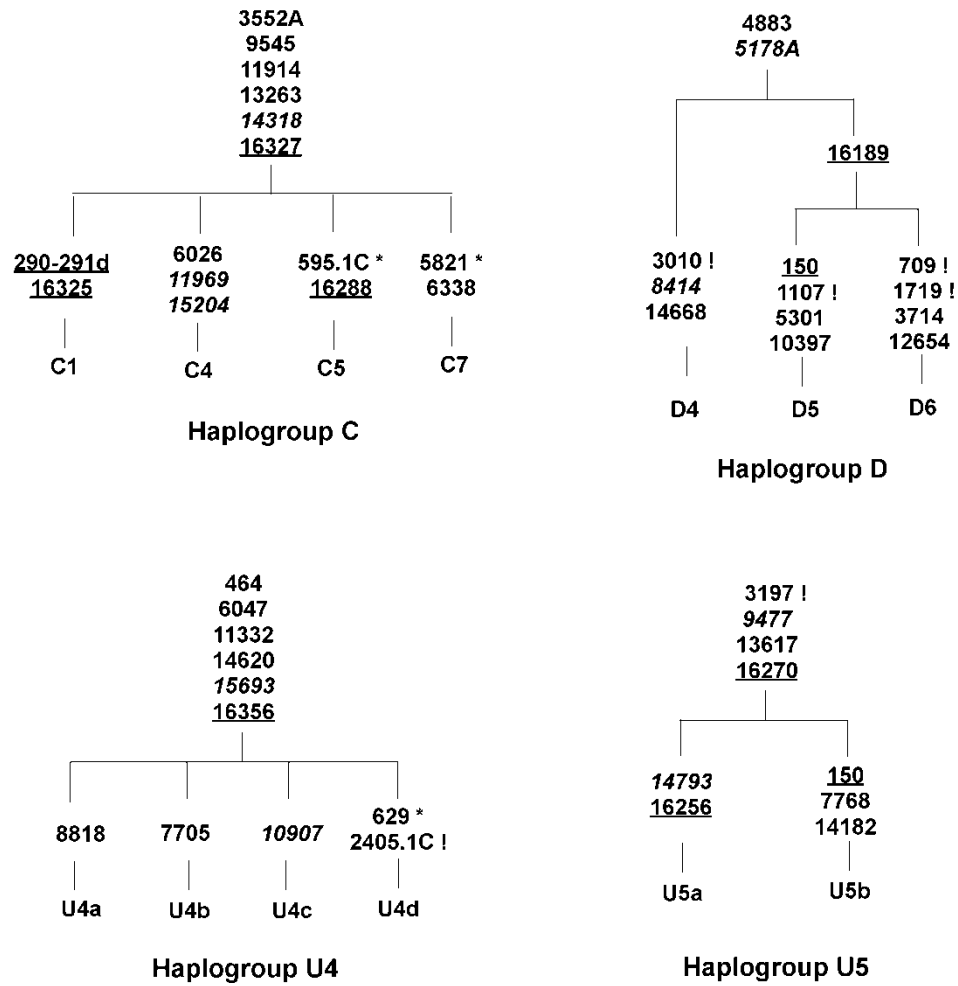


Figure 5.1 Basal mtDNA haplogroup phylogenies. Synonymous mutations have no additional formatting. Nonsynonymous polymorphisms are italicized. Control region polymorphisms are underlined. Polymorphisms occurring in tRNA and rRNA regions are indicated by an asterisk and exclamation point, respectively.

The current haplogroup C phylogeny consists of four major branches (C1, C4, C5 and C7). In addition to the core motif, C1 is defined by polymorphisms at 16325 and deletions at 290 and 291. The highest frequency of C1 occurs in the Americas, but it is also found in Asia (Achilli et al., 2008; Derenko et al., 2003; Kolman et al., 1996; Schurr

& Wallace, 2003; Tamm et al., 2007). C4, C5 and C7 are found throughout eastern Eurasia and parts of South Asia (Chandrasekar et al., 2009; Comas et al., 1998; Comas et al., 2004; Derbeneva, Starikovskaia, Volod'ko, Wallace, & Sukernik, 2002; Derbeneva, Starikovskaya, Wallace, & Sukernik, 2002; Derenko et al., 2003; Derenko, Malyarchuk, Grzybowski et al., 2007; Heyer et al., 2009; Kivisild et al., 2002; Kolman et al., 1996; Kong, Yao, Liu et al., 2003; Kong, Yao, Sun et al., 2003; Pakendorf et al., 2003; Pimenoff et al., 2008; Schurr et al., 1999; Starikovskaya et al., 1998; Tanaka et al., 2004; Torroni, Sukernik et al., 1993; Yao, Kong, Bandelt, Kivisild, & Zhang, 2002; Yao & Zhang, 2002). The difficulty with two of these branches is that the diagnostic mutations for C4 and C7 branches are defined by polymorphisms located in the coding region or HVS2. Most studies only use a combination of HVS1 sequence and RFLPs specific at the haplogroup level to characterize the mtDNAs. Therefore, knowledge of markers that could define these sub-haplogroups is not available in many studies. As a result, all available haplogroup C mtDNAs cannot be accurately placed into one of the four known branches.

Moving to haplogroup D, this lineage is one of the oldest mtDNA haplogroups in East Asia. Its distribution stretches from Eastern Europe to the Pacific and is one of the founding mtDNA haplogroups for New World populations. It has been reported throughout China and makes up a significant portion of Siberian mtDNAs. Haplogroup D was first identified by a PCR-RFLP test (Torroni et al., 1992). Correlated with this RFLP was a less distinctive HVS1 motif (16223-16362). We now know that haplogroup D is defined by three mutations: one synonymous mutation at 4883, one control region mutation at 16362 and one non-synonymous mutation at 5178, which is a C to A

transversion. The 5178A mutation changes the amino acid from a leucine to a methionine; both are neutral non-polar amino acids.

Analysis of complete mtDNA genomes allowed for identification of three large branches belonging to haplogroup D (Kivisild et al., 2002; Kong, Yao, Sun et al., 2003; Tanaka et al., 2004). These distinct branches are D4, D5 and D6. D4 is unique in that it is comprised of about 15 different haplotype clusters, far more than typically found for mtDNA haplogroups (van Oven & Kayser, 2009). All fifteen clusters share three polymorphisms by which D4 sub-haplogroups are differentiated from all other haplogroup D mtDNAs. One polymorphism is located at 3010 is located in the 16s rRNA gene locus and is a recurrent mutation. The second mutation is a non-synonymous mutation located at 8414 in the ATP8 gene. Here, a leucine is changed to a phenylalanine, which like the previous non-synonymous mutation involves two neutral non-polar amino acids. The final polymorphism is a synonymous mutation that occurs at 14668 in the ND6 gene.

The remaining two branches of haplogroup D both share a mutation at 16189 in the control region. D5 is also defined by four additional polymorphisms. The first is located at 150 in the control region. The second is located at 1107 in the 12s rRNA gene. The third is a non-synonymous mutation at 5301 in the ND2 gene at position 5301. In this instance, an isoleucine is replaced by a valine; again, both are neutral non-polar amino acids. The final mutation is a synonymous mutation occurring at 10397 in the ND3 gene.

D5 has three clusters of its own. The first two (D5a and D5b) share a synonymous mutation at 9180. D5a is designated by three polymorphisms: one mutation



at 752 in the 12s rRNA gene at position 752, one synonymous mutation at 11944 and one non-synonymous mutation at 12046 in the ND4 gene. The two amino acids that were replaced are neutral non-polar (isoleucine to valine). D5b is defined by five polymorphisms. These include one located at 456 in the control region, 681 and 1048 are in the 12s rRNA gene, and 5153 and 15724 are synonymous mutations in the ND2 and CytB, respectively.

D5c is distinct from the other D5 clusters. It possesses four mutations in the coding region and five in the control region. Those in the control region are at 151 and 152, plus a back mutation at 16189, flanked by two base insertions in each of the polycytosine tracks before and after the nucleotide position at 16189. Of the four coding region mutations, two are synonymous mutations (at 4200T and 15622) and two are non-synonymous mutations (at 4216 in ND1 and 14927 in CytB). Both of the D5c non-synonymous mutations involve alterations in amino acid classes. The 4216 polymorphism is a replacement of a tyrosine (neutral polar) with a histidine (basic polar), and the 14927 polymorphism is a replacement of a threonine (neutral polar) with an alanine (neutral non-polar).

D6 is defined by four polymorphisms. Two are located in the 12s rRNA gene (709 and 1719), and two are synonymous mutations (3714 and 12654).

The two West Eurasian haplogroups analyzed here belong to haplogroup U. Haplogroup U was first identified by a PCR-RFLP test (Torroni et al., 1996). The presence of a Hinf I cut at nucleotide position 12308 marked the presence of this haplogroup. It was later determined through additional RFLP and sequencing analysis that haplogroup U is extremely diverse. Many of the branches of U actually have distinct

geographic distributions as well as varying levels of diversity (Achilli et al., 2005; Finnila, Hassinen, Ala-Kokko, & Majamaa, 2000; Malyarchuk, 2004; Richards et al., 1998; Tambets et al., 2004). Among these branches, two are significant for Siberian and northern European populations – U4 and U5. U4 was first identified by Hofmann et al. (1997) with the discovery of the 4646 mutation in the coding region, although it was officially named “U4” at a later date (Hofmann et al., 1997; Richards et al., 1998). U5 was first called “Group 5” by Richards et al. (1996) and was defined by HVS1 mutation 16270 (Richards et al., 1996). Later, this cluster was named “U5” once Group 5 was determined to also possess the 12308 Hinf I RFLP site (Richards et al., 1998).

The current haplogroup U phylogeny consists of four main branches. These include U1, U5, U6 and all other U sub-haplogroups. U4 falls into the last of these categories, as it shares a mutation at 1811 in the 16S rRNA gene with U2, U3, U7, U8, and U9. U4 and U9 also share three mutations: 195 and 499 in the control region and 5999, which is a synonymous mutation. U4 is further differentiated by six mutations: 4646, 6047, 11332, 14620, 15693 and 16356. The polymorphism at 16356 is found in the control region, while 15693 is a nonsynonymous mutation in CytB, changing methionine to threonine (neutral non-polar to neutral polar amino acid). The rest of the diagnostic mutations for U4 are synonymous.

U4 has four branches, U4a – U4d. Mutations for U4a and U4b (7705 and 8818) are synonymous. The mutation for U4c (10907) is a nonsynonymous mutation, changing a phenylalanine to leucine in ND4 (both are neutral non-polar). U4d has two mutations – one in the tRNA for phenylalanine (629), and the other in 16s rRNA (2405.1C).

Haplogroup U5 is considered one of the oldest mtDNA haplogroups in Europe, being associated with the Paleolithic modern humans who first inhabited Europe (Richards et al., 2000). U5 is defined by three coding region and one control region mutation. The first coding region mutation is located in the 16S rRNA encoded section of the mtDNA genome (3197). The second is a nonsynonymous mutation at 9477 in the CO3 gene. It encodes for a change from valine to isoleucine; both are neutral non-polar amino acids. The third coding region mutation is a synonymous mutation at 13617. The single control region mutation occurs at 16270.

U5 only has two main branches, U5a and U5b. U5a is defined by a nonsynonymous mutation at 14793 in the CytB gene and a mutation in the control region at 16256. The non-synonymous mutation changes a histidine to an arginine; both are basic and polar amino acids. One sub-cluster of U5a is identified by yet another non-synonymous mutation at 15218 in the CytB gene (a change from a threonine to an alanine) and one control region mutation at 16399. Like the previous non-synonymous mutations, there is no difference in acidity (neutral) with this change, but there is in polarity (polar to non-polar).

U5b is defined by three mutations, including one control region mutation at 150, and two synonymous mutations at 7768 and 14182. U5b has three major branches (U5b1, U5b2 and U5b3). U5b1 is defined by a mutation at 5656, which is located in one of the few non-coding regions outside of the control region. U5b2 is defined by two mutations: 1721 and 13637. The 721 polymorphism is located in the 16S rRNA encoded region, while 13637 is a non-synonymous mutation in the ND5 gene. This mutation causes a change from a glutamine to an arginine. Both amino acids are polar, but

glutamine is neutral, while arginine is basic. The final branch, U5b3, is defined by a control region mutation at 16304 and a synonymous mutation at 7226.

To gain a better understanding of the phylogeography of each haplogroup, haplotypes were retrieved from GenBank and their sequences analyzed to determine the possible effects of selection and non-clocklike evolution. Once the effects of selection were assessed, TMRCA's were estimated for each haplogroup, as well as for some specific sub-haplogroup branches (see below).

As of March 2010, there were 162 complete genomes in GenBank belonging to mtDNA haplogroup C. Of these 162 genomes, 121 were unique. Haplogroup D was represented by 642 complete genomes; 521 were unique. Haplogroup U4 had the fewest complete mtDNA genomes with 82, of which, 76 represented unique haplotypes. Finally, there were 158 complete U5 genomes available in GenBank, represented by 135 unique haplotypes.

## **5.2 Assessment of Natural Selection on mtDNA Haplogroups**

To assess the effects of natural selection on each haplogroup, all 13 protein encoded regions were analyzed separately to determine the effects (if any) of selection on the mtDNAs belonging to each haplogroup. Two tests were employed that utilized the nonsynonymous versus synonymous mutation approach. The first was a  $K_A/K_S$  ratio test. In this test, the  $K_A$  (the number of nonsynonymous sites per number of possible nonsynonymous sites) was compared to  $K_S$  (the number of synonymous sites compared to the number of possible synonymous sites). The estimates of possible nonsynonymous and synonymous sites were obtained from DnaSP v5, and Fisher's Exact Tests were used

to check for significant differences. The second was the  $d_N/d_S$  estimate. This is similar to the  $K_A/K_S$  ratio tests, but uses a different method for determining the number of differences between nonsynonymous and synonymous mutations.

### 5.3 $K_A/K_S$ Ratios

Looking first at the  $K_A/K_S$  ratios of haplogroup C, significant differences in the number of synonymous and nonsynonymous mutations were present for eleven of the thirteen protein encoded regions (Table 5.1). The exceptions were sequences that encode for ATP6 and ND6. They had  $K_A/K_S$  ratios around 0.8 and non-significant P values ( $P = 0.494$  and  $0.509$ , respectively). Thus, the numbers of synonymous to non-synonymous mutations were not significantly different from each other, suggesting neutrality for only ATP6 and ND6. The overwhelming majority of genes showed evidence of purifying selection. Previously, Mishmar et al. (2003) identified ATP6, CO3 and ND6 as having greater variability in arctic “populations.” While ATP6 and ND6 did have higher numbers of non-synonymous mutations, the pattern actually showed a lack of purifying selection in these regions, not adaptive selection. Considering the overall number of synonymous and nonsynonymous mutations from all genes, a clear pattern of purifying selection is evident.

The  $K_A/K_S$  ratios for the thirteen genes from haplogroup D mtDNAs were mostly between 0.08 and 0.20 (Table 5.2). The only exception was ATP6 with a value of 0.57, being the only gene with more non-synonymous than synonymous mutations. The total  $K_A/K_S$  ratio for all genes also fell between 0.08 to 0.20. All Fisher’s Exact Tests indicated significantly different numbers of synonymous to non-synonymous mutations.

The ratios close to zero and the preponderance of synonymous mutations suggested the presence of purifying selection on the haplogroup D mtDNA genomes at all genes.

Table 5.1  $K_A/K_S$  ratios for haplogroup C

Gene	# Syn	# Possible Syn	# NonSyn	# Possible NonSyn	$K_S$	$K_A$	$K_A/K_S$	Fisher's Exact Test
All	144	2900.92	84	8475.08	0.0496	0.0099	0.1997	P = 0.000
ATP6	3	175.36	11	502.64	0.0171	0.0219	0.7815	P = 0.494
ATP8	7	47.37	1	156.63	0.1478	0.0064	0.0432	P = 0.000
CO1	21	390.33	2	1151.67	0.0538	0.0017	0.0323	P = 0.000
CO2	9	169.91	6	514.09	0.0530	0.0117	0.2203	P = 0.005
CO3	10	196.32	3	586.68	0.0509	0.0051	0.1003	P = 0.000
CytB	15	286.34	15	853.66	0.0524	0.0176	0.3354	P = 0.003
ND1	15	254.29	15	696.71	0.0590	0.0215	0.3650	P = 0.006
ND2	10	264.40	7	776.60	0.0378	0.0090	0.2393	P = 0.004
ND3	5	84.02	2	260.98	0.0595	0.0077	0.1287	P = 0.013
ND4	14	357.36	5	1019.64	0.0392	0.0049	0.1251	P = 0.000
ND4L	4	81.67	0	215.33	0.0490	0.0000	0.0000	P = 0.006
ND5	28	457.90	10	1354.10	0.0612	0.0074	0.1207	P = 0.000
ND6	3	135.65	7	386.35	0.0221	0.0181	0.8192	P = 0.509

For U4, only five of thirteen protein-encoded loci showed significant differences between  $K_A$  and  $K_S$  estimates. These included CO1, CO2, CO3, ND4, and ND5 (Table 5.3). For the most part, the number of nonsynonymous versus synonymous mutations was relatively similar, leading to a higher overall  $K_A/K_S$  ratio. These results implicated purifying selection as acting on the mtDNAs from this haplogroup, but not to the same degree as the other three haplogroups being investigated here.

The  $K_A/K_S$  ratios for haplogroup U5 were around 0.10, except for CO1, CO3 and ATP6, which were 0.04, 0.38 and 0.81, respectively (Table 5.4). ATP6 was remarkable in that twice as many non-synonymous mutations were present as compared to

Table 5.2  $K_A/K_S$  ratios for haplogroup D

Gene	# Syn	# Possible Syn	# NonSyn	# Possible NonSyn	$K_S$	$K_A$	$K_A/K_S$	Fisher's Exact Test
All	468	2900.99	233	8482.01	0.1613	0.0275	0.1703	P = 0.000
ATP6	23	178.37	37	502.63	0.1289	0.0736	0.5709	P = 0.021
ATP8	12	49.77	4	157.23	0.2411	0.0254	0.1055	P = 0.000
CO1	52	389.98	21	1152.02	0.1333	0.0182	0.1367	P = 0.000
CO2	24	170.01	8	513.99	0.1412	0.0156	0.1103	P = 0.000
CO3	27	196.40	15	586.60	0.1375	0.0256	0.1860	P = 0.000
CytB	54	286.10	33	853.90	0.1887	0.0386	0.2048	P = 0.000
ND1	49	255.26	25	698.74	0.1920	0.0358	0.1864	P = 0.000
ND2	44	263.77	17	777.23	0.1668	0.0219	0.1311	P = 0.000
ND3	14	83.99	4	261.01	0.1667	0.0153	0.0919	P = 0.000
ND4	55	357.32	16	1019.68	0.1539	0.0157	0.1019	P = 0.000
ND4L	10	81.67	2	215.33	0.1224	0.0093	0.0759	P = 0.000
ND5	83	455.02	41	1353.98	0.1824	0.0303	0.1660	P = 0.000
ND6	21	133.33	10	389.67	0.1575	0.0257	0.1629	P = 0.000

Table 5.3  $K_A/K_S$  ratios for haplogroup U4

Gene	# Syn	# Possible Syn	# NonSyn	# Possible NonSyn	$K_S$	$K_A$	$K_A/K_S$	Fisher's Exact Test
All	80	2907.39	53	8480.61	0.0275	0.0062	0.2271	P = 0.000
ATP6	3	178.02	7	502.98	0.0169	0.0139	0.8261	P = 0.352
ATP8	2	50.38	3	156.62	0.0397	0.0192	0.4824	P = 0.349
CO1	19	390.02	3	1151.98	0.0487	0.0026	0.0534	P = 0.000
CO2	7	169.98	4	514.02	0.0412	0.0078	0.1889	P = 0.007
CO3	4	195.66	2	587.34	0.0204	0.0034	0.1668	P = 0.037
CytB	6	285.96	8	854.04	0.0210	0.0094	0.4466	P = 0.112
ND1	2	255.31	3	698.69	0.0078	0.0043	0.5479	P = 0.404
ND2	5	264.67	5	776.33	0.0189	0.0064	0.3409	P = 0.083
ND3	1	84.00	0	261.00	0.0119	0.0000	0.0000	P = 0.244
ND4	15	357.55	3	1019.45	0.0420	0.0029	0.0701	P = 0.000
ND4L	3	81.64	1	215.36	0.0368	0.0046	0.1263	P = 0.065
ND5	10	458.17	10	1353.83	0.0218	0.0074	0.3385	P = 0.014
ND6	3	136.03	4	388.97	0.0221	0.0103	0.4662	P = 0.262

The  $K_A/K_S$  ratios for haplogroup U5 were around 0.10, except for CO1, CO3 and ATP6, which were 0.04, 0.38 and 0.81, respectively (Table 5.4). ATP6 was remarkable in that twice as many non-synonymous mutations were present as compared to synonymous, although because there is a greater number of possible non-synonymous

sites, the Fisher's Exact Test did not show a significant difference between the frequency of each type of mutation ( $P = 0.396$ ). Only two other genes were shown to have a non-significant difference in number of synonymous and non-synonymous mutations: CO3 ( $P = 0.054$ ) and ND3 ( $P = 0.148$ ). All other genes showed significantly more synonymous mutations than non-synonymous, suggesting the action of purifying selection.

Table 5.4  $K_A/K_S$  ratios for haplogroup U5

Gene	# Syn	# Possible Syn	# NonSyn	# Possible NonSyn	$K_S$	$K_A$	$K_A/K_S$	Fisher's Exact Test
All	154	2903.92	77	8472.08	0.0530	0.0091	0.1714	$P = 0.000$
ATP6	7	178.36	16	502.64	0.0393	0.0318	0.8111	$P = 0.396$
ATP8	4	50.33	1	156.67	0.0795	0.0064	0.0803	$P = 0.013$
CO1	18	390.01	2	1151.99	0.0462	0.0017	0.0376	$P = 0.000$
CO2	13	170.00	3	514.00	0.0765	0.0058	0.0763	$P = 0.000$
CO3	7	195.05	8	587.95	0.0359	0.0136	0.3791	$P = 0.054$
CytB	20	285.56	15	854.44	0.0700	0.0176	0.2507	$P = 0.000$
ND1	16	255.33	8	698.67	0.0627	0.0115	0.1827	$P = 0.000$
ND2	16	264.70	7	776.30	0.0605	0.0090	0.1492	$P = 0.000$
ND3	2	84.00	1	261.00	0.0238	0.0038	0.1609	$P = 0.148$
ND4	16	357.33	5	1019.67	0.0448	0.0049	0.1095	$P = 0.000$
ND4L	4	81.67	1	215.33	0.0490	0.0046	0.0948	$P = 0.022$
ND5	24	458.13	8	1353.87	0.0524	0.0059	0.1128	$P = 0.000$
ND6	7	136.00	2	386.00	0.0515	0.0052	0.1007	$P = 0.002$

With the exception of a few genes in each haplogroup, the vast majority of these tests showed that purifying selection has shaped the current mutational landscape of mtDNA haplogroup variation.

#### 5.4 $d_N/d_S$ Analysis

In addition to  $K_A/K_S$  estimates, codon-based Z tests ( $d_N/d_S$ ) were used to assess the effects of selection on the protein-encoded loci. For the majority of cases, the null hypothesis (neutrality;  $d_N/d_S = 1$ ) could not be rejected when testing positive selection on the genes ( $d_N/d_S > 1$ ). To the contrary, the p-values for all tests of purifying selection



( $d_N/d_S < 1$ ) were lower than the p-values for all tests of strict neutrality. For haplogroup C, the null hypothesis was rejected in favor of the presence of purifying selection for ATP8, CO1, CO3, ND1, ND2, ND3, ND4L, and ND5 (p-value < 0.05). The null hypothesis could not be rejected for the remaining loci: ATP6, CO2, CytB, ND4, and ND6.

Tests for haplogroup D consistently show evidence of either purifying selection or strict neutrality. ATP6, CO1, CO2, CO3, CytB, ND2, ND4, ND4L, ND5 and ND6 all had significant p-values for the test of purifying selection, where as tests for strict neutrality were significant for CO1, CO3, CytB, ND1, ND2, ND4, ND4L, ND5 and ND6. Tests for positive selection had p-values of 1.0 for every gene.

For U4, significant results were obtained from only three genes, these being CO1, ND2 and ND4. CO1 and ND4 showed the presence of either purifying selection or strict neutrality. The p-values for both estimates were below the 0.05 threshold, but the p-value for the purifying selection test was lower than the test for strict neutrality. The only exception was ATP8, where the lowest p-value was for the test of positive selection. In this case, the p-value was only 0.429, which is far from being significant. This estimate was probably obtained because only five mutations were found in this gene for U4 (3 nonsynonymous and 2 synonymous). The  $K_A/K_S$  analysis did not indicate that these values were significantly different. Thus, if selection is acting on the U4 mtDNA genomes, then it is doing so through purifying selection, albeit only weakly.

Tests of  $d_N$  and  $d_S$  for U5 showed similar results. Tests indicated that ATP8, CO1, ND1, ND2, ND4 and ND5 likely were affected by purifying selection. Tests for strict neutrality were also significant for ND1, ND2 and ND4. ATP6 and ND6 had lower

p-values for tests of positive selection instead of purifying selection. While neither p-value was below the 0.05 significance level, the dN/dS p-value for ND6 was close ( $p = 0.072$ ). The p-values for the tests of strict neutrality were always higher than for tests of purifying selection (often twice as high).

Overall, these tests indicated that if selection has acted on the protein-encoded regions of these haplogroups, then it has done so through purifying selection, and not adaptive (positive) selection.

## **5.5 Tests for Clocklike Behavior**

To assess the relative branch lengths within each haplogroup, maximum likelihood trees were constructed using two models (one with and one without clocklike rates). Likelihood ratio tests were used to test the significance in likelihood estimates between the two models. Because the control region is hypervariable, less importance is placed on this portion of the genome. Much more important is the neutrality of the coding region, where positive selection has the greatest potential to have acted. If positive selection were acting on these phylogenies, then there should be differences between branch lengths from the same haplogroup. Under neutral evolution, the branch lengths should be relatively similar.

Because the entire mitochondrial genome will be used to estimate coalescence dates, entire genomes were used to create the ML trees. However, each was split into coding and control regions to investigate the effects of each mtDNA segment on the clocklike behavior of the four haplogroups. Therefore, three different data sets were used to assess the effects of the control versus coding regions on the results of the entire

genome: (1) entire genome for all haplotypes, (2) coding region only and (3) control region only. Deviation from clocklike behavior in the coding region could signal evidence for positive selection.

For the entire haplogroup C tree (all 121 haplotypes), the simpler tree (clocklike rate) was rejected. Thus, ML branch lengths were not consistent across the haplogroup. One problem with all of these trees was that the major branches were not fully resolved. As a consequence, the four primary branches of haplogroup C were not identifiable. For example, C1 and C5 are defined by single mutations in the control region (16325 and 16288, respectively), which were obviously missing in the “coding region only” trees. Also, a subbranch of C1 (C1b) is defined by a single mutation in the control region (493), causing the haplotypes for this subbranch to be positioned diffusely throughout the trees. Thus, it is not clear whether the differences between haplogroup C branches were caused by unresolved branches or the effects of positive selection.

To gain some clarity on this issue for our purposes, I ran a second set of tests that included the primary branches of haplogroup C found in southern Siberia (C4 and C5). Trees were generated as above using the entire genome, the control region and the coding region. The simpler model (clocklike rate) could not be rejected for any of these trees. Thus, the branches of haplogroup C that were found in Siberia and that are the focus in this dissertation evolved in a clocklike manner. It is possible that the rare C7 or the abundant C1 has been affected by positive selection, but these two sub-haplogroups were not included in any other analysis because they are not found in southern Siberia.

The ML trees generated from entire mtDNA genomes of haplogroup D showed that the clocklike model could be rejected, as it was for haplogroup C. The clocklike

model can also be rejected for D4 and D5, but not D6. To determine if the control or coding regions had non-clocklike behavior, each segment was assessed separately. The clocklike model could not be rejected for D5 coding region, but was for the control region. Overall, positive selection does not seem to have played a role here because the coding region appears to have evolved in a clocklike manner. It is possible that the control region has variable branch lengths because of sequence instability.

In a previous study, I identified several mutations that appear to cause mutations at additional mutations in nearby positions, similar to that identified in haplogroup T (Dulik, Gokcumen, Zhadanov, Osipova, & Schurr, 2007; Malyarchuk & Derenko, 1999). In this case, D5 haplotypes that have a mutation at 16164 appear to also have greater levels of recurrent mutation at 16092 and 16172. Malyarchuk and Derenko noted a similar phenomenon in haplogroup T, with mutations at 16292 and 16296. Additional experiments are necessary to confirm these findings, yet the recurrence of several mutations at these positions cause reticulations in the network analysis. These mutations can also make accurate ML tree construction difficult.

D4 was slightly problematic for this analysis because it has so many branches, and some of them have very small sample sizes. Therefore, to verify that the branches of the D4 phylogeny that I am interested in are evolving in a clocklike manner, only three of the defined D4 clusters were examined. These clusters were D4b1, D4m and D4o, as will be discussed in the sections below. The clocklike model could not be rejected for any of these branches, no matter which mtDNA genome region was used (complete, coding only or control only), even though a non-clocklike model was a better fit of all D4 data considered together.

To show that no differences existed between the three D4 clusters in question, I pooled all the sequences from these clusters into a single file, generated ML trees and assessed them using both a clocklike and non-clocklike model. The clocklike model could not be rejected for the ML trees using complete genome data or coding region only data. The control region data could, but this was not surprising given the high mutation rate for this region of the mtDNA. This observation could explain why the ML trees with all D4 branches appeared to exhibit non-clocklike behavior.

Tests for the clocklike evolution of haplogroup U4 phylogenies showed that the hypothesis of clocklike evolution could not be rejected, whereas that for the coding region could. Closer inspection of the coding ML trees showed a large number of unresolved branches. This problem was likely due to the importance of control region mutations for the identification of several internal branch nodes.

To explore this finding further, an additional set of analyses were run specifically for U4 branches found in Siberia. Generally, the U4 mtDNAs are not well defined in Siberia, but one study found U4a and U4c. Therefore, I generated ML trees for U4a and U4c to determine if they possessed clocklike mutation rates. All three tests for U4a (complete sequence, coding sequence only, and control sequence only) did not support clocklike mutation rates. This observation was again likely due to the number of unresolved branches in the ML trees, which were large (nearly 22%, 50%, 29% for ML trees generated from complete sequences, coding sequences and control sequences, respectively). In contrast to U4a, U4c clocklike models could not be rejected for any ML tree. The unresolved branches for these trees were much lower (as small as 3% for complete sequences, 5% for coding region sequences, but about the same for control

region sequences). Thus, it appears that poor ML tree construction resulted in the rejection of the clocklike hypothesis for the U4a branch.

For tests on the U5 mtDNA genomes, the clocklike model could be rejected at the 5% level. Further investigation of the coding and control region trees shows that those generated from only the coding region could be rejected, but the trees generated from the control region could not. Both models were assessed for U5a, and indicated that a model of clocklike evolution cannot be rejected for this branch. This finding is consistent with all three data sets (the entire genome, coding region only, and control region only). For U5b, clocklike ML trees cannot be rejected for the entire genomes and control region data sets, but it can for the coding region data set. The coding regions of each of the three U5b clusters were analyzed by generating ML trees from only two clusters at a time. The trees for each of these pairs (U5b1'2, U5b1'3 and U5b2'3) can all be rejected at the 5% significance level. When each cluster was examined individually, the clocklike model can be rejected for only U5b1.

It appears then that several U5b branches are evolving in a non-clocklike manner. Therefore, coalescent dates that rely solely on the coding region of U5b could be inaccurate. Ultimately, these tests showed that a clocklike tree was not significantly different from a non-clocklike tree when generated with the entire U5b genomes. Therefore, dates that rely on the entire genome data set should generally be accurate, while dates for U5b1 that rely solely on synonymous mutations should be viewed with caution.

It is possible that the non-clocklike evolution of U5b coding region ML trees could be attributable to positive selection. Of the U5 clusters, only U5b2 has a diagnostic

mutation that is nonsynonymous. This mutation is located in ND5. Recall, however, that the dN/dS estimates for U5 showed that ND6 and ATP6 were the most likely candidates for genes affected by positive selection. Therefore, this issue is not fully resolved, and caution is clearly needed in any interpretation involving this haplogroup branch.

To summarize, while ML trees showed evidence of non-clocklike evolution for haplogroups C, D, and U5, they did not for U4. Closer examination of Siberian branches of C and D showed that the hypothesis of clocklike behavior could not be rejected. One branch of U4 (U4a) appeared to evolve in a non-clocklike manner, but closer examination of the ML tree showed many unresolved branches. Thus, poor ML tree construction is the likely culprit for hypothesis rejection. Finally, of the two U5 branches, one (U5a) appeared to evolve in a clocklike fashion, but the other (U5b) did not. The cause of this pattern is not entirely clear, although it does not appear to be due to poor ML tree construction, like U4a. Instead, this branch was likely affected by selection, but identification of the selected locus was not determined. Therefore, with the exception of U5b, the coalescence dates of these haplogroups can be estimated accurately.

## **5.6 Coalescence Dating and Phylogeography**

Network v4.5.1.6 was used to create six different reduced median-median joining (RM-MJ) networks with different classes of mutations and segments of the mtDNA genome, as described by Soares et al. (2009). This ensured the consistency of coalescence dates from rho statistics across several different segments of the genome.<sup>8</sup>

---

<sup>8</sup> All six networks are shown for haplogroup C, but since the structure of the networks is redundant for other haplogroups, I will only provide the network using complete genome data for haplogroups D, U4 and U5.

The first network utilized all substitutions from each genome except 16519, which is extremely mutable. Length polymorphisms surrounding the 16189C mutation were ignored, as these are commonly the result of sequencing errors and therefore not phylogenetically informative (Bandelt & Parson, 2008; Bandelt et al., 2002). The second network was generated from only synonymous substitutions. These networks generally were not as well resolved as those based on complete genomes because the control region (which is not included) contains a considerable number of polymorphisms useful for differentiating sub-haplogroups.

Control region data were used exclusively to create the final four networks. The first of these included all transitions between nucleotide positions 16090 and 16365. This sequence range was used as the first HVS1 standard (Forster et al., 1996). The second control region networks includes the HVS1 region commonly used in publications today (16051-16400), although the actual HVS1 range is from 16024 to 16383 (MITOMAP, 2009). The third network used only a portion of HVS2 (68 through 263). This network was the least resolved in all cases because even though there is high mutability among HVS2 sites, these often occur at the same positions, and therefore are not as phylogenetically informative as HVS1 polymorphisms. The final network used the entire control region (except mutations at 16519 and length polymorphisms for the reasons listed above).

## **5.7 Haplogroup C – Coalescence Dating and Phylogeography**

Overall, major branches of haplogroup C in the complete genome RM-MJ network were well defined (Figure 5.2). C4 and C6 branches were still identifiable with



the “synonymous mutations only” network, but the roots of C1 and C5 were not because each of these branches is defined by polymorphisms at positions outside the protein-encoded regions (Figure 5.3). Regardless, they could be differentiated based on derived mutations even though membership in each branch was not as obvious as in the first network.

Branches of haplogroup C in the four control region networks were not always readily identifiable. Multiple sequences that belong to different branches may actually be combined in these last four networks, making coalescent dating for branches within these networks less reliable. These dates are also less reliable because recurrent mutations occur in this region of the mtDNA genome more often, and therefore, reticulations occurred more often, as expected (Heyer et al., 2001; Stoneking, 2000; Yang, 1996).

The coalescent dates for haplogroup C obtained from the six networks were compared to ensure the consistency of estimates based on different portions of the mtDNA genome (Table 5.5). Haplogroup C coalesced to an MRCA sometime between about 19 and 33 kya. The estimate obtained from the entire genome had the smallest 95% confidence interval and was centered at about 26 kya. The dates calculated from all synonymous mutations and the entire control region agreed with the first estimate, but had slightly larger confidence intervals. The date obtained from all synonymous mutations was slightly lower than the one obtained from the entire genome, but this was expected if some portion of the nonsynonymous mutations (likely, those from the internal branches of the network) were also neutral.

Estimates obtained exclusively from either HVS1 or HVS2 were not precise. Both HVS1 dates were centered at about 35 kya (above the upper limit of the 95%

confidence interval for the coding region estimates), but their 95% confidence intervals were extremely large [19 – 50 kya]. HVS2 provided a date of 11 kya [18 – 37 kya]. The lower dates for HVS2 were also expected because that section of the mtDNA genome – although hypervariable – does not provide sufficient phylogenetic resolution on its own. Many of the mutations in that section are recurrent, thereby causing it to lose much of its phylogenetic utility. When HVS1 and 2 were included together, however, the estimates were congruent with those from the coding region.

Table 5.5 Coalescent estimates for haplogroup C

<b>Networks</b>	<b>Rho</b>	<b>Sigma</b>	<b>KYA</b>	<b>95% C.I.</b>
Complete Sequences	9.5372	1.0734	26.4	20.3 – 32.7
Synonymous Mutations Only	3.0413	0.6051	23.9	14.6 – 33.3
Entire Control Region	3.0331	0.5251	27.5	18.2 – 36.8
HVS1: 16051-16400	2.0661	0.4604	34.5	19.4 – 49.5
HVS1: Transitions between 16090-16365	1.8512	0.4564	34.9	18.0 – 51.7
HVS2: 68-263	0.5041	0.1790	11.3	3.4 – 36.8

The entire mtDNA genomes were especially useful for answering phylogeographic questions concerning haplogroup C. Sequence variation in the C4 branch dates to around 23 kya, not much different from the entire C phylogeny (Table 5.6). The 95% CI falls between 16 and 28.6 kya when considering all substitutions and 11.4 to 36.2 kya when only considering synonymous substitutions. Calculations were not attempted with the control region data because not all C4 haplotypes could be readily identified based on control region mutations alone. Moreover, some haplotypes in the C4 and C6 branches were identical. Consequently, the coalescence estimates for the entire phylogeny overlapped when comparing networks generated from control region,

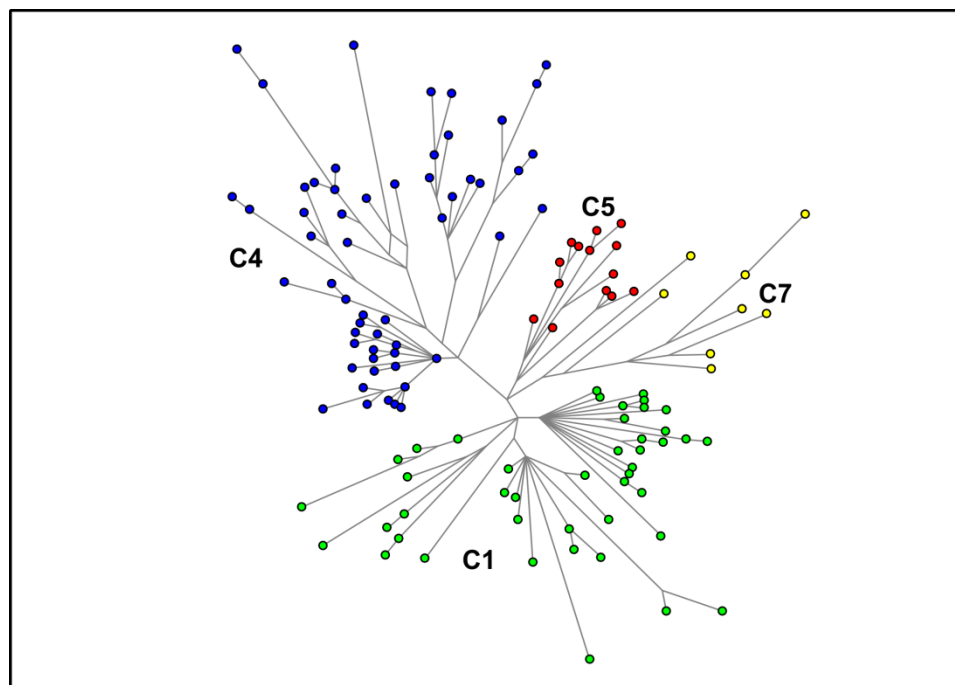


Figure 5.2 RM-MJ network of complete haplogroup C mtDNA genomes

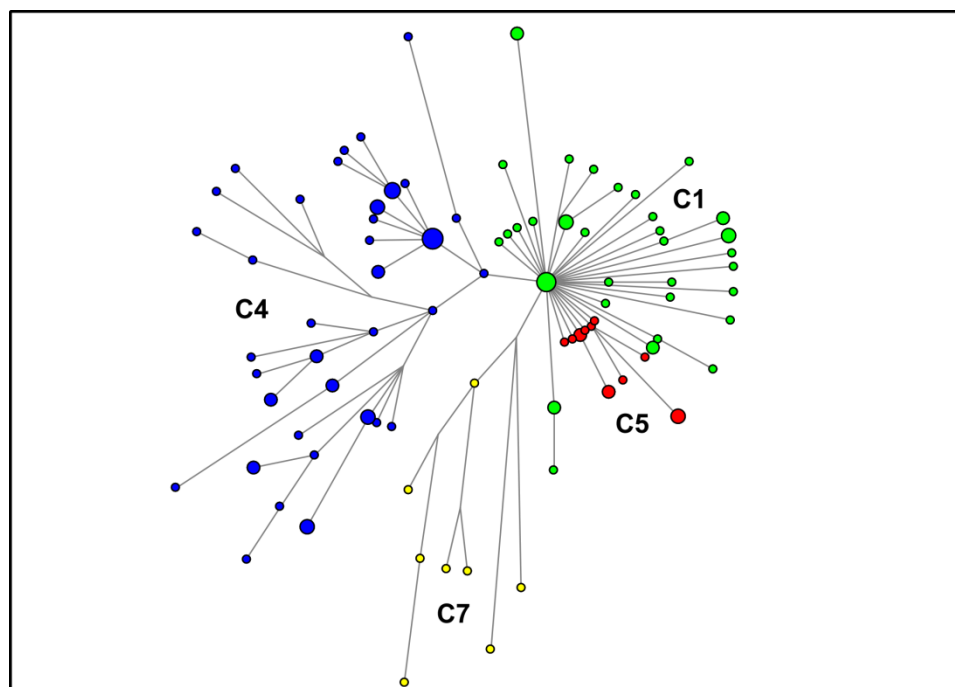


Figure 5.3 RM-MJ network of only haplogroup C synonymous mutations

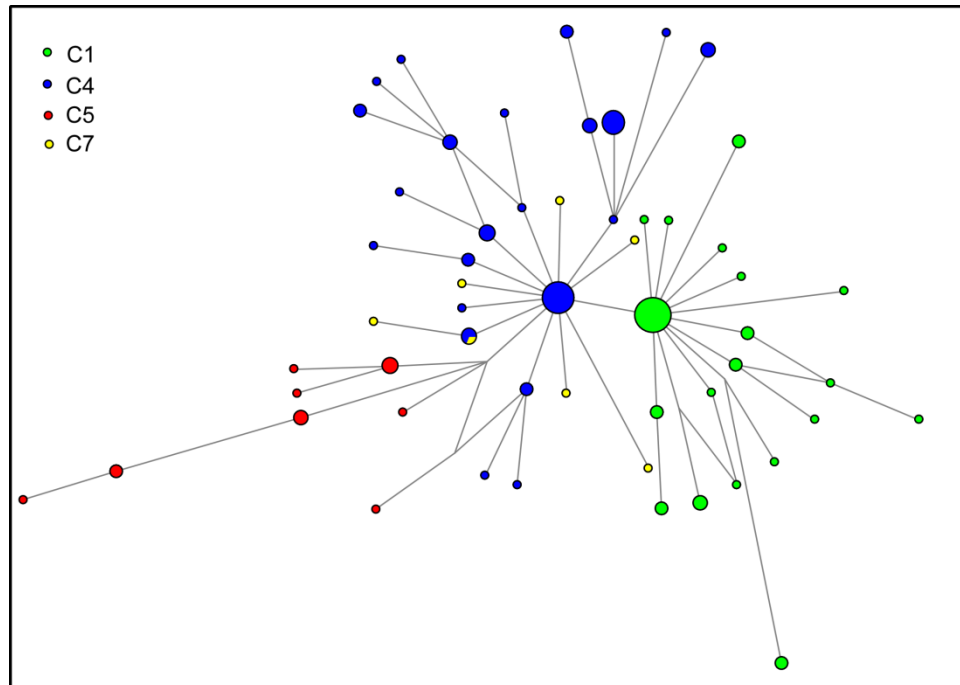


Figure 5.4 RM-MJ network of haplogroup C HVS1 from 16090-16365

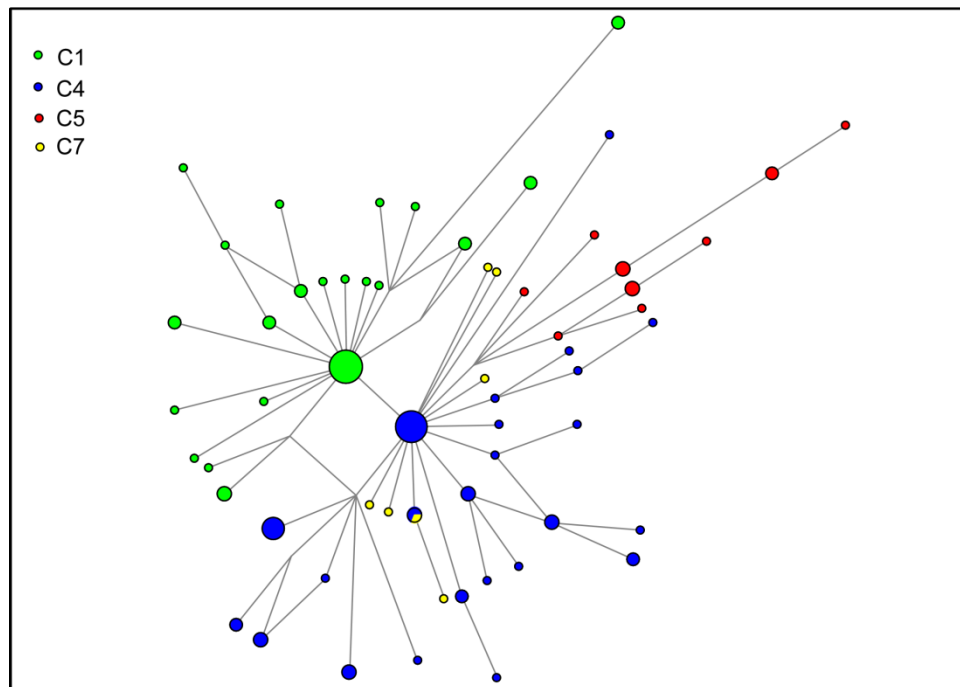


Figure 5.5 RM-MJ network of haplogroup C HVS1 from 16051-16400

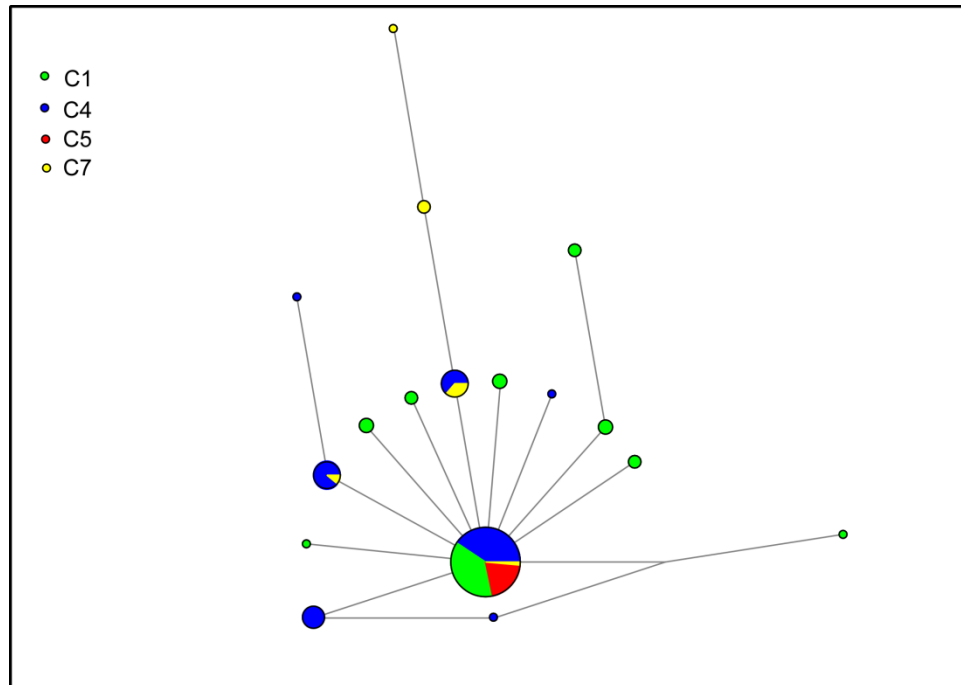


Figure 5.6 RM-MJ network of haplogroup C HVS2 from 68-263

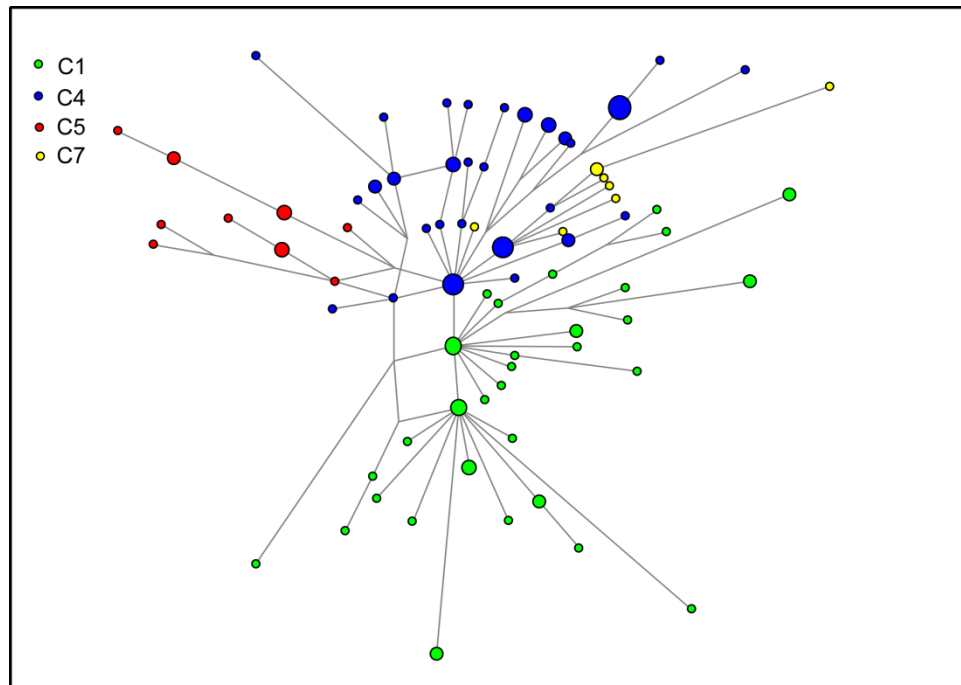


Figure 5.7 RM-MJ network of haplogroup C complete control regions

synonymous mutations and all polymorphisms.

Table 5.6 Coalescent estimates of haplogroup C branches

<b>Networks</b>	<b>Rho</b>	<b>Sigma</b>	<b>KYA</b>	<b>95% C.I.</b>
All Branches	9.5372	1.0734	26.4	20.3 – 32.7
C1	6.8409	0.8109	18.6	14.1 – 23.1
C4	8.1091	1.1095	22.2	16.0 – 28.6
C5	6.3571	1.2392	17.2	10.4 – 24.2
C7	9.8750	1.9244	27.4	16.5 – 38.8

Because C4 is roughly the same age as the entire C phylogeny, it is not surprising that this branch is the most widespread in Asia. It was found in populations from East Asia, Siberia, Central Asia, and even the Americas (Derbeneva, Starikovskaia et al., 2002; Derenko et al., 2003; Derenko, Malyarchuk, Grzybowski et al., 2007; Jin, Kim, & Kim, 2010; Naumova, Khaiat, & Rychkov, 2009; Naumova et al., 2008; Pakendorf, Novgorodov, Osakovskij, & Stoneking, 2007; Pimenoff et al., 2008; Schurr et al., 1999; Starikovskaya et al., 2005; Starikovskaya et al., 1998; Torroni, Sukernik et al., 1993; Volod'ko et al., 2009; Volodko et al., 2008). C4 is defined by an insertion in the 16S rRNA gene (2232.1A), one synonymous mutation (6026) and two non-synonymous mutations (11969 and 15204). The first nonsynonymous change occurred in the ND4 gene and resulted in an alanine to threonine change. The second nonsynonymous change occurred in CytB and resulted in an isoleucine to threonine change. Both of these mutations cause neutral non-polar to neutral polar amino acid changes.

C4 is made up of three primary clusters: C4a, C4b and C4c. C4c was found only in the Americas, being identified by two complete sequences (Malhi et al., 2010; Perego

et al., 2009). The coalescent date for this branch was around 13 kya, but the 95% CI was extremely large (5 to 22.1 kya) because only two haplotypes have been identified to date.

C4a and C4b are more abundant than C4c. C4a is designated by a synonymous mutation at 12672. There are two branches of C4a – C4a1 and C4a2 – both of which are defined by five mutations each. Control region mutations differentiate these branches, with C4a1 possessing a mutation at 16129 and C4a2 having mutations at 16171, 16344 and 16357. C4a (and both of its branches) dated to roughly the same TMRCA as the entire haplogroup.

Based on the network of complete haplogroup C genomes, one cluster within C4a1 is defined by a synonymous mutation at 15607. The full genomes that constituted this cluster came from northern and southern Siberia (Tubalar, Tofalar, Evenk, Nganasan and Yukaghir). The TMRCA for this cluster was 8.4 kya [2.9 – 14.1 kya]. What is notable about this cluster is the large number of southern Siberian and Central Asian samples that have the same or very similar HVS1 motifs (16093-16129-16223-16298-16327). Its definition is problematic because both diagnostic mutations (16093 and 16129) undergo recurrent mutation more often than other HVS1 sites, thus making unambiguous identification of C4a1 difficult from control region data alone. Therefore, more complete genomes are needed to confirm the presence of this cluster in Central Asia.

The remainder of the C4a1 sequences fell into another cluster that is defined by a control region mutation at 16150. Like the previous C4a1 cluster, this one encompasses both southern Siberians and Central Asians. However, only a single complete genome represented this branch. Because the HVS1 motif is so distinctive, it was possible to use

the network of control region mutations to identify and estimate the coalescence of this cluster. Its TMRCA estimate was 3.2 kya [0.5 – 6.0 kya], which is slightly younger than that of the other C4a1 cluster, although it had an equally large distribution. Additional complete genomes for this cluster are needed (especially from Central Asian populations) to determine a more precise TMRCA for it.

The relatively recent dates for these two clusters indicate that C4a1 mtDNAs likely originated in the Neolithic and/or Bronze Ages. Based on ancient DNA studies of Siberian and Mongolian human remains, the C4a1 cluster with 16129 first appears in a Pazyryk kurgan from the southern Altai (Voevoda et al., 1998). It was also found in burials attributed to the Xiongnu in Mongolia, the Tashyk culture in Minusinsk and the Iron Age Tuoba Xianbei (also excavated in Mongolia) (Changchun, Li, Xiaolei, Hui, & Hong, 2006; Keyser-Tracqui, Crubezy, & Ludes, 2003; Keyser et al., 2009), but has not been found in the Tarim Basin, nor among the Neolithic Baikal inhabitants (C. Li et al., 2010; Zhang et al., 2010). These results indicate that the C4a1 haplotypes may have expanded west toward southern Siberia after the adoption of pastoral nomadism, although the C4a1 cluster defined by 16150 was not found in the ancient DNA samples.

C4a2, the sister branch of C4a1, had a coalescence date of 12.2 kya [5.2 – 19.5 kya]. The current distribution for C4a2 is decipherable, given the unique mutations in the control region of these mtDNAs. These lineages were found primarily in southern Siberian populations, but had also been identified in central and northeastern Siberian populations (Yakut, Evenk, Even, Yukaghir), northern Siberia (Khanty, Nganasan and Ket), Central Asia (Kazakh, Uzbek, Kara-kalpak, Kyrgyz) and Mongolia and northern China (Mongolians, Oroqen, Ewenki, Kalmyk, Buryat) (Comas et al., 1998; Comas et



al., 2004; Derbeneva, Starikovskaia et al., 2002; Derenko et al., 2003; Pimenoff et al., 2008; Volodko et al., 2008; Yao, Kong, Wang, Zhu, & Zhang, 2004; Yao & Zhang, 2002). It was also found in a C4a2 single Xiongnu individual, but was not found in the 250-plus Han Chinese sampled (Yao et al., 2002). The coalescence date of this branch suggests that it arose in southern Siberia and spread to the surrounding regions, probably during the Neolithic.

C4b is defined by a synonymous mutation at 3816. Unlike C4a, the derived branches of C4b show a star-like pattern, indicating a possible recent expansion. In fact, the TMRCA for all C4b lineages was around 8.3 kya [4.8 – 11.9 kya], roughly 25-50% of the age of its sister branch (and the rest of the C phylogeny). The difference between the two C4 branches cannot be due to positive selection. Both branches are defined by synonymous mutations, and there is no evidence of positive selection in the analyses mentioned above. Thus, this discrepancy is likely the result of demographic differences in populations where these two branches reside.

The TMRCA places this origin of C4b firmly in the Neolithic. One cluster of C4b named C4b1a, defined by a synonymous mutation at 8251, was found among Tuvinian, Nganasan, Mansi, Kyrgyz, Yukaghir, Udegei and Mongolians. Among this cluster are a set of haplotypes that possess an insertion in the control region (16259.1A) and are readily identifiable among populations of southeastern Siberia (Oroqen, Ulchi), Mongolia (Mongolians) and Baikal area of Siberia (Mongolian, Buryat, Soyot, Yakut), as well as one Nganasan and a few Western Evenks (Derbeneva, Starikovskaia et al., 2002; Derenko, Malyarchuk, Grzybowski et al., 2007; Kolman et al., 1996; Starikovskaya et al., 2005). The TMRCA for this cluster was about 4.1 kya [1.1 – 7.2 kya]. While this

haplogroup probably emerged in the Bronze Age, it cannot be reliably tied to any particular culture, but could have spread to Western Evenks and Nganasans during the migrations of Yakut to their current geographic distributions in central Siberia. No ancient samples belonging to this cluster were found, like the C4a1 cluster with 16150.

C4b3 is defined by a mutation in the control region, 16291. Based on HVS1 haplotypes, this branch of C4b was found in southern Siberia (largely Tubalar, but also Altai-kizhi, Telenghit, Tuvinian, Todzhan, and Khakass), central and northern Siberia (Evenk, Yakut and particularly Yukaghir) and at low frequencies in Mongolia and the Baikal area of Siberia (Mongolian and Buryat) (Derenko, Malyarchuk, Grzybowski et al., 2007; Starikovskaya et al., 2005; Volodko et al., 2008). This cluster had a TMRCA of 5.8 kya [0.6 – 11.4 kya]. The 11,000-year range was too large of a window to attribute these to any particular time period, although the TMRCA estimate was just slightly older than C4b1a. One Xiongnu individual possessed this haplotype (Keyser-Tracqui et al., 2003).

The final C4b cluster is C4b2, which is defined by a transversion mutation in the control region (16318T). This cluster is restricted to northeastern Siberia among the Chukchi, Koryak and Siberian Eskimo (Schurr et al., 1999; Starikovskaya et al., 1998; Volodko et al., 2008). The TMRCA was 1.7 kya (95% CI = 0 – 4.1 kya). Given the specific range of this cluster, it is highly likely that it originated in northeastern Siberia relatively recently (and after peopling of New World).

One problem with the phylogeography of C4b is that many of the haplotypes lack any defining feature in the control region. Most have the root HVS1 motif for haplogroup C (16223-16298-16362), but this root is also found in the C7 branch. C7 is

less common than C4b, and was found in India, China (only one Han Chinese) and Siberia (only one Evenk) and one simply identified as “Asian” (Chandrasekar et al., 2009; Kivisild et al., 2006; Kong, Yao, Sun et al., 2003; Volodko et al., 2008). For this reason, its full range is not precisely known.

The second primary branch found in southern Siberia is C5. It had a TMRCA of 17.2 kya [10.4 – 24.2 kya]. It is defined by an insertion at 595 in the sequence encoding for tRNA phenylalanine and a substitution at 16288 in the control region. C5 is made up of three clusters, with C5a being defined by a synonymous mutation at 3591 and two control region mutations - one at 16261 and a back mutation of 16327. Haplotypes from this cluster are found throughout Siberia, northern China and Central Asia (Chaix et al., 2007; Derenko, Maliarchuk, Denisova, M. et al., 2002; Kolman et al., 1996; Kong, Yao, Liu et al., 2003; Schurr et al., 1999; Starikovskaya et al., 2005; Yao et al., 2004) and dated to around 3.5 kya [ 0.07 – 6.9 kya].

C5b is defined by a mutation in tRNA arginine (10454) and control region mutations 16518T and 16527. This cluster was found almost exclusively in northern Altaians and dated to about 9.2 kya, but had a large 95% CI (2.3 – 16.4 kya) because it was estimated from only two complete genomes. A third genome was not included because it only consists of the coding region sequences (Herrnstadt et al., 2002). The provenience of only one of the three complete genomes is known. It comes from a Tubalar (northern Altaian) from Volodko et al. (2008). As for the other two, one is listed as Asian and the other comes from a private genetic ancestry company that submitted their data to GenBank. Further screening of our samples may narrow the confidence interval to a more precise period.

I also sequenced a C5b genome from a northern Altaian Chelkan. The genome from this sample matched the Tubalar genome, except that the Tubalar had additional mutations at 3754 and 10967. A coalescent date for only these two northern Altaian genomes was 5.2 kya [0 – 12.7 kya]. Human remains dating to the late Bronze and early Iron Age (one Tagar, one Xiongnu and one Tuoba Xianbei) have haplotypes belonging to C5, but they do not match the C5a and C5b clusters (Changchun et al., 2006; Keyser-Tracqui et al., 2003; Keyser et al., 2009).

Haplogroup C certainly originated in eastern Eurasia, but the exact location cannot be determined at this time. All populations from Siberia, Mongolia, Northern China, and Central Asia possess haplotypes from this haplogroup. It is also found in Han Chinese, but at much lower frequencies than the populations from these other regions. The logical assumption, then, is that it arose in this region. The haplogroup's coalescence dates place its origins firmly in the Late Paleolithic. Of its four branches, C4 and C5 are the most important for Siberian, Central Asian and Mongolian populations. Branches of each of these sub-haplogroups date anywhere from the Neolithic to the Iron Ages.

The earliest appearance of this haplogroup as seen in the ancient DNA data is in archaeological sites in Mongolia and the Baikal area – the eastern extent of its current distribution. Haplogroup C has not been found in the Neolithic sites in the Altai, Tuva, Minusinsk, or Kazakhstan and does not appear in this area until roughly the Bronze or Iron Ages, at which time they are also found at sites attributed to Xiongnu and Xianbei – both were nomadic groups with their origins in eastern Mongolia. However, there are currently no aDNA data from Neolithic sites in southern Siberia, where haplogroup C

could have already been present in these populations. Given the coalescence dates and distributions of modern and ancient haplogroup C mtDNAs, their presence in southern Siberia is likely due, in part, to the westward migrations of nomadic peoples during the Bronze and Iron Ages, although an indigenous origin of some is not out of the question.

## **5.8 Haplogroup D – Coalescence and Phylogeography**

Six networks were generated for the haplogroup D complete genomes, as previously described for haplogroup C (Figure 5.8). The three branches of the D phylogeny were readily apparent in the RM-MJ network generated from complete mtDNA genomes (D4, D5 and D6). The vast diversity of D4 overwhelmed the network, but identification of major branches was unambiguous. In the “synonymous mutation only” network, D5 and D6 were differentiated from the D4 clusters, but some of the defining mutations for D4 were not present, thus collapsing a number of the sub-haplogroups into the primary nodes of the network. For those D4 clusters, TMRCA estimates relied only on the entire mtDNA networks. The control region networks generally lacked distinction between the various primary and secondary branches. Therefore, these networks were only used to assess the consistency across all TMRCA estimates for the entire phylogeny.

Haplogroup D had a TMRCA of 32.3 kya [22.2 – 42.4 kya] when all six estimates were averaged together (Table 5.7). The TMRCA from the HVS2 data – 22.5 kya [8.9 - 36.1 kya] – was significantly lower than estimates generated from other networks. This pattern was consistent across haplogroups and was due to the recurrent mutations in the HVS2 region, which make it less phylogenetically important. If the HVS2 TMRCA is

removed from the average, then the TMRCA for haplogroup D increased slightly to 34.2 kya [24.9 – 43.6 kya]. This number was probably lower than previously calculated estimated because other estimates relied only on HVS1 mutations.

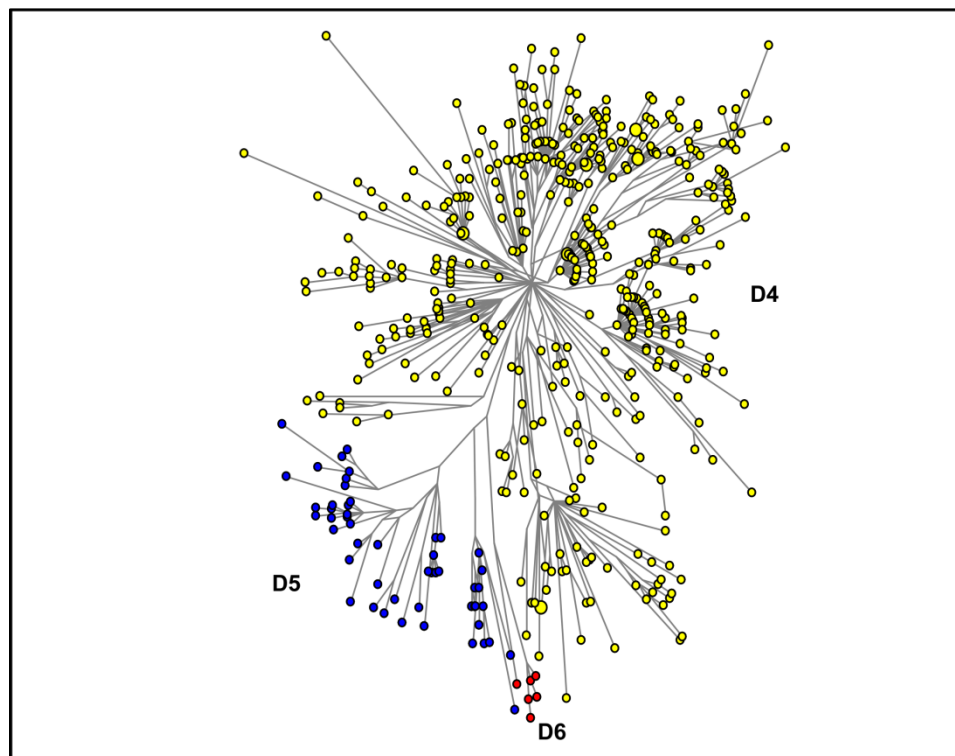


Figure 5.8 RM-MJ network of complete haplogroup D mtDNA genomes

Table 5.7 Coalescent estimates for haplogroup D

<b>Networks</b>	<b>Rho</b>	<b>Sigma</b>	<b>KYA</b>	<b>95% C.I.</b>
Complete Sequences	13.4900	1.6666	38.3	28.4 – 48.6
Synonymous Mutations Only	4.1214	0.4760	32.4	25.1 – 39.8
Entire Control Region	3.3512	0.3935	30.4	23.4 – 37.3
HVS1: 16051-16400	2.0405	0.3178	34.0	23.6 – 44.4
HVS1: Transitions between 16090-16365	1.9072	0.3263	35.9	23.9 – 48.0
HVS2: 68-263	1.0038	0.3095	22.5	8.9 – 36.1

The TMRCA for the three branches of haplogroup D were compared between the network generated from entire mtDNA genomes and that from only synonymous mutations, because these three branches could not be identified unambiguously in the control region networks. Haplogroup D4 had an average TMRCA of 28.5 kya [22.6 – 34.4 kya], D5 had a TMRCA of 33.8 kya [19.1 – 48.7 kya], and D6 had a TMRCA of 33.2 kya [18.3 – 48.4 kya]. The two estimates for D6 were quite different. The one based on the entire genomes was 24.8 kya, but the one based on “synonymous mutations only” was 41.6 kya. The entire genome estimate was more precise than the “synonymous mutation only” estimate (14.8 – 35.3 kya versus 21.8 – 61.4 kya), indicating that, in this case, the entire genome estimate was the more accurate of the two (Table 5.8).

Table 5.8 Coalescent estimates of haplogroup D branches from complete genomes

<b>Networks</b>	<b>Rho</b>	<b>Sigma</b>	<b>KYA</b>	<b>95% C.I.</b>
All Branches	13.4900	1.6666	38.3	28.4 – 48.6
D4	9.8689	0.7056	27.4	23.3 – 31.5
D5	12.2110	1.8372	34.4	23.7 – 45.6
D6	9.0000	1.7786	24.8	14.8 – 35.3

Haplogroup D mtDNAs are distributed widely and, as a result, there are many haplotype clusters. Because of this fact, I only focused on those branches which were relevant to the understanding of Altaian and (more broadly) Siberian population histories. Northern Altaians were represented by five D4 haplotypes and two D5 haplotypes. The first haplotype is the root haplotype for D (16223-16362). There are no additional mutations in HVS1 or HVS2 to help classify these samples. Therefore, complete genome analysis is necessary to place them into the proper phylogenetic context.

The second haplotype (16093-16172-16173-16215-16223-16319-16362) belongs to D4b1a2a1. The mutation at 16319 places this haplotype into D4b1. The mutations at 16093 and 16173 define D4b1a2a1. The highest frequency of this haplotype was seen in the northern Altaians, specifically in the Tubalar and Kumandin. It was also found in one Shor, and two Uzbeks (Derenko, Malyarchuk, Grzybowski et al., 2007; Yao et al., 2004). A similar haplotype with an apparent back mutation at 16093 was found in Siberian Tatar, highland Kyrgyz and Mongolians (Comas et al., 1998; Naumova et al., 2008; Yao et al., 2004). The ancestral haplotype (16093-16173-16223-16319-16362) was found in northeastern Siberia (Koryak, Chukchi, and Siberian Eskimo) and in one Kalmyk (Derenko, Malyarchuk, Grzybowski et al., 2007; Schurr et al., 1999; Volodko et al., 2008). One haplotype (16129-16173-16223-16319-16362) is specific for the Baikal area of Siberia (Buryat, Khamnigan, Mongolian populations) (Derenko, Malyarchuk, Grzybowski et al., 2007). The remaining haplotypes all occur in northeastern Siberia.

The TMRCA for this branch was 11.2 kya (4.1 – 18.5 kya). The confidence interval was wide because only five genomes from this cluster were represented in the network. Of these, only one genome represented the haplotypes that have 16172 and 16215 mutations, making it impossible to determine the TMRCA for this cluster, which was found almost exclusively in Altaians and their Central Asian neighbors. Conversely, the other cluster of haplotypes that are found largely in northeastern Siberians had four genomes and a TMRCA of 5.9 kya [2.0 – 9.8 kya], suggesting an origin in the Neolithic. No ancient mtDNAs were found to belong within this cluster.

The next haplotype (16042-16214-16223-16362) belonged to D4m2, identified by mutations at 16042 and 16214. This haplotype was well represented among the northern



Altaians and was found in all three ethnic groups. However, it was not specific to the Altai. The same haplotype was found in the Shor in southern Siberia, the Buryat and Khamnigans of the Baikal region, and the Central Even and Nivki in central and eastern Siberia (Derenko, Malyarchuk, Grzybowski et al., 2007; Pakendorf et al., 2007; Starikovskaya et al., 2005). It was also found among the Kara-kalpak, Kazakh, and Turkmen of Central Asia (Chaix et al., 2007; Derenko, Malyarchuk, Grzybowski et al., 2007; Pakendorf et al., 2007; Starikovskaya et al., 2005; Yao et al., 2004). Additional mutations were found with the root haplotype (like 16093 and 16192A), which occurred in central and northeastern Siberian populations (Derenko, Malyarchuk, Grzybowski et al., 2007; Pakendorf et al., 2007; Volodko et al., 2008). The 16295 mutation was found in two Turkmen samples, and 16042 was lost in a Shor sample (Chaix et al., 2007; Derenko, Malyarchuk, Grzybowski et al., 2007). Even though variations on the root D4m2 haplotype have been noted, they do not come near the frequency of the type found in Altaians.

The TMRCA for this cluster was calculated from only three complete genomes (10.6 kya [4.0 – 17.4 kya]). Based on the available data, it is difficult to place the origins of this branch in any historical period. It could have arisen as long ago as the LGM, or it could have emerged as recently as the Bronze Age in southern Siberia. Given its distribution, an older (probably Neolithic) origin is more likely, although gene flow could have spread these lineages more recently.

The next haplotype (16176-16183-16223-16274-16290-16319-16342-16362) belongs to D4o1. The mutation at 16290 defines the D4o cluster. The version found among the Chelkan and Tubalar of the northern Altai belong to one of the more derived

branches. This haplotype was only found among Altaians and Uzbek, while a slight variant was found in a Teleut from the southern Altai (Derenko, Malyarchuk, Grzybowski et al., 2007; Yao et al., 2004). A less derived version that lacks mutations at 16176 and 16342 was found in Buryat and Nivki populations, albeit in low frequencies (Derenko, Malyarchuk, Grzybowski et al., 2007; Starikovskaya et al., 2005).

The other D4o branch, D4o2, is defined by 16093 and sometimes 16232. These haplotypes were found mostly in southeastern Siberian populations, but also Yakut, Nganasan, Mongolian, Uzbek, Buryat, Kalmyk, Ewenki and Han Chinese (Bermisheva et al., 2005; Comas et al., 2004; Derenko, Malyarchuk, Grzybowski et al., 2007; Kolman et al., 1996; Kong, Yao, Liu et al., 2003; Pakendorf et al., 2006; Puzyrev et al., 2003; Starikovskaya et al., 2005; Volodko et al., 2008; Yao et al., 2002). The ancestral D4o haplotype (16223-16290-16362) was found only in Buryat and Mongolian populations (one person each) (Derenko, Malyarchuk, Grzybowski et al., 2007; Kolman et al., 1996).

The TMRCA for D4o and its sub-branches yielded insights into the age and dispersal of these mtDNAs. The TMRCA for D4o was 19.7 kya [10.5 – 29.4 kya], that for D4o1 was 8.4 kya [2.7 – 14.3 kya], and that for D4o2 was 6.5 kya [2.5 – 10.7 kya]. While D4o has an ancient origin, both its clusters likely began differentiating around 7,000 years ago. Both versions of D4o2 have been found in ancient Xiongnu and Tuoba Xianbei individuals (Changchun et al., 2006; Keyser-Tracqui et al., 2003). It appears that its current distribution in Mongolia and southeastern Siberia overlap with the ancient distribution with the only exceptions being the Uzbek and Nganasan. To date, the D4o1 branch has not been identified in any ancient samples.

The final distinctive D4 haplotype among northern Altaians (16082-16147A-16223-16362) lacks classification to any previously described D4 cluster. This is not to say that it does not belong to one of the defined branches, but only that complete genomes have not been generated yet for this sub-haplogroup. Such an analysis would allow for the proper placement and TMRCA estimation for this haplotype. This haplotype was exclusive to Altaians and has not been found in any ancient sample (Keyser-Tracqui et al., 2003; Keyser et al., 2009; C. Li et al., 2010; Zhang et al., 2010).

A greater variety of D4 haplotypes was found in southern Altaians, but they lack distinctive motifs that would clearly classify them as belonging to specific D4 clusters. The coalescence analysis of the haplogroup D phylogeny showed that, as a whole, haplogroup D had an ancient presence in China, Siberia and Central Asia. Even so, the clusters of D haplotypes provide a means at understanding more recent migration/gene flow events.

Most northern Altaian haplogroup D mtDNAs are shared with other Siberian or Central Asian groups. Thus, unlike haplogroup C, the different haplogroup D4 branches are rather distinctive among Altaians and show little resemblance to haplotypes appearing in ancient samples from Mongolia, southern Siberia and northwestern China. Furthermore, most of these shared types have origins in the Neolithic. Thus, it may be that these are representative of Neolithic Siberian mtDNAs.

Two very different haplotypes of haplogroup D5 were found in the northern Altaians. The first one (16092-16126-16164-16189-16223-16266-16362) belongs to D5a2a, defined by the mutations at 16092, 16164 and 16266. This haplotype (with 16126) is not common, but it was closely related to D5a haplotypes characterized from

populations throughout Siberia and China. The haplotype ancestral to this one was found in a Tuvinian and Buryat (Derenko, Malyarchuk, Grzybowski et al., 2007). Among the southern Siberians, most of the D5a2 mtDNAs were found in populations further to the east in the Baikal region (Buryat, Khamnigan), northeastern China (Daur, Ewenki, Oroqen, Han Chinese) or Mongolia (Kalmyk, Mongolian) (Derenko, Malyarchuk, Grzybowski et al., 2007; Kolman et al., 1996; Yao et al., 2002; Yao & Zhang, 2002). D5a2 was also prevalent in central Siberia, particularly among the Yakut (Pakendorf et al., 2006). The D5a branch had a TMRCA of 20.3 kya [12.9 – 27.9 kya], but D5a2 had a TMRCA of 12.6 kya [5.5 – 20.0 kya].

The final haplotype (16129-16188.1C-16193.1C-16362-16390) belongs to D5c2. MtDNAs belonging to D5c are rare worldwide. The TMRCA for the three branches of D5c are 28.3 kya [17.3 – 39.8 kya]. D5c1 (defined by 16311, 16316 and two insertions in the poly-cytosine tract) was found in two Altaian Kazakh, two Japanese and possibly one Mongolian (Gokcumen et al., 2008; Kong, Yao, Liu et al., 2003; Tanaka et al., 2004). The TMRCA for this branch was 12.7 kya [5.0 – 20.6 kya]. More importantly, the D5c2 branch had a coalescence of 6.0 kya (1.9 – 10.4 kya). This sub-lineage was found at high frequency in northern Altaian Chelkan and Tubalar, but it had also been described for one Shor and several Japanese (Derenko, Malyarchuk, Grzybowski et al., 2007; Tanaka et al., 2004; Volodko et al., 2008). It was also misidentified in a Tubalar sample (Starikovskaya et al., 2005). The high frequency of this unique haplotype in northern Altaians was probably the result of genetic drift.

Complete mtDNA sequencing of several of our D5c2 revealed no differences in haplotypes between samples or ethnic groups. The D5c2 found among Japanese,

however, had greater diversity than the Altaians. Gene flow from Japan to the Altai seems unlikely, as there is no historical record of such an event. There also were no D5c2 representatives among the populations between the Altai and Japan, making any migration between the two regions less probable. The ancestral populations that carried these haplotypes likely no longer exist, making the exact location of its origin impossible to know based only on the current available data. Nevertheless, the relatively recent coalescence of this cluster (probably Neolithic but could be as recent as Iron Age) suggests a closer genetic relationship between these regions, which previously had not been recognized. The occurrence of essentially the same haplotype in the northern Altai and Japan links these regions and is suggestive for Altaic expansions toward Japan.

Ancient D5 mtDNAs were found in Siberia, Mongolia and China. The only ancient samples belonging to D5a came from historical period Yakut (17<sup>th</sup> – 18<sup>th</sup> c. CE). Modern D5a haplotypes in southern Siberia generally have little variation, suggesting a recent introduction of these haplotypes to the region. However, one D5a haplotype was found in Russian Old Believers, some Slavs, Saami and Mansi in northwestern Siberia (Delghandi, Utsi, & Krauss, 1998; Lahermo et al., 1996; Malyarchuk, Grzybowski et al., 2002; Rubinstein et al., 2008) (Derenko et al., 2003) (Derbeneva, Starikovskaya et al., 2002). The divergence of this haplotype from the other D5a haplotypes occurred in the Paleolithic. Thus, despite this lack of ancient DNA, the widespread distribution of D5a suggests a late Paleolithic or early Neolithic origin and expansion of different branches of this sub-haplogroup.

Ancient D5b mtDNAs were found in the Tarim Basin and Mongolia (Xiongnu and Tuoba Xianbei). These mtDNAs were not found in southern Siberia. Today, the

distribution of D5b is still centered in northern China. Ancient D5c mtDNAs have not been identified yet.

Unlike haplogroup C, the phylogeography of haplogroup D is not as informative, in part, because there were not as many distinctive haplotypes among the Altaians for these mtDNAs. More complete genomes are needed for these branches, and the phylogenetic placement of several Altaian mtDNAs needs greater clarification. Still, of those haplogroup D mtDNA clusters for which coalescence estimates were generated, they showed likely Neolithic origins. These haplotypes were not found in any aDNA samples. It is possible that these lineages represent indigenous Siberian mtDNAs.

## **5.9 Haplogroup U4 – Coalescence Dating**

The four primary branches of U4 were clearly defined in the first network (Figure 5.9). One haplotype contained the primary U4 diagnostic mutations (195, 4646, 6047, 14620, 15693 and 16356), but lacked 11332. It also had private mutations, but because there was only one representative of this branch, it was not possible to determine its age or distribution. To reflect this placement, I designated this haplotype as U4\*.

The network of synonymous mutations provided enough resolution to distinguish U4a and U4b from the rest, but U4c and U4d did not form distinct clusters. Networks generated from control region substitutions also lacked any identifying characteristics of U4c and U4d. Unlike the previous two networks, the U4b cluster could not be identified. Only two clear clusters existed. The first was U4a1, defined by 16134, and the other was U4a2, defined by 310. Therefore, throughout the analysis of this haplogroup, TMRCAs of branches (like U4a) were calculated using only the complete genome data.

Overall, U4 (which includes U4\*, U4a, U4b, U4c and U4d) had a TMRCA of around 20.1 kya [13.2 – 23.8 kya] (Table 5.9). The coalescence dates estimated from the other networks all gave approximately the same TMRCA, but the 95% CI varied from about 9 kya to 30 kya, with HVS2 having the widest range (3kya – 40 kya).

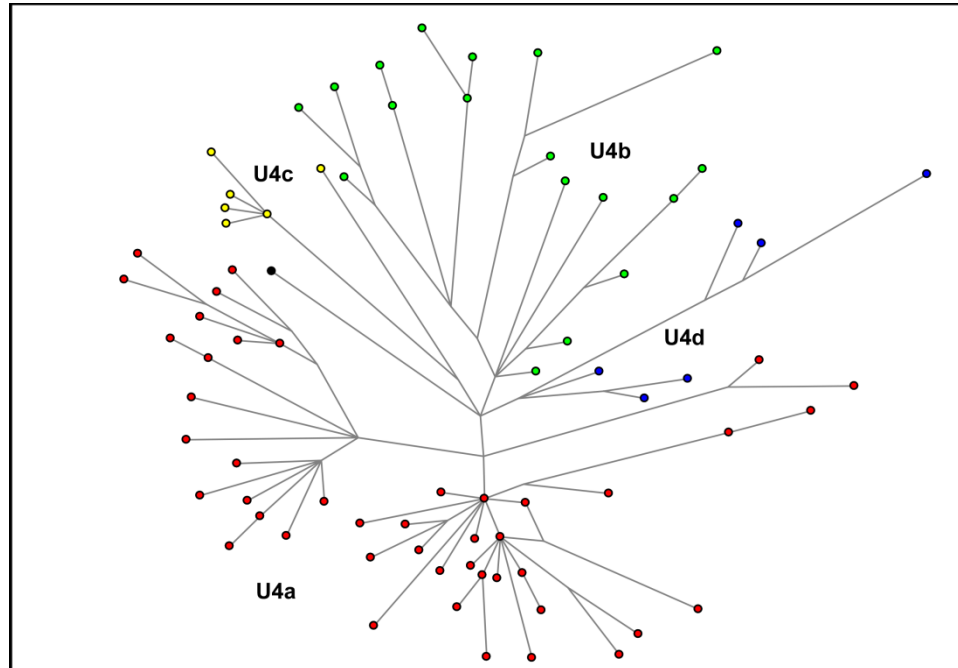


Figure 5.9 RM-MJ network of complete haplogroup U4 mtDNA genomes

The most geographically widespread of the U4 branches was U4a. This branch is defined by a synonymous mutation at 8818. It was found throughout much of Europe, with greater frequencies in northern regions and in the Volga-Ural area of Russia (Malyarchuk et al., 2008). The TMRCA for this branch was 14.6 kya [9.1 – 20.2 kya], placing its origin and expansion after the LGM (Table 5.10).

Table 5.9 Coalescent estimates for haplogroup U4

<b>Networks</b>	<b>Rho</b>	<b>Sigma</b>	<b>KYA</b>	<b>95% C.I.</b>
Complete Sequences	6.8026	0.9453	18.4	13.2 – 23.8
Synonymous Mutations Only	2.4868	0.6941	19.6	8.9 – 30.3
Entire Control Region	2.2368	0.5422	20.3	10.6 – 29.9
HVS1: 16051-16400	1.2500	0.3347	20.8	9.9 – 31.8
HVS1: Transitions between 16090-16365	1.8512	0.4581	34.9	18.0 – 51.8
HVS2: 68-263	0.9605	0.4249	21.5	2.9 – 40.1

Table 5.10 Coalescent estimates of haplogroup U4 branches

<b>Networks</b>	<b>Rho</b>	<b>Sigma</b>	<b>KYA</b>	<b>95% C.I.</b>
All Branches	6.8026	0.9453	18.4	13.2 – 23.8
U4a	5.4444	1.0196	14.6	9.1 – 20.3
U4b	6.4444	1.1083	17.4	11.4 – 23.7
U4c	6.8333	2.1148	18.5	7.1 – 30.7
U4d	5.6667	1.3944	15.2	7.7 – 23.1

Most of the U4a haplotypes fell into one of two distinct clusters. The first is U4a1, which is defined by two control region mutations (152 and 16134) and one nonsynonymous mutation in ND5 (12937). The nonsynonymous mutation replaces a methionine with a valine; both are neutral non-polar amino acids. U4a1 had a coalescence date of 11.1 kya [6.5 – 15.7 kya]. Because this cluster was easily identifiable by the 16134 polymorphism in the control region, TMRCA estimates were calculated for this cluster using networks of the full control region and HVS1. Both of those networks provided estimates comparable the entire molecule estimates at about 11.0 kya [3.3 – 17.0 kya]. The highest frequencies of this cluster were found in western Siberia (Mari, Chuvash and Ket), but also populations in Eastern Europe (Malyarchuk et al., 2008).

U4a2 is defined by a mutation in the control region at 310. The TMRCA for this cluster was younger than the U4a1 cluster, with a coalescence date of 7.9 kya [4.6 – 11.2



kya]. The other cluster, U4a3, defined by a nonsynonymous mutation at 8567 in ATP6, consisted of only three samples. The TMRCA for this cluster was 12.4 kya [3.8 – 21.5 kya].

The next branch of U4 phylogeny is U4b. It is defined by a synonymous mutation at 7705. The TMRCA for this entire branch was 17.4 kya [11.4 – 23.7 kya]. The distribution of U4b seems restricted to Eastern Europe, even though it is slightly older than U4a.

Less is known about the last two U4 branches. U4c is defined by a non-synonymous mutation at 10907 in the ND4 gene. The substitution changes a phenylalanine to a leucine; both are neutral non-polar amino acids. This branch had a TMRCA of 18.5 kya [7.1 – 30.7 kya], but had a large confidence interval because it is composed of so few haplotypes. The U4d branch is defined by a mutation at 629 in the segment encoding for tRNA phenylalanine. This branch had a TMRCA of 15.2 kya [7.7 – 23.1 kya], and like U4c, had a large confidence interval due to the few haplotypes that belong to this cluster.

### **5.10 Haplogroup U4 - Phylogeography**

The absence of diagnostic mutations or motifs in the control regions of U4 haplotypes make it impossible to unambiguously classify many of the published U4 mtDNAs into their respective context in the U4 phylogeny. This problem is evident for northern Altaian samples as well. There were essentially only three haplotypes found in Altaians, although these occurred at relatively high frequencies. These do not belong to U4a1 or U4a2, because the diagnostic mutations for those are not present (16134 and

310, respectively). Therefore, it is possible that they belong to U4d. In this regard, the diagnostic marker for U4d (629) was found in Ket and Nganasan (Derbeneva et al. 2002c) and Mansi (Derbeneva et al. 2002a). However, all of those haplotypes also had a mutation at 16189 in HVS1, which was not found in the Altaian U4 haplotypes.

One of the HVS1 motifs appearing in northern Altaian U4 mtDNAs has been noted by other researchers. This motif (16311-16356) was found in Kets, Nentsi, Mansi, Chuvash, Mari, Komi, Khakass, Teleut and Altai-kizhi (Derenko, Malyarchuk, Grzybowski et al., 2007; Malyarchuk, 2004). The rho estimates calculated from the non-southern Siberian groups in Malyarchuk (2004) was  $\rho=0.3$  with a sigma ( $\sigma$ ) of 0.18. Using the Soares et al. (2009) method, the TMRCA for this cluster was 5 kya [0 – 10.1 kya]. Another possibility is that these lineages belong to U4b1b given that they possess the HVS2 motif of 146-152-195, but the two complete genome U4b1b haplotypes lack the 16311 mutation. The coalescent date for these two haplotypes matched the TMRCA from HVS1 network (5.2 kya [0.1 – 10.5 kya]).

To attain a more precise estimate, complete mtDNA genomes are needed from these populations. In addition, the southern Siberian population samples need to be examined to place them properly into the U4 phylogeny. The current estimates suggest either an Eneolithic or a Bronze Age origin for this haplotype cluster, which was shared among the Altaian populations and Uralic speakers of northern and western Siberia.

Ancient U4 mtDNAs were identified in the Tarim Basin and in southern and western Siberia. The only exact matches between modern and ancient mtDNAs involved those U4s that only had the root mutation (16356). This haplotype was found in Andronovo burials in Krasnoyarsk, down the Yenisei River from the Altai. Another

haplotype (16356-16362) was found in the Karasuk in Khakassia. Modern populations possess derived versions of this haplotype. In particular, Siberian Tatars, Khanty and Mansi all share the 16113C-16356-16362 haplotype.

Although there are few U4 haplotypes in the Altaians, this haplogroup does make up a significant portion of the Altaian mtDNAs, particularly for the Tubalar. Ancient DNA showed that this haplogroup was present in southern Siberia since at least the Andronovo and could have been brought into the area with the migration associated with that culture. The similarity in haplotypes among the Altaians, Uralic-speakers and Kets in northwestern Siberia suggest that these populations derived at least in to some degree from the Bronze Age populations of southern Siberia.

### **5.11 Haplogroup U5 – Coalescence Dating**

Network v 4.5.1.6 was used to generate six networks, following the methods previously described. In the first network using all mutations, the major U5 branches were clearly identifiable (Figure 5.10). The structure of the U5a portion of the network was severely hampered when only synonymous mutations were used. This is because mutations that define U5a and its branches are not synonymous mutations. Similarly, the main branches of U5b are, for the most part, not defined by synonymous mutations, leaving U5b3 as the only one of the three U5b branches to be clearly defined by them. U5 networks constructed from the control region generally showed the same branching patterns as the complete genome network. As expected, the network based on only HVS2 was hardly adequate for deciphering phylogenetic signals from haplotype data.

The TMRCA for the entire U5 phylogeny was 33.9 kya [16.3 to 51.3 kya], as an average from rho estimates for all six networks (Table 5.11). The most precise estimate came from the first network containing all substitutions. This network placed the TMRCA for U5 at 27.9 kya

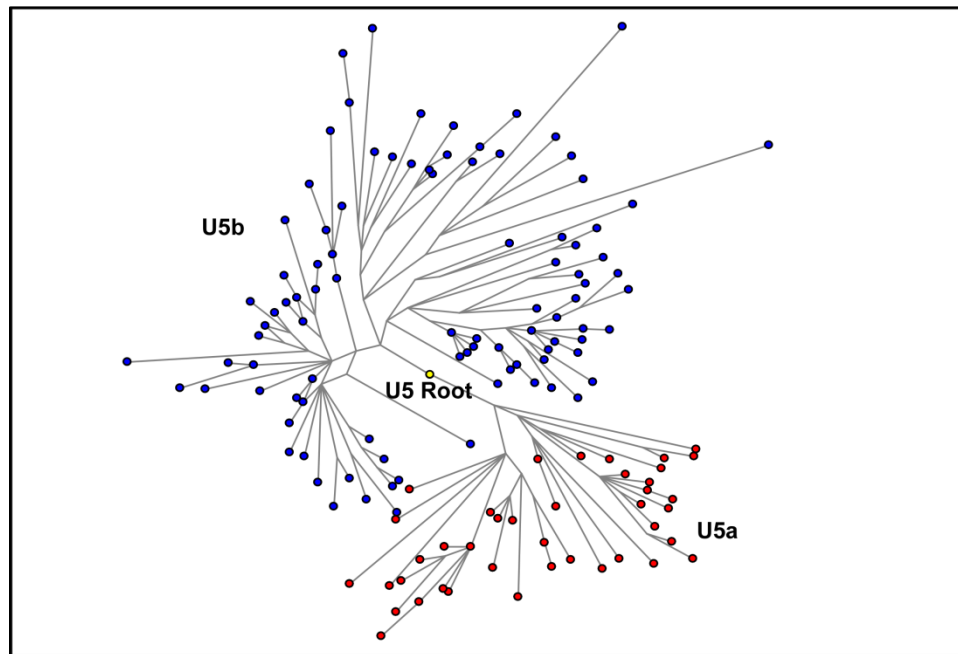


Figure 5.10 RM-MJ network of complete haplogroup U5 mtDNA genomes

[20.2 to 35.9 kya]. This estimate was similar to those obtained from the synonymous and the control region networks. Like estimates for haplogroup D, these ages were slightly more recent than those traditionally attributed to U5. The two estimates obtained from sections of HVS1 were similar and averaged to 43.2 kya [24.2 to 62.1 kya]. Because previous analyses used only HVS1 sequences, it is not surprising that these estimates were much closer to those that have been published.

U5a had a TMRCA of 19.9 kya [13.6 to 26.5 kya] (Table 5.12). Remarkably, there was no evidence that this branch evolved irregularly or that positive selection played a direct role on it, in spite of the fact that a large fraction of its diagnostic mutations were nonsynonymous. U5a comprises two major clusters – U5a1 and U5a2. These clusters were similar in age. U5a1 had a TMRCA of 14.8 kya [9.1 to 20.7 kya], while U5a2 had a TMRCA of 16.7 kya [11.0 – 22.7 kya]. Both clusters likely differentiated around the same time, just after the LGM.

Table 5.11 Coalescent estimates for haplogroup U5

<b>Networks</b>	<b>Rho</b>	<b>Sigma</b>	<b>KYA</b>	<b>95% C.I.</b>
Complete Sequences	10.0590	1.3540	27.9	20.2 – 35.9
Synonymous Mutations Only	3.7333	1.0710	29.4	12.9 – 45.9
Entire Control Region	3.6535	0.6262	33.1	22.0 – 44.2
HVS1: 16051-16400	2.5588	0.5559	42.7	24.5 – 60.8
HVS1: Transitions between 16090-16365	2.3150	0.5365	43.6	23.8 – 63.4
HVS2: 68-263	1.1630	0.6262	26.0	0 – 53.5

U5b had a TMRCA of 21.7 kya [16.2 – 27.4 kya], which was similar to that for U5a. Even though the U5b coding region did not appear to evolve in a clocklike manner, the TMRCA estimated from only synonymous mutations was similar to the date calculated from complete genomes (19.2 kya [16.2 – 27.4 kya]). Still, these estimates should be considered cautiously. The three U5b branches had different TMRCA, possibly accounting for the non-clocklike characteristics of the ML trees. The TMRCA for U5b1 was 19.3 kya [11.0 – 28.9 kya], that for U5b2 was 24.4 kya [17.6 – 31.4 kya], and that for U5b3 was 14.2 kya [8.7 – 19.9 kya]. ML tests for clocklike evolution showed that within the U5b3 cluster, the clocklike model cannot be rejected. These dates

suggested that U5b1 and U5b2 likely originated during the LGM, while U5b3 occurred afterwards. This interpretation was supported by analyses using additional complete genome sequences not deposited in GenBank (Malyarchuk, Derenko, Grzybowski et al., 2010).

Table 5.12 Coalescent estimates of haplogroup U5 branches

<b>Networks</b>	<b>Rho</b>	<b>Sigma</b>	<b>KYA</b>	<b>95% C.I.</b>
All Branches	10.0590	1.3540	27.9	20.2 – 35.9
U5a	7.3171	1.1435	19.9	13.6 – 26.5
U5b	7.9468	0.9806	21.7	16.2 – 27.4

### 5.12 Haplogroup U5 - Phylogeography

Recently, two papers explored U5 phylogeography through complete genome sequences (Malyarchuk, Derenko, Grzybowski et al., 2010; Pala et al., 2009). Originally, when haplogroup U (and U5 in particular) was identified, the common belief was that these haplotypes came from the first modern human inhabitants of Europe (Richards et al., 2000). However, the distributions and coalescence dates of haplotypes from complete genomes suggest that U5 originated in the Franco-Cantabrian refugium during the LGM and subsequently spread throughout southern and central Europe (Malyarchuk, Derenko, Grzybowski et al., 2010; Pala et al., 2009). U5a and U5b can be found throughout Europe, albeit often only in low frequencies. The furthest eastern extent of the U5 range is Siberia. There, it is found in southern (Altaians, Tuvinians and Buryats) and northwest (Khanty and Mansi) Siberia (Derenko et al., 2003; Derenko, Malyarchuk, Grzybowski et al., 2007; Pimenoff et al., 2008; Starikovskaya et al., 2005).

Both U5a and U5b were present among Altaians. U5a was represented by two haplotypes (16192-16241-16256-16270-16287-16304-16325-16399 and 16192-16256-

16270-16319-16320-16399). The second haplotype was found only found in one Tubalar, while the other was found in all three northern Altaian ethnic groups and Altai-kizhi. Both haplotypes belong to U5a1a, since the 16399 and 16192 mutations are present, although they lack 15924. These data suggest that they belong to U5a1a1b. Complete genome sequences are needed to verify their placement in the phylogeny, as 16192 recurs in this haplogroup. Since neither of these branches can be accurately placed, all that can be stated about them is that they belong to U5a1, whose TMRCA is around 15 kya.

U5b is represented by only one haplotype in several Altai-kizhi (16192-16249-16311). The back mutation at 16270 places this haplotype in U5b2a1, and the mutation at 16311 puts it within the U5b2a1a1 cluster. If the dating can be trusted for this cluster, then the TMRCA is about 4.6 kya [0.7 – 8.5 kya], well within the Neolithic.

None of the haplotypes found in Altaian populations were identified in the ancient samples from southern Siberia, Mongolia or northern China. Root haplotypes for U5a (16256-16270) were identified in the Neolithic and Bronze Age Baikal samples (Mooder, Schurr, Bamforth, Bazaliiski, & Savel'ev, 2006), indicating an early presence of U5 mtDNAs in southeastern Siberia. Therefore, if U5 arose during the LGM in a refugium, then it spread quickly to Siberia, because it was found in the Baikal region at roughly the sixth millennium BP.

### **5.13 Chapter Conclusions**

Phylogenetic analysis of these specific mtDNA haplogroups was possible after determining that natural selection on these mtDNAs was largely effective in the form of

purifying selection. Coalescence dating was carried out for these haplogroups to determine how long ago they originated. By comparing the distributions of mtDNA haplotypes over time, it is possible to gain a better understanding of when haplogroups or haplotype clusters arose in or entered southern Siberia. One major limitation of this effort is not having complete genome sequences for all comparable data sets. While HVS1 data are informative, they do not always provide the level of resolution necessary to disentangle the membership of each sub-haplogroup.

East Eurasian haplogroups tended to be more informative than West Eurasian ones. This occurred in part because there was greater diversity in East Eurasian haplogroups. Haplogroup C, for example, was not found in any burials in southern Siberia until the Pazyryk, Tagar and Tashyk periods. These periods are associated with the Scytho-Siberian cultures and later the Xiongnu. Based on the cranial morphological analysis from this period, southern Siberian populations had greater affinities with peoples to the east. In fact, haplogroup C was found in eastern Siberia in Neolithic and later Xiongnu and Tuoba Xianbei burials. Thus, one could argue that many of the haplogroup C mtDNAs arrived in southern Siberia around this time.

Haplogroup D provided less information with regard to when particular haplotypes arrived in southern Siberia. Many of the haplogroup D lineages had origins dating to the Neolithic and/or Bronze Age, and likely represent indigenous Siberian mtDNAs because they were not found outside of this region.

West Eurasian haplogroup U5 provided little information as to when this haplogroup appeared in Siberia. The branches of U5 and U4 found in southern Siberia, likely have an origin in the late Paleolithic of West Eurasia. The higher frequencies and



greater diversity in West Eurasia supports the view that U5 was certainly in Siberia by the Neolithic, as evidenced from the Baikal aDNA studies. One U4 cluster that is shared among southern and northwestern Siberian populations likely has an origin around 3000 BCE (Malyarchuk, 2004), and was likely spread among the Yeniseian and Samoyedic-speaking populations living along the Yenisei River, of which later descendants included the Ket, Khanty and Mansi and Shor, Northern Altaians.

Based on the phylogeographic analysis presented here, it appears that the current mitochondrial gene pools of Altaians were shaped by their complex history. Some of the C and D mtDNAs found in those populations likely were indigenous during the Neolithic. Their appearance in New World populations also suggests an ancient presence of these haplogroups in Siberia. The presence of at least one West Eurasian haplogroup likely reflects the migrations of people from the steppe into southern Siberia during the Eneolithic and Bronze Ages. Another series of haplogroup C mtDNAs appeared to have arrived from Mongolia/eastern Siberia during the Bronze and Iron Ages as nomadic peoples continuously moved westward.

Similarities in the southern Siberian and Central Asian mitochondrial gene pools could have persisted throughout history. There is ample evidence to suggest that southern and western Siberia interacted with the steppe as far back as the Neolithic. Furthermore, the political expansions of the Turkic and Kyrgyz Khanates would have helped to redistribute lineages across the region (Gokcumen et al., 2008). The impact of the Mongol expansions in the 12<sup>th</sup> through 14<sup>th</sup> centuries was certainly felt by Altaians. The greater similarities in haplotypes between the Altai-kizhi and Mongolian populations

provide support for this hypothesis, although this historical event occurred too recently to see much of a genetic signature in the mtDNAs of these populations.

One ambiguity that is important to note is the high frequency of haplogroups F1a, F1b, F2 and N9a in the Chelkan. These haplogroups likely originated in China, but all of their coalescence times date to around 15 kya, with 95% confidence intervals between 8.8 and 22 kya. Therefore, these mtDNAs were probably present in southern Siberia before the Eneolithic Period. Because they were not found in the Americas, we can place their arrival after the first inhabitants of the New World left Siberia.

Thus, the earliest Siberian mtDNA gene pool likely included C and D (and A). Later, after the LGM, Siberian groups received more mtDNAs from the south, including F-derived and N9 haplogroups. The N9 mtDNAs later differentiated into N9a in southern Siberia and Central Asia, while Y became prevalent in eastern Siberia. One haplogroup that remains problematic is haplogroup B. Very little variation in this haplogroup has been found in Siberia. It also is not widespread, only being found in southern Siberia. Thus, if haplogroup B was part of the ancient Siberian gene pool, then it was likely lost and then reintroduced only recently, possibly from Mongol expansions.

Haplogroups from West Eurasia were also present in the ancient Siberian Paleolithic. Certainly, haplogroup X was present, as it has also been found in North America. Without aDNA data, however, it is not possible to infer how many lineages may have become extinct between the initial colonization of Siberia (and the Americas) and today. U5 was also likely present in Siberia relatively early, but if this haplogroup expanded from a glacial refugium, then it must have appeared in Siberia during a second

wave of migration, possibly around the same time that N9 and F moved north into the region.

Even as long ago as the Paleolithic, it seems southern Siberia has played a crucial role as a crossroads of sorts, with its indigenous populations being influenced by groups from the steppe and from the south in northern China. It is only by examining the phylogeography of these lineages that such a genetic history can be examined. It is remarkable that events occurring thousands of years ago still leave their marks upon a region that has long forgotten them.

## Chapter 6: NRY Variation and Population Histories

The Y-chromosome analysis of human populations involves the use of a series of genetic markers that are valuable for making inferences about the relationships between populations and the way in which they evolve through time. As such, these data are necessary for uncovering the Y-chromosome phylogeny. By knowing the phylogenetic position and haplogroup frequencies of all Y-chromosomes in a population, the NRY data are invaluable for revealing aspects of population history and migrations. It also provides the opportunity to infer when these (prehistoric and historic) events occurred by examining paternal lineages within populations (Jobling et al., 2004).

These genetic markers can be used to produce gene trees that show the phylogenetic relationships among haplogroups and lineages (Hammer, 1995; Hammer et al., 2001; Hammer et al., 1997; Hammer & Zegura, 1996; Underhill et al., 1997; Underhill & Kivisild, 2007; Underhill et al., 2001; Underhill et al., 2000). The paternal haplogroups are defined by a series of unique event polymorphisms (UEPs), mostly in the form of indels (insertions or deletions of nucleotides) and single nucleotide polymorphisms (SNPs), where one nucleotide (A, G, T, or C) is substituted for another. A second type of mutation defines the haplotype of a Y-chromosome, this being short tandem repeats (STRs). STRs are sets of 2-5 nucleotides that are repeated over and over at specific locations in the DNA sequence, and usually change by the addition or subtraction of a single repeat run (Moran, 1975; Ohta & Kimura, 1973). Because the mechanisms for creating these two classes of mutation are different, the rates at which they occur are also different. Given that the UEPs are just that, unique, the rate at which they occur is much slower than that of STRs (de Knijff, 2000). Therefore, SNP data

provide the backbone (or trunk) of the phylogenetic tree, and STR data compose the twigs at the ends of the branches. The combination of haplogroup (SNP) and haplotype (STR) data defines a Y chromosome lineage.

To explore the paternal genetic ancestry among Altaian populations, I characterized 85 biallelic markers found on the NRY in 308 Altaian individuals. This analysis resulted in samples being assigned to one of 20 distinct haplogroups (Table 6.1). NRY nomenclature followed the conventions of the YCC, and used the current version of the NRY phylogeny (ISOGG, 2010; Karafet et al., 2008; Y Chromosome Consortium, 2002).

Table 6.1 High-resolution haplogroup frequencies of Altaian populations

Haplogroup	Chelkan	Kumandin	Tubalar	Altai-kizhi	Altaian Kazakh
C3*				19 (0.158)	24 (0.202)
C3c1				5 (0.042)	47 (0.395)
D3a				6 (0.050)	
E1b1b1c			1 (0.037)		
G1					4 (0.034)
G2a					2 (0.017)
I2a			1 (0.037)		
J2a				3 (0.025)	5 (0.042)
L	1 (0.040)				
N1*		1 (0.059)	3 (0.111)		
N1b*	5 (0.200)	8 (0.471)		2 (0.017)	
N1c*				1 (0.008)	
N1c1				2 (0.017)	
O3a3c*				1 (0.008)	31 (0.261)
O3a3c1			1 (0.037)	1 (0.008)	
Q1a2			1 (0.037)		
Q1a3*	15 (0.600)		10 (0.370)	20 (0.167)	1 (0.008)
R1a1a*	4 (0.160)	2 (0.118)	10 (0.370)	60 (0.500)	1 (0.008)
R1b1b1		6 (0.353)			3 (0.025)
T					1 (0.008)
<b>Total</b>	<b>25</b>	<b>17</b>	<b>27</b>	<b>120</b>	<b>119</b>

Samples were categorized using ethnic group identification that was self-reported at the time of sample collection, and verified through genealogies of the past 2 to 3 generations for each participant. Individuals related through paternal lineages were

removed from the analysis. As seen in the mtDNA analysis, the northern Altaian group comprises samples from the Chelkan, Kumandin and Tubalar ethnic groups. The Altai-kizhi is the only southern Altaian ethnic group represented in our sample set. In addition, I characterized the Y-chromosomes from Kazakhs living in the Altai Republic. While they are not indigenous to the Altai per se, they do make up a significant portion of the current Y-chromosome genetic diversity in the southern Altai.

### **6.1 Northern Versus Southern Altaian NRY Variation: Haplogroup Diversity**

The NRY variation showed similarities among indigenous Altaian populations. In general, Altaians were characterized by having a large proportion of Y-chromosomes that derived from the P-M45 haplogroup. This haplogroup was estimated to be approximately 34.0 kya (26.6 - 41.4 kya) (Karafet et al., 2008), which roughly corresponds to the Middle Paleolithic and the initial modern human presence in southern Siberia (Goebel, 1999; Vasilev et al., 2002). Both northern and southern groups also shared high frequencies of a derived branch of haplogroup P, namely, R1a1a\* (23.2% and 50%, respectively), but this is where their similarities end.

Marked differences in haplogroup composition and frequency were noted among indigenous northern and southern Altaian and Altaian Kazakh populations. Northern Altaian populations were largely composed of haplogroups N1b\* (18.8%) and Q1a3\* (36.2%). Q1a3\* was also present in southern Altaian populations but at a much lower frequency (16.7%). In stark contrast, south Altaians had a high proportion of Y-chromosomes that lack derived alleles at the M89 locus. These haplogroups (C3\*, C3c1 and D3a) were absent in northern Altaian groups.

N1b\* is defined by the SNP marker P43. An additional marker, P63, designates N1b1. This marker is characterized by an insertion of a C at nucleotide 566 in the PCR amplicon (Karafet et al., 2008). The Altaian N1b derived samples were tested for the P63 marker, but did not have it, instead exhibiting a C to T transition at nucleotide 567. The most likely scenario for this discrepancy is simply a mistake in the information provided by the publication (Karafet et al., 2008). Until this issue is resolved, these samples will be listed as N1b\*. If, in fact, this discrepancy is determined to be a typographic error, then all the samples listed as N1b\* in the dissertation should be renamed as N1b1. Ultimately, this problem does not affect the statistical analyses or comparisons, as P63 was not tested in any of the samples from other publications.

Altaian Kazakhs showed a completely different haplogroup profile from indigenous Altaians. The majority of Kazakh Y-chromosomes fell into one of three haplogroups, with over half of the Y-chromosomes belonging to C3\* (20.2%) and C3c1 (39.5%) and over one quarter belonging to O3a3c\* (26.1%). Of the Altai inhabitants characterized here, haplogroups G1, G2a and T were found exclusively in Kazakhs. J2a was found in both Kazakhs and Altai-kizhi, and R1b1b1 was found in both Kumandins and Kazakhs. All of these populations lacked significant numbers of haplogroups E and I (which are typical in Western European populations) and H and L (found in South Asia) (Hammer et al., 2001; Kivisild et al., 2003; Quintana-Murci, Krausz, Zerjal et al., 2001; Rosser et al., 2000; Wells et al., 2001). Haplogroup O, which predominates in East Asian populations (Xue et al., 2006), was nearly absent in indigenous Altaians, but was common in Altaian Kazakhs. The frequency differences between the northern and

southern Altaians were statistically significant ( $p = 0.000$ ), as were the differences between Kazakhs and indigenous Altaians ( $p = 0.000$ ).

## **6.2 Northern Altaian NRY Haplogroup Variation**

The haplogroup profiles found in northern Altaian populations suggest that the genetic diversity in these groups was structured, seemingly along ethnic boundaries. The Chelkan Y-chromosomes belonged mostly to three haplogroups, with 96% of the samples falling into N1b\*, Q1a3\*, or R1a1a\*. Tubalars and Chelkan both had Q1a3\* and R1a1a\*, but at different frequencies. The Chelkan possessed many more Q1a3\* than Tubalar (60% versus 37%), but the Tubalar had a considerably higher number of R1a1a\* (37% versus 16%). Tubalars lacked N1b\* entirely, but instead had N1\*, a precursor to N1b Y-chromosomes. From our sample set, the Kumandins had the most disparate haplogroup frequencies of the northern Altaians. It had a similar number of N1b\* chromosomes as the Chelkan (eight Kumandins versus five Chelkan), but only two R1a1a\* Y-chromosomes and a large portion of R1b1b1 (35.3%), while also lacking Q1a3\* entirely.

## **6.3 Haplotype-Sharing Among Altaians**

Out of the entire data set (including indigenous Altaians and men with non-indigenous paternal ancestors), 178 haplotypes were identified in 335 men. Of these 178 haplotypes, only one was shared between a northern and southern Altaian ethnic group (Figure 6.1). This haplotype belongs to haplogroup O3a3c1\* and was found in one Tubalar and one Altai-kizhi. Although these samples were collected from different



villages, they were exact matches and the only two O3a3c1\* lineages in the entire data set. For these reasons, it is likely that they shared a common ancestor in the recent past and represent recent additions to their respective ethnic groups.

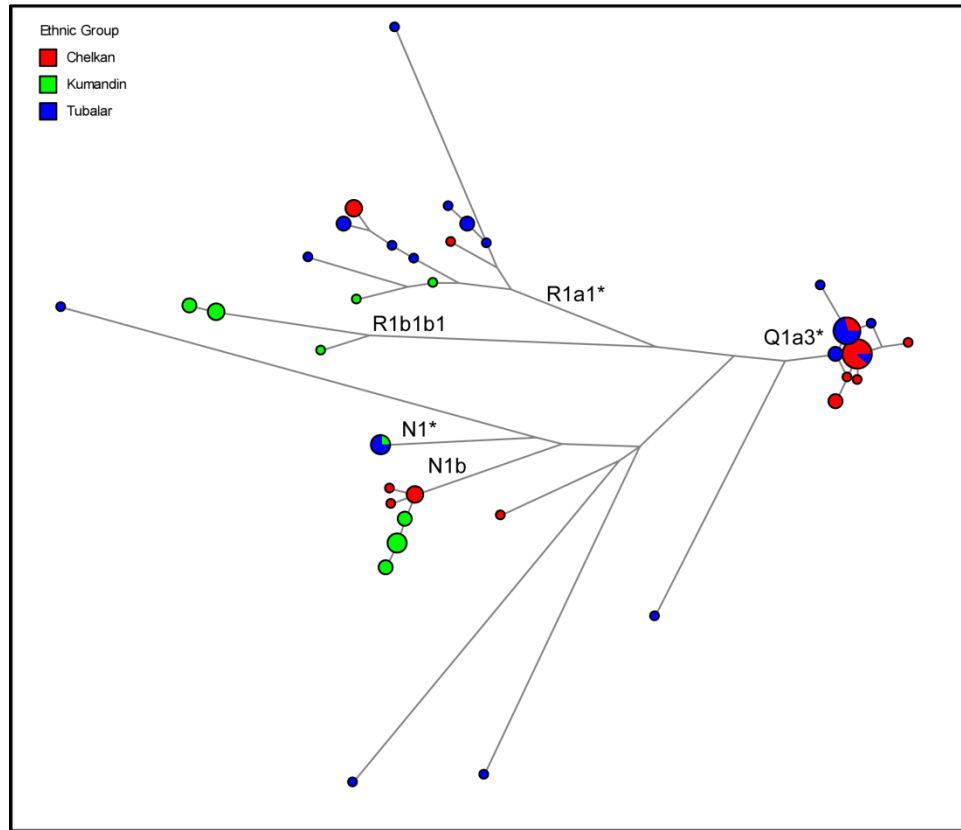


Figure 6.1 MJ-RM network of northern Altaian NRY lineages

Haplotype sharing did occur among the northern Altaian ethnic groups. The Tubalar shared two haplotypes with Chelkan. They belonged to Q1a3\* and were only a single repeat difference from each other. These were the two most common haplotypes for Q1a3\*. The third haplotype belonged to haplogroup N1\* and was found in one Kumandin and several Tubalars. In this instance, the Kumandin participant noted that

both of his parents were also Kumandin. The three Tubalars, however, listed their fathers as Tubalar and their mothers as Russian, but none of these individuals were related. Based on the genealogical information, there is no clear connection amongst these four men. Yet, given that this haplotype was found only in one village and that it is an exact match, this evidence suggests that its distribution in multiple ethnic groups is due to recent intermarriage. No other instances of haplotype-sharing between ethnic groups were noted for the remaining haplotypes. These observations do not suggest that the haplotypes among northern Altaians are not similar. In fact, clusters of ethnic group-specific haplotypes usually only had one or two mutations separating them from each other.

Haplotype sharing was also noted among the southern Altaian populations (Figure 6.2). Three haplotypes were shared between Altai-kizhi and Altaian Kazakhs. One of them belonged to J2a and was found in three Altai-kizhi from Mendur-Sokkon and only one Kazakh from Cherny Anuy. The second haplotype belonged to haplogroup C3c1 and was shared between an Altai-kizhi from Kosh Agach and a Kazakh from Cherny Anuy. The third haplotype belonged to haplogroup O3a3c\* and was shared between many Kazakhs and only one Altai-kizhi from Cherny Anuy. Therefore, in every instance of haplotype sharing between ethnic groups, none of the cases occurred in the same location. It is likely that the J2a lineage came from the Altai-kizhi and the O3a3c\* lineage came from the Kazakhs, given the higher occurrences of these lineages in those populations. One haplotype in Figure 6.2 that appears to be shared between ethnic groups was actually different. The differences between the haplotypes occurred at the DYS19 locus, but this locus was not included in this network (see Methods).

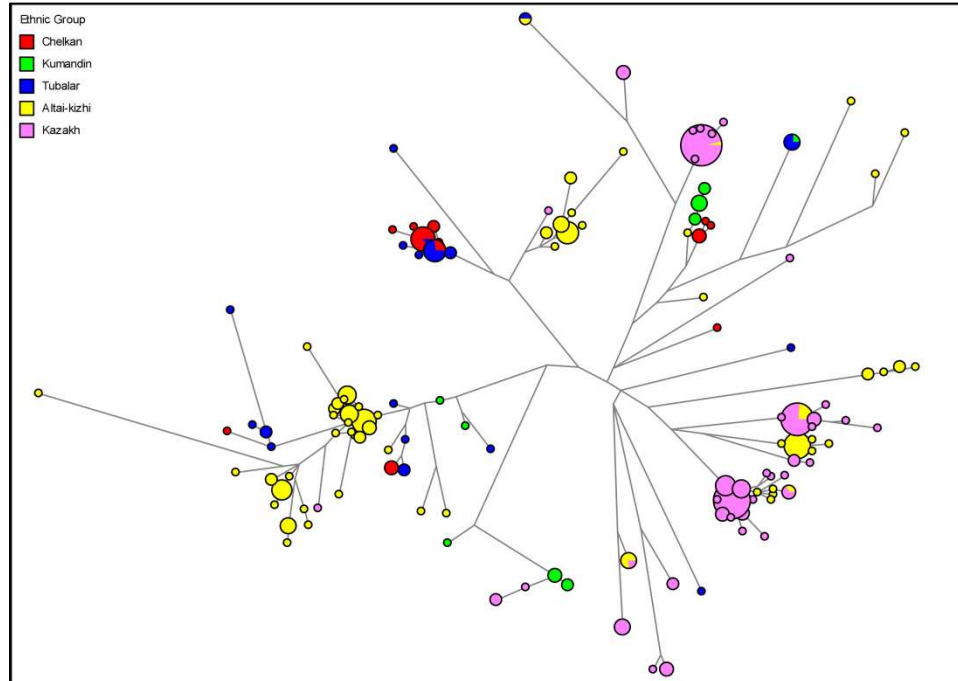


Figure 6.2 RM-MJ network of all Altaian NRY lineages

## 6.4 Genetic Structure among Altaian Populations

### 6.4.1 Haplogroup Analysis

To understand the nature of genetic diversity in the northern Altaians, I first considered whether haplogroup composition varied within an ethnic group relative to its spatial distribution. To this end, Fisher's Exact Tests were employed to assess whether the samples collected for each northern Altaian ethnic group were similar regardless of the village from which they originated. The null hypothesis was that haplogroup frequencies for each ethnic group would be the same in each village (i.e., the haplogroup frequencies of Chelkans in each village are roughly equal). In all three cases, the null hypothesis could not be rejected (Chelkan,  $p = 0.839$ ; Kumandin,  $p = 0.221$ ; Tubalar,  $p = 0.410$ ). Hence, there were no significant differences in the haplogroup composition for

each ethnic group between villages. Because there were no haplogroup differences between Chelkan (as an example), all Chelkan individuals could be pooled together. In other words, it is justifiable to pool all samples into ethnic groups based on the NRY haplogroup data.

An additional Fisher's Exact Test was performed to test whether there were differences between the northern Altaian ethnic groups, with the null hypothesis being that each ethnic group had the same haplogroup composition. In this case, the haplogroup frequencies were significantly different ( $p = 0.000$ ). Given these results, all subsequent analyses treated the Chelkans, Tubalars and Kumandins as separate populations.

A Fisher's Exact Test was also performed on the Altai-kizhi data. The null hypothesis was that there were no differences in haplogroup frequencies between Altai-kizhi at different locations. The test was not significant ( $p = 0.119$ ). Therefore, as with the northern Altaians, data from these populations were combined into a single population for all subsequent genetic analyses.

#### 6.4.2 Haplotype Analysis

Much like the mtDNA analysis, the first priority was determining if NRY diversity was structured in Altaian populations. To this end, the genetic variation was examined first through a geographic lens. Each village was analyzed as a separate population. Pairwise  $R_{ST}$  estimates were calculated from the STR haplotype data and displayed in an MDS plot (Figure 6.3). The three villages from the northwestern portion of the Altai (Dmitrievka, Sank-Ino and Shunarak) all clustered together in the upper left

hand portion of the plot. The genetic distances between these three villages were not significant ( $p$ -value  $> 0.05$ ). The genetic distances between Dmitrievka and the second cluster (Tandoshka and Biika) were also not significant. In fact, none of the genetic distances between the remaining northern Altaian villages were significant. The final cluster was located at the bottom right hand side of the plot and comprised the southern Altaian villages. The genetic distances between Kosh Agach and Cherny Anuy and between Kosh Agach and Zhan-aul were not significant. The only non-significant genetic distance between northern and southern village was between Mendur-Sokkon and Kebezen.

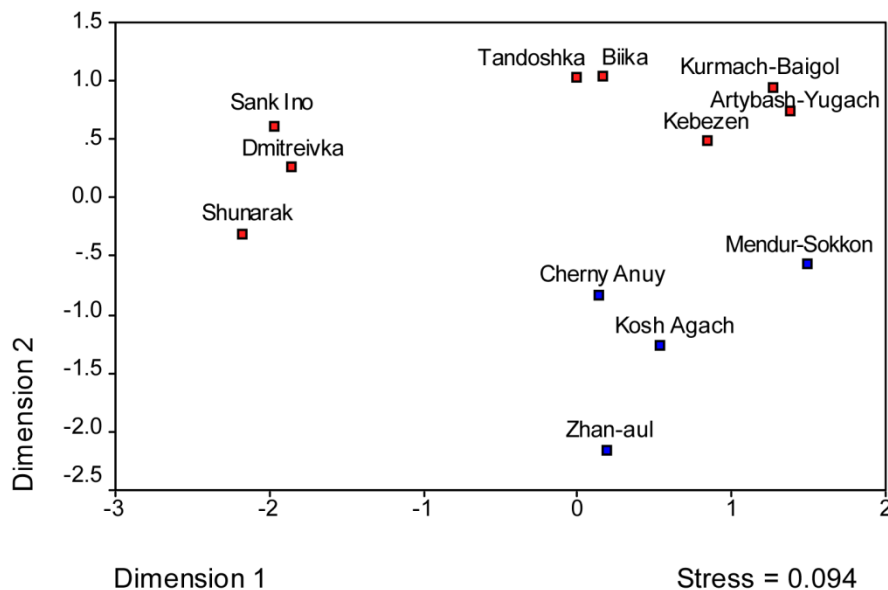


Figure 6.3 MDS plot of village  $R_{ST}$  estimates. Northern locations are represented in red; southern locations are shown in blue.

Separation between northern and southern villages was not found using the mtDNA data (see Chapter 4). Yet, with the Y-chromosome data, it appears that there were three clusters – at least two groups of northern Altaian villages (northwest and

other) and a distinction between most of the northern and southern villages. AMOVA analysis of northern versus southern villages showed that 14.5% of the genetic variation could be attributed to “among group” differences, while 13.8% of the variation was due to differences between villages within the same geographic region (Table 6.2).

Table 6.2 AMOVA of northern versus southern Altaian villages

Groups	Percentage of Variation	P-value
<b>Geography</b>		
<i>Among group</i>	14.50	0.008
<i>Among population within group</i>	13.79	0
<i>Within population</i>	71.72	0

Table 6.2 The “Northern” category comprises Artybash-Yugach, Biika, Dmitrievka, Kebezen, Kurmach-Baigol, Sank-Ino, Shunarak, and Tandoshka. The “Southern” category includes Mendur-Sokkon, Cherny Anuy, Kosh Agach and Zhan-aul.

As in the mtDNA analysis, the composition of each village can be problematic because there were multiple ethnic groups residing together in several locations. Thus, the question whether a village or sample location represent multiple populations was again raised by these results.

To explore the relationships between ethnic groups and villages, each village was separated along ethnic group membership. These ethnic groups per village were used as populations. For example, Tubalar, Chelkan and Kumandin from Tandoshka were designated as three different populations and listed as “Tandoshka Tubalar”, “Tandoshka Chelkan” and “Tandoshka Kumandin.”  $R_{ST}$  estimates were calculated and displayed in an MDS plot (Figure 6.4).

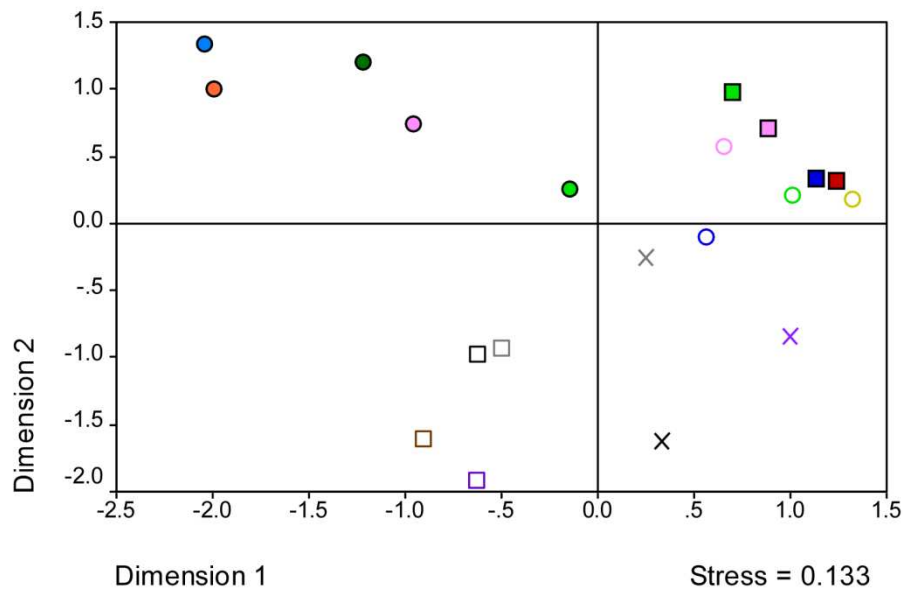


Figure 6.4 MDS plot of  $F_{ST}$  values - ethnic groups and villages. Ethnic groups are designated by symbols: Chelkan (filled square), Kumandin (filled circle), Tubalar (open circle), Altai-kizhi (X), and Altaian Kazakhs (open squares). Villages are designated by color: Artybash (gold), Biika (light green), Dmitrievka (dark green), Kebezen (dark blue), Kurmach-Baigol (red), Sank-Ino (light blue), Shunarak (orange), Tandoshka (light purple), Mendur-Sokkon (dark purple), Cherny Anuy (Gray), Kosh Agach (black), and Zhan-aul (brown).

The MDS plot provided clear evidence that populations grouped together based on ethnic group membership. Symbols represented ethnic groups, and different colors indicated village residence. Filled circles represented the Kumandin populations, which were largely clustered into the upper left hand corner of the MDS plot. Kumandin from Biika were the closest to other ethnic groups. The Tubalar took an intermediate position between the Chelkan in the upper right hand corner of the plot and the Altai-kizhi in the lower right hand side. The Chelkan and Tubalar clustered quite close to each other, with the Kumandins being the outliers. The Altaian Kazakhs also were segregated in the lower middle portion of the plot from the remainder of the populations.

The  $R_{ST}$  values were generally small, and many of them had non-significant p-values (Supplementary Digital File). This result is likely due to the small sample sizes in several of the groups. Nonetheless, the manner in which these populations grouped according to ethnic group membership indicated that the genetic structure might be organized by ethnicity in addition to geography. This certainly seemed to be the case for the Altai-kizhi and Altaian Kazakhs, who resided in many of the same villages.

Two AMOVA tests were run to examine how the northern Altaian NRY variation was partitioned among these populations (Table 6.3). The first AMOVA, which tested whether geography still explained the genetic variation best, grouped ethnic groups by village. For example, “Tandoshka Tubalar”, “Tandoshka Chelkan” and “Tandoshka Kumandin” were designated as three different populations and grouped together as “Tandoshka.” The second AMOVA, which tested whether ethnic group membership better describes the genetic variation, clustered villages by ethnic group. In this case, each ethnic group within a village was considered a separate population and grouped by ethnic group. For example, Artybash-Yugach Tubalar, Biika Tubalar, Kebezen Tubalar and Tandoshka Tubalar were assigned to the “Tubalar” group.

The geography test showed that the “among group” component totaled 16.6%, but the p-value was non-significant (p-value > 0.05). The ethnicity test showed that the Altaian NRY diversity was structured along ethnic group boundaries, with the “among group” component accounting for 21.5% of the total variation. The variation within ethnic groups was non-significant. These results mirror the Fisher’s Exact Tests which showed that differences between the haplogroup frequencies of ethnic groups was significant, and differences within an ethnic group, but between villages, was not (see



page 8). SAMOVA was also used to assess these results, but the findings were not statistically significant.

Table 6.3 AMOVA of northern Altaians

Groups	Percentage of Variation	P-value
<b>Northern Villages</b>		
<i>Among group</i>	16.60	0.002
<i>Among population within group</i>	2.13	0.398
<i>Within population</i>	81.26	0.058
<b>Northern Ethnic Groups</b>		
<i>Among group</i>	21.52	0.003
<i>Among population within group</i>	0.90	0.440
<i>Within population</i>	77.58	0

Table 6.3 Group membership: Northern Villages: Artybash (Tubalar), Biika (Chelkan, Kumandin, Tubalar), Dmitrievka (Kumandin), Kebezen (Chelkan, Tubalar), Kurmach-Baigol (Chelkan), Sank-Ino (Kumandin), Shunarak (Kumandin) and Tandoshka (Chelkan, Kumandin, Tubalar); Northern Ethnic Groups Chelkan (Biika, Kebezen, Kurmach-Baigol, Tandoshka), Kumandin (Biika, Dmitrievka, Sank-Ino, Shunarak, Tandoshka), and Tubalar (Artybash, Biika, Kebezen, Tandoshka).

Once evidence supporting the usefulness of ethnic self-identification was determined,  $R_{ST}$  values were estimated from the STR haplotype data (Table 6.4). These genetic distances showed a close relationship between the Chelkan and Tubalar. All other comparisons were significantly different, with the Kumandin being outliers to all the rest.

Table 6.4  $R_{ST}$  values between Altaian ethnic groups

	Chelkan	Kumandin	Tubalar	Altai-kizhi	Altaian Kazakh
Chelkan	*	0.000	0.146	0.000	0.000
Kumandin	0.316	*	0.000	0.000	0.000
Tubalar	0.024	0.331	*	0.000	0.000
Altai-kizhi	0.237	0.414	0.124	*	0.000
Alt Kazakh	0.350	0.356	0.301	0.237	*

Table 6.4  $R_{ST}$  values are displayed in the lower matrix. P-values are located in the upper matrix.

## 6.5 Within Population Variation

Molecular diversity estimates were calculated to quantify the levels of genetic variation within each of the five populations (Table 6.5). Gene diversity estimates indicated that the Tubalar population had the greatest amount of diversity relative to its population size. Altai-kizhi displayed moderate levels of variation at the haplogroup level, but high variation at the haplotype level. While opposite pattern was true for the Altaian Kazakh. The Chelkan and Kumandin were the least diverse of the five populations. Small population sizes likely had profound effects on the overall genetic composition of northern Altaian populations.

Table 6.5 Summary statistics for Altaian NRY variation

Group	Northern Altaian			Southern Altaian	Altaian Kazakh
	Chelkan	Kumandin	Tubalar	Altai-kizhi	Kazakh
Population					
# of Samples	25	17	27	120	119
# of Haplogroups	4	4	7	10	10
Haplogroup Diversity	0.597 ± 0.088	0.677 ± 0.075	0.735 ± 0.056	0.697 ± 0.036	0.738 ± 0.023
# of Haplotypes	12	9	18	60	38
Haplotype Diversity	0.880 ± 0.050	0.912 ± 0.042	0.954 ± 0.025	0.973 ± 0.006	0.919 ± 0.014

## 6.6 Altaian Local Context

### 6.6.1 Altaian Diversity – Haplogroup (Biallelic Marker) Analysis

Now that the Y-chromosome composition of these five populations has been discussed, we can examine them in relation to surrounding populations. The biallelic marker (SNP) data are discussed first, followed by the STR data. The analyses of these mutation classes were done separately because many of the populations used for these comparisons only had one type of mutation class characterized.

Previous studies of southern Siberian populations provide additional data with which to compare our own (Appendix 1). The Altai-kizhi was represented by three haplogroup data sets. The first was generated for use in this dissertation. Participants came from three locations in the southern Altaian region: Mendur-Sokkon, Cherny Anuy/Turata and Kosh Agach, as mentioned above. The other data sets were taken from the published literature (Derenko, Maliarchuk, & Solovenchuk, 1996; Karafet et al., 2002; Tambets et al., 2004). Karafet et al. (2002) and Tambets et al. (2004) published sample information on indigenous Altaians, although they made no distinction between northern or southern groups in their papers. Derenko et al. (2006) published haplogroup profiles of Altai-kizhi (southern Altaian), Teleut (southern Altaian) and Shor (northern Altaian) populations. In addition, Kharkov et al. (2008) published a study of northern and southern Altaians, but did not investigate the structure of northern Altaian populations. In both cases, only a limited number of SNPs were tested in the population samples, thereby providing only a moderate level of resolution for the data.

To compare the data from these populations, it was necessary to collapse the high-resolution SNP data into a 10-haplogroup profile (Table 6.6). This profile included haplogroups C, D, and E, which are defined by the RPS4Y<sub>711</sub>, M174 and M96 markers, respectively. Haplogroup F (xJ,K) includes all M89 derived Y-chromosomes, except those that had M9 (haplogroup K) or M134 (haplogroup J) derived alleles. Therefore, all Y-chromosomes that belong to haplogroups G, H, or I were placed in this category. Haplogroup J is designated by Y-chromosomes with the derived allele at the M134 marker. Haplogroup K (xN1c,O,P) includes L, M or N derived branches, except for N1c. Haplogroup O describes any M175 derived alleles. Any Y-chromosome with derived

M45 SNP and either the M17 or M198 SNP was defined as R1a1a. M45 derived alleles lacking the M17 or M198 markers were placed into the P (xR1a1a) category.

Table 6.6 Frequencies of 10-haplogroup profiles for Altaian populations

Hg	Chelkan	Kumandin	Tubalar	Altai-kizhi1	Altai-kizhi2	Teleut1	Teleut2	Shor	Altaian Kazakh
C				20.0	13.0	8.5	5.7	2.0	59.7
D				5.0	3.3				
E			3.7						
F (xJ,K)			3.7		3.3	10.7		2.0	5.0
J				2.5	2.2	2.1			4.2
K (xN1c,O,P)	24.0	52.9	11.1	1.7	2.2			13.7	0.8
N1c				2.5	5.4	10.6	28.6	2.0	
O			3.7	1.7					26.1
P (xR1a1a)	60.0	35.3	40.7	16.7	28.3		34.3	2.0	3.4
R1a1a	16.0	11.8	37.0	50.0	42.4	68.1	31.4	78.4	0.8
<b>Total</b>	<b>25</b>	<b>17</b>	<b>27</b>	<b>120</b>	<b>92</b>	<b>47</b>	<b>35</b>	<b>51</b>	<b>119</b>

The Altai-kizhi data from Derenko et al. (2006) and Karafet et al. (2002) were quite similar to those found in our Altai-kizhi populations.<sup>9</sup> High frequencies of R1a1a and moderate amounts C and Q characterized these southern Altaian populations. Derenko et al. (2006) did not test their samples for haplogroup Q, but instead used a P (xR1a1a) category in which Hg Q Y-chromosomes would fall. The frequencies from Derenko et al. were slightly higher for this category and lower for C3 and R1a1a, but the overall haplogroup profile was similar. The high degree of similarity, including a significant amount of haplotype sharing between the two data sets, suggested that the published data from Derenko et al. (2006) derived from southern Altaian populations.

The Teleut (the other southern Altaian population analyzed) was represented by two populations. Teleut1 was made up of R1a1a, F\*(xJ,K), N1c, and C Y-chromosomes. In this case, it was difficult to infer which haplogroups were actually represented by

<sup>9</sup> The Altaians from Karafet et al. (2002) and Tambets et al. (2004) were not listed in Table 6.5, because these publications did not list the haplogroup frequencies for all of their Altaian samples.

F\*(xJ,K). These could include F\*, G, H, or I derived lineages. G was found in Altaian Kazakhs, and I at low frequencies in Siberians, but high frequency in Russians (Balanovsky et al., 2008; Derenko et al., 2006). Therefore, its presence in the Teleut could represent admixture between the relatively recent arriving Russians or a more distant connection to Central Asia. Unfortunately, there was not enough information to determine what specific haplogroups were actually included in this category. By contrast, Teleut2 had significant differences relative to the other Teleut population. Considerably fewer R1a1a Y-chromosomes were present, but a substantially higher proportion of N1c1, R1b (xR1b1b2) and R1b1b2 were also observed. The differences between the two Teleut groups are likely due to genetic drift, as the population size of this ethnic group decreased dramatically in the past century or two. Only 2650 Teleuts live in Russia today (Russian Census 2002).

The Shor population was composed of R1a1a and K\*(xN1c,O,P) Y-chromosomes. Given what is known about surrounding populations, the K\* category was most likely made up of N derived lineages (N1\* and/or N1b), as N1b is prevalent in portions of Siberia, and L, M and O are not commonly seen in this region.

Summary statistics were calculated to evaluate the genetic diversity within each population (ethnic group or regional/ethnographic group) (Table 6.7). Overall, the Tubalar had the greatest haplogroup diversity followed closely by the Altai-kizhi, Shor and Kumandin populations. The lower diversity in Chelkans and Teleuts was not surprising, given the fact that they had high frequencies of only a few haplogroups. The Altai-kizhi had large population sizes within the southern Altai, thus their diversity indices were not surprising either.

The Tubalar were distinctive, possessing six out of the ten haplogroups for only 27 individuals. This finding likely represents the interaction of Tubalar with other groups. For instance, Tubalars were the only northern population to have haplogroups E and I (which were most likely of Russian origin (Balanovsky et al., 2008)) and haplogroup O3a3c1-M117 (found in southwest and northeast China (Xue et al., 2006)).

Table 6.7 Haplogroup diversities in Altaian populations with 10-haplogroup profiles

Group	Northern Altaian				Southern Altaian			
Population	Chelkan	Kumandin	Tubalar	Shor	Altai-kizhi1	Altai-kizhi2	Teleut1	Teleut2
N	25	17	27	51	120	92	47	35
H	3	3	6	6	8	8	5	4
Haplogroup Diversity	0.580 ± 0.081	0.618 ± 0.077	0.707 ± 0.056	0.636 ± 0.080	0.689 ± 0.033	0.725 ± 0.031	0.517 ± 0.080	0.719 ± 0.029
Reference	[1]	[1]	[1]	[2]	[1]	[2]	[2]	[3]

References: [1] this study, [2] Derenko et al. (2006), [3] Kharkov et al. (2009)

Conventional  $F_{ST}$  values were calculated to evaluate the genetic diversity and similarities among Altaian populations. Tests of significance indicated that there were significant difference between the Chelkan, Tubalar and Kumandin, based on the 10-haplogroup profile ( $p < 0.05$ ). The MDS plot of  $F_{ST}$  values supported this interpretation. The northern Altaian groups were scattered across the plot (Figure 6.5). The Shor population was found on the opposite side of the plot, while the rest of the northern Altaians were separated to the right. The two Altai-kizhi populations clustered near one another and, according to the  $F_{ST}$  values, were not significantly different. Alternatively, the Teleut were located far apart and differed significantly ( $p$ -value = 0.000). Of all of the northern populations, the Tubalar were positioned closest to the Altai-kizhi population from Derenko et al. (2006).

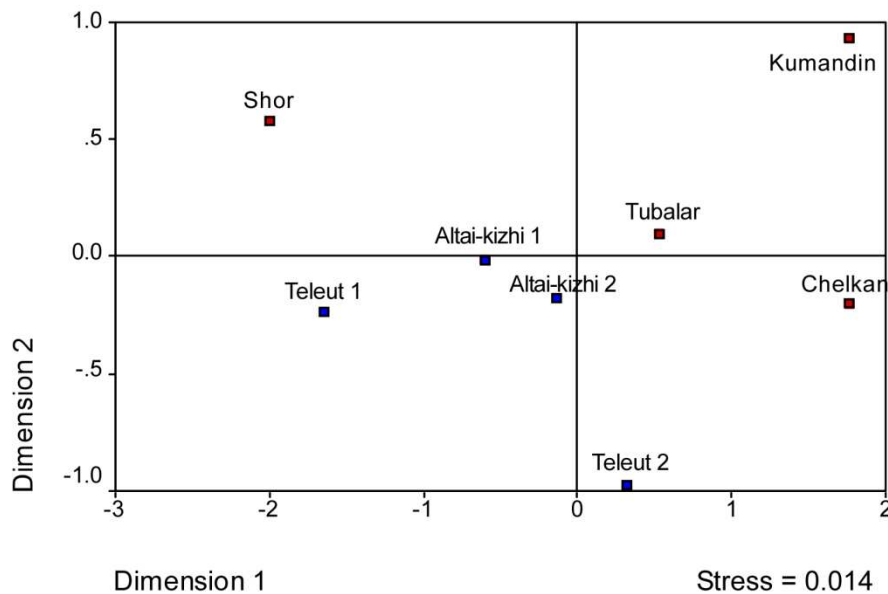


Figure 6.5 MDS plot of conventional  $F_{ST}$  values among Altaian populations

The Tubalar were in closest proximity to the southern Altaians, probably because of the high frequencies of Q and R1a1a in both populations. This relationship was first suggested by Potapov (1962), who noted their shared cultural characteristics and geographic proximity to each other. The Shor and Teleut1 (although belonging to northern and southern categories, respectively) actually clustered near one another, most likely due to the large portion of R1a1a making up each group. Teleut2, however, were quite different from all other populations.

AMOVA analysis was carried out to determine how the genetic variance was divided among and within groups. In this instance, the groups were classified by geography, with Chelkan, Kumandin, Shor and Tubalar classified as “Northern Altaian” and the three Altai-kizhi and Teleut populations as “Southern Altaian” (Table 6.8).

AMOVA analysis of the Altaian populations indicated greater variance for the “among

populations within group” component than for the “among group” component (11.9% versus 0.0%). That is, there was greater variance within northern Altaians or within southern Altaians than there was between northern and southern groups. Given that the Shor and Teleut were so similar in their haplogroup profiles, these populations were combined together in their own group in a second AMOVA test run. The “among group” differences for these three groups (Northern Altaian, Southern Altaian and Shor-Teleut) increased to 12.1%, and the variation “among populations within a group” decreased from 11.9% to 3.47%, thus emphasizing the great similarities of Shor and Teleut1 populations.

Table 6.8 AMOVA results among Altaian populations

<b>Groups</b>	<b>Percentage of Variation</b>	<b>P-value</b>
<b>Northern v. Southern</b>		
<i>Among group</i>	0.51	0.174
<i>Among population within group</i>	11.91	0
<i>Within population</i>	77.58	0
<b>Northern v. Southern v. Shor-Teleut1</b>		
<i>Among group</i>	12.11	0.003
<i>Among population within group</i>	3.47	0
<i>Within population</i>	84.62	0

Table 6.8 Group membership: northern group (Chelkan, Kumandin, Tubalar and Shor); southern group (Altai-kizhi1, Altai-kizhi2, Teleut1 and Teleut2).

The categorization of ethnic groups as either northern or southern Altaian has been a primary distinction of indigenous groups from the Altai-Sayan region in all of the literature discussing the anthropology, linguistics, and prehistory of these peoples. It was therefore surprising that the Chelkan, Kumandin, Tubalar and Shor did not cluster together to the exclusion of all other groups. The close association of Shor and Teleut1 suggested that they shared a common paternal source or had experienced recent gene



flow. The Chelkan, Tubalar, and Kumandin still clustered together and the Altai-kizhi were distinctive from them. Therefore, a north/south division still exists despite the aberration in this pattern for the Shor and Teleut1 populations, which appear to make up an “Eastern Altaian” group. Unfortunately, STR data were not available for the Altai-kizhi2, Teleut1 and Shor, thereby preventing these relationships from being confirmed with haplotype analysis.

The SNP and STR analyses did provide different perspectives on relationships among the Altaian populations. Although the Chelkan and Tubalar had significantly different haplogroup frequencies, their haplotypes were so similar that  $R_{ST}$  genetic distances were non-significant. The reverse pattern was seen with the Tubalar and Altai-kizhi, where their haplogroup frequency profiles were quite similar, yet they significantly differed at the haplotype level. It is not clear what the relationship between the Teleut1 and Shor is (especially considering the position of Teleut2). The haplogroup data indicated similarity between them, but haplotype information would certainly help to resolve this issue.

## **6.7 South Siberian and Regional Contexts**

To understand how Altaian populations are related to their neighbors, I used published data to compare our samples to those from other populations in southern Siberia, northwest Siberia, central Siberia, Mongolia, northern China and Xinjiang, and Central Asia (Appendix 1). The high-resolution data from our sample set was reduced again to the 10-haplogroup profile to allow comparisons with previously published data sets. The addition of southern Siberian populations to the analysis helps us to understand

the Altaian populations in the proper context by delineating their relationships with neighboring groups.

Southern Siberian populations had a haplogroup distribution very similar to southern Altaians. Relatively high frequencies of R1a1a and P (xR1a1a) were noted, while some also had moderate frequencies of K (xN1c1,O,P), most of which presumably belong to N1b Y-chromosomes. One difference, however, was the higher frequencies of haplogroup O in Tuvinian and Tuvinian-derived populations. It was nearly absent in Altai-kizhi and Tubalar and was completely missing from Chelkan, Kumandin, and Shor. The Buryats were the only populations to deviate drastically from this southern Siberian haplogroup profile. Buryats tended to have very high frequencies of C-derived Y-chromosomes and moderate frequencies of N1c1 and O.

The MDS plot that included southern Siberian populations resembled the plot of Altaians only (Figure 6.6). Chelkan and Kumandin were still separated from the main cluster. The Tubalar fell near the Altai-kizhi and Teleut2. The Tuvinian and Todzhan also clustered near the Altai-kizhi populations, but the Khakass, Soyot and Tofalar extended away from this cluster, up the y-axis. Buryats were excluded from this plot because they were extreme outliers in an MDS plot that included them, making the relationships between other populations difficult to ascertain (data not shown).

A number of populations shared non-significant values at the 0.05 level. The northern Altaian Chelkan, Tubalar and Kumandin had non-significant p-values. The other main group was the central cluster of populations – the Altai-kizhi1, Altai-kizhi2, Teleut2, Tubalar, Todzhan and Tuvinian. This cluster represented the largest populations in southern Siberia, which essentially showed a common paternal ancestry based on the

biallelic markers. After applying a Bonferroni correction (significance level  $p = 0.003$ ), some of these groups became more inclusive. For instance, the Teleut and Shor clustered together as did the Chelkan and Kumandin with Tuvinians. The Kumandins also had closer affiliations with the Tofalar/Khakass group, while the Tubalars were not significantly different from Todzhans, Khakass and Altai-kizhi.

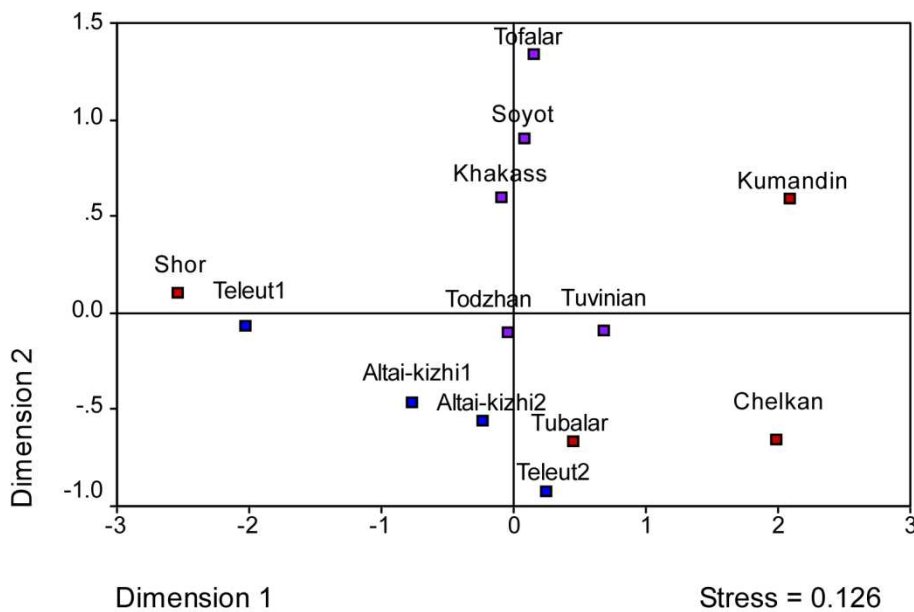


Figure 6.6 MDS plot of conventional  $F_{ST}$  values for southern Siberian populations

These results emphasized the genetic similarities of Tuvinian populations with many of the other southern Siberian groups. This finding was expected, as several of these populations were believed to originate from the Tuvan region. In addition, the same general pattern was found in the mtDNA analysis.

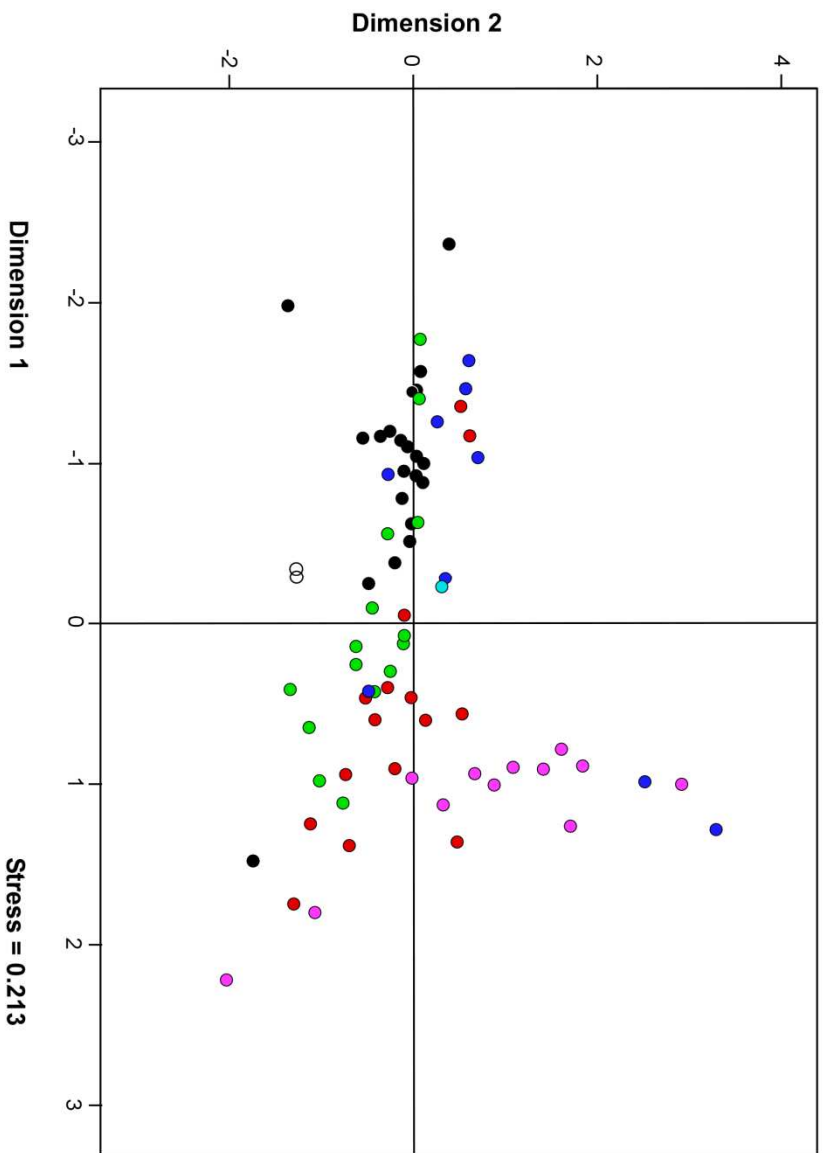
Unfortunately, as for the NRY analysis of Altaian populations, STR data were not available for many of the southern Siberian populations, except the Teleut2. Therefore,

the relationships identified with the haplogroup data could not be verified with haplotype analysis.

## **6.8 Altaians from a Global Perspective**

Again, to gain a better understanding of the placement of Altaian and southern Siberian populations in relation to a broader geographic context, these populations were compared to other Siberian and Central Asian populations. Conventional  $F_{ST}$  values were visualized on an MDS plot (Figure 6.7). The major pattern of the MDS plot was the separation of northwestern Siberian populations and Northern China / Mongolian populations. Located between these groups were the southern Siberians and Central Asians, which largely overlapped with each other. The central Siberian populations were mostly found just above the Mongolian population cluster. This group included Evenks, Evens, and two southern Siberian Buryat groups. This pattern is consistent with the mtDNA data and historical records that show affinity between Mongolian, Buryat and central Siberian populations. The northwest Siberians were largely made up of Finno-Ugric speakers and Samoyedic speakers. The two central Siberian populations in this cluster were the Yakut and Western Evenk.

In this case, the Central Asians and southern Siberians clearly fell into the same general cluster. The Altai-kizhi and Tuvinians were found right next to Uyghur and Uzbek populations. The northern Altaians still did not cluster together, with the Tubalar being placed near the Altai-kizhi and Central Asians, the Kumandin up near the northwestern Siberians and the Chelkan down near the Shor and Teleut1. Additionally, the Sel'kup and Ket were found in this cluster at the lower right-hand side of the plot. As



- Green – Central Asia
- Red – Southern Siberia
- Dark Blue – Central Siberia
- Light Blue – Northeastern Siberia
- Violet – Northwestern Siberia
- White – Tibet
- Black – Mongolia / Northern China

Figure 6.7 MDS plot of  $F_{ST}$  values for Siberian and Central Asian populations

mentioned before, these groups were historically situated along the Yenisei River, far to the south of the land they inhabit now. These analyses showed a similarity between the groups from that region and provided support for the notion that they derived from a common paternal source.

The Chelkan were not significantly different from Sel'kups and Dolgans ( $p$ -value  $> 0.05$ ). When a Bonferroni correction was applied (0.001 significance level), the Uyghur and Mari groups were also included. The high frequency of haplogroup P\*(xR1a1a) in the Chelkan was a contributing factor to this pattern, although it should be noted that it also makes the Kumandin and Chelkan appear more similar than they actually are. When considering the high-resolution SNP data, it is clear that the Kumandin lacked haplogroup Q1a3\* and had high frequencies of R1b1b1, whereas the opposite was true for the Chelkan, although both Q1a3\* and R1b1b1 were counted together in the P (xR1a1a) category. If differences like these also exist between the Chelkan, Sel'kup and Uyghur haplogroups, then they too would be masked in this analysis.

A separate cluster of Tajik, Kyrgyz, Shor and Teleut populations were located in the bottom, right area of the plot. In the center was a large conglomerate of Central Asian (Uyghurs and Uzbeks) and southern Siberian (Altai-kizhi, Tubalar, Tuvinian, Todzhan, Soyot, Tofalar, Khakass) groups. Remarkably, the Altai-kizhi had the greatest affinity with populations not found in southern Siberia, according to the conventional  $F_{ST}$  values. However, the MDS plot placed them at the edge of the central cluster near the Uzbek, Uyghur and Todzhan. Both the  $F_{ST}$  values and MDS plots indicated a stronger affiliation of this group with Central Asian populations. The Altai-kizhi shared the

smallest  $F_{ST}$  values with the Central Asian Kyrgyz, Uzbek and Tajik populations. It was only when Bonferroni corrections were applied to the data that the Tubalar, Teleut and Todzhan genetic distances became non-significant ( $p > 0.001$ ). The large portion of haplogroup R1a in these populations certainly had much to do with these results.

Similarly, Kumandins were closest to Chelkans, Tubalars and Tuvinians. However, they also shared low  $F_{ST}$  values with other southern Siberians (Tofalar and Khakass), Uyghurs and northwestern Siberian populations (Dolgans, Mansi and Khanty). Interestingly, Kumandins were most similar to the Chelkan and the Mansi. These affinities were most likely due to the prevalence of haplogroup N1b in these three populations (K (xN1c,O,P) in this analysis).

## **6.9 Population Haplotype Analysis**

While haplogroup observations are acceptable at a cursory level, the Y-chromosome lineage data that provide a haplotype-based assessment allow for a nuanced analysis of the relationships among individuals, ethnic groups and regions. Appraisal of Y-STRs in addition to the haplogroup defining SNPs accomplished this task. The populations used in the haplotype analysis were not always the same as those used in the previous analysis of haplogroup frequency differences. This discrepancy occurred because the level of resolution for the data and the type of mutations characterized were often different in publications describing Y-chromosome variation in Siberian and Asian populations. Many of the papers used only STR data or only SNP data, with very few providing information from both types of Y chromosome mutations. Many of the

publications that did provide such STR information, did so for only select haplogroups, thereby making it impossible to use these data at a population level.

In short, population data sets were used only when all haplogroups were represented with STR data. As a consequence, the following haplotype analysis utilized Teleut, Khanty and Mansi, Central and Western Evens, Yukaghir, Iengra Evenks and Stony Tunguska Evenks, Uyghur from Urumqi and Yili, Mongolian, Yakut, Hazara, and Turkish in addition to those populations that I have characterized (Chelkan, Kumandin, Tubalar, Altai-kizhi, and Altaian Kazakh) (Appendices 1 and 2). These analyses focused on a 7-STR profile that included DYS19, DYS389I, DYS389b, DYS390, DYS391, DYS392, and DYS393. Genetic distance estimates were calculated in the form of  $R_{ST}$  values. These values were calculated using microsatellite data, and are the equivalent of  $F_{ST}$  values obtained in DNA sequence and SNP analysis.

$R_{ST}$  values between populations were displayed on an MDS plot (Figure 6.8). In this figure, the Yakut population immediately stood out as an outlier. This population is composed of mostly N1c haplotypes, which were shared among other populations, but at much lower frequencies. The Chelkan and Kumandin were found in a cluster among the Mansi and Khanty populations. Their position must be due to the significant presence of N1b in these ethnic groups. Other populations that also had high frequencies of N1b, such as the Nganasan, Nenet, Udmurt and Dolgan, were not included due to the lack of STR information for their Y haplotypes (Karafet et al., 2002; Rosser et al., 2000).

The Tubalar (the third northern Altaian population in this analysis) was actually positioned closer to the southern Altaian Teleuts (Teleut2). The other southern Altaian ethnic group, the Altai-kizhi, was positioned away from all other populations. Uyghur



populations fell in a central cluster with the Iengra Evenks and Hazara. Nearby were the Mongolians and Stony Tunguska Evenks, which were closest to the Altaian Kazakhs.

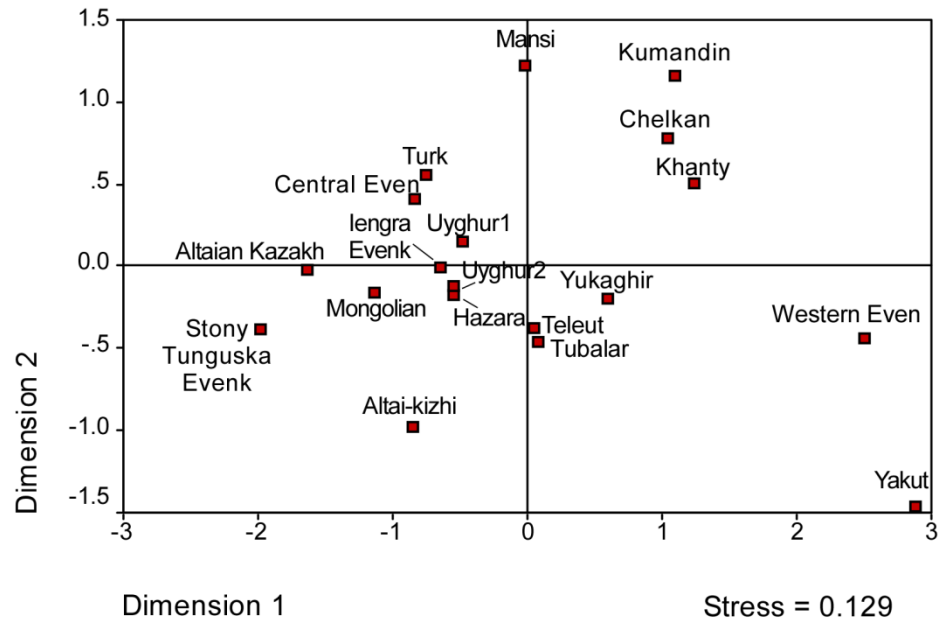


Figure 6.8 MDS plot of  $R_{ST}$  Values for Siberian populations

$R_{ST}$  values indicated that the Kumandin were distinct from all other groups, while Tubalars were similar to Iengra Evenks ( $p > 0.05$ ). Both Chelkan and Tubalar were similar to Yukaghirs, probably due to the presence of haplogroup Q. When a Bonferroni correction was applied ( $p = 0.003$  significance level), the Chelkan were not significantly different from the Ugric-speaking Khanty and Mansi, or from the Iengra Evenk and Yukaghirs. These groups, with the addition of Teleut and Hazara, were also similar to the Tubalars. By contrast, Kumandins only showed affinities to the Iengra Evenk and Yukaghirs, regardless of applying a Bonferroni correction.

Among the northern Altaians, the Chelkan and Kumandin share no haplotypes. The Chelkan and Tubalar share just a single R1a1 haplotype (16-14-17-25-11-11-13). This haplotype is widely distributed and is found in several other populations. In particular, five Altai-kizhi, five Teleut, one Mansi and one Uyghur (Urumqi) also have this haplotype. According to the YHRD, this haplotype is also found in Uyghurs from Yining and in Tuvinians from Tuva (Willuweit & Roewer, 2007). A second Chelkan R1a1 lineage (16-14-18-24-9-11-13) was also found in Mongolians. A third R1a1 lineage found in Tubalars (16-14-19-24-9-11-13) was shared with the Uyghur (Urumqi), Hazara, and Turks. Finally, a fourth Kumandin R1a1 lineage (15-13-17-25-11-11-13) was also characterized in Uyghur (Urumqi) individuals and a single Mansi. Therefore, many of the closely related R1a1 lineages found in northern Altaians were not seen in southern Altaians, but appeared in other populations from southern Siberia, northwest Siberia, Mongolia and the Xinjiang region on China.

Chelkan N1b haplotypes also have a wide distribution in northern and northwestern regions of Siberia. They are shared with Khanty, Mansi, Evenks from Stony Tunguska, Central Evens, Uyghur (Urumqi), Turks, and a single Western Even. YHRD shows an identical haplotype being present in Tuvinians (Willuweit & Roewer, 2007). Unlike the Chelkan, the Kumandin N1b was only shared with a single Altai-kizhi. Furthermore, a one-step variant of this haplotype was found in a single Nepalese sample (Gayden et al., 2007).

The Tubalar and Kumandin share only one haplotype, which belongs to N1\* (17-14-17-24-10-14-14). No other haplotypes with this motif have been found, which, as mentioned in a previous section, is restricted to a single village in the northern Altai

region. More intriguing, Kumandin had two R1b1b1 haplotypes. One of them was shared with a few populations, including Teleuts and, according to the YHRD, Kazakh and Ukrainian populations. The second (and more abundant) R1b1b1 haplotype was found exclusively in the Kumandin population.

Multiple Q1a3\* lineages were shared between the Chelkan and Tubalar. However, no other populations are currently known to have the same Q haplotypes. A haplotype from a Turkish individual (YHRD) had a single step difference from one Tubalar (13-14-17-24-10-14-13) haplotype. However, the Altai-kizhi shared two Q1a3\* lineages with Tuvinians (with a difference at the DYS385 locus). This basal haplotype (13-13-18-23-10-14-13) was also seen in Yakut and Mongolian populations. Given the similarities between northern Altaians and Tuvinians for R1a1a lineages, it was surprising that the same pattern was not noticed for the Q1a3\* lineages.

C3 and C3c lineages represented many of the Altaian Kazakh haplotypes, while C3 contributed significantly to the Altai-kizhi NRY diversity. Altai-kizhi C3 lineages were shared with the Stony Tunguska Evenk and Mongolians. According to the YHRD, they also shared haplotypes with Uyghur (Yining), Inner Mongolians and Buryat. Altaian Kazakh C3 had a broader, more widely dispersed distribution, specifically among Uyghur (Urumqi), Mongolians, Turks, Hazara and Madjar, as well as the Altai-kizhi. Notably, the Genghis Khan modal haplotype was found in Altaian Kazakh and Altai-kizhi and was shared with groups throughout Eurasia (Zerjal et al., 2002; Zerjal et al., 2003). Altaian Kazakh C3c lineages were observed in Uyghur (Yili and Yining), Mongolian, Kalmyk, Kazakh, Chinese and Tuvinian. By contrast, Altai-kizhi C3c

haplotypes were only found in Mongolian, Kazakh and Kalmyk populations – most likely because of recent Mongol political hegemony over the southern Altai.

One of the Altai-kizhi and Altaian Kazakh J2a haplotypes are the same; they also appear in Uyghur (Urumqi), Turkish, and Madjar populations. A different Altaian Kazakh J2a haplotype is also found in Turkish, Nepalese and Czech populations. Most of the O3a3c haplotypes in Altaian Kazakhs are also found in Kazakh and Uyghur populations. This trend is also true for the Altai-kizhi O3a3c haplotype, but it was not the same as that seen in Altaian Kazakhs.

When using the 7-STR profile, haplotype sharing does occur between the Altai-kizhi and northern Altaians. For example, an R1a1a haplotype was shared between Chelkan, Tubalar, Altai-kizhi, Teleut, Mansi, and Uyghur (Urumqi) populations. Tubalar also shared a single O3a3c haplotype with the Altai-kizhi, while a single N1b haplotype was shared between the Kumandin and Altai-kizhi. However, the haplotype sharing between northern and southern Altaian groups did not persist for the most part when the data set was expanded to a 17-STR profile, as discussed above.

Altai-kizhi had greater affinities to Central Asians in part because of the presumed Mongol influence on their genetic make-up. C3 derived lineages are found in Tuva as well, and may also be the consequence of continued a Mongol presence and political influence throughout much of the historical record of this country (Barfield, 1989; Golden, 1992; Grousset, 1970; Potapov, 1964e).

As for the northern Altaians, nothing can be stated about Samoyedic and Yeniseian-speaking elements in these populations because there was no STR population data from these modern populations or their historical antecedents. Full STR profiles of

Khakass, Ket and Sel'kup populations would surely help to provide the comparative information necessary to infer about the genetic relatedness between northern Altaians and their presumed Siberian forest belt origins. The fact that the Chelkan and Kumandin showed some affinities to the Khanty and Mansi supports this origin hypothesis, but the results are far from conclusive.

Chelkan also have R1a1a\* haplotypes that appear in Tuvinians and Mongolians. They have a single 7-STR haplotype that appears in Altai-kizhi, Teleut, Mansi, and Uyghurs, although it has already been shown that Chelkan and Altai-kizhi do not share any lineages when using all 17 STRs. One might anticipate that the Chelkan and Tubalar would share a number of haplotypes with Tuvinians, but as it turns out, the set of Q haplotypes shared among these northern Altaian groups are not found anywhere else in Siberia, including the Tuvan region. The Q lineages from Tuvinians actually match several of the haplotypes found in Altai-kizhi. These findings seem to indicate at least two sources of haplogroup Q exist for populations in southern Siberia. One of them supplied the Tuvinians and Altai-kizhi with their haplotypes and maybe also populations from Central Asia, while the other provided Q lineages to northern Altaian groups. In this regard, having Q haplotype data from northwestern Siberia would be very useful, since it could be a potential source area for the northern haplotypes (or link the Altai as source of northwestern Siberia types). This scenario will be discussed in more detail in the next chapter, in which the phylogeography of haplogroup Q is described.

## 6.10 Chapter Conclusions

As the preceding discussion attests, the first objective of this study was clearly met. Through this analysis, I generated high-resolution NRY data for Altaian populations, the level of which has not yet been published for Siberian populations. The second objective was also achieved in that at least two distinct population histories in the Altai were verified, those of the northern and southern ethnic groups. Ethnographic, linguistic and historical evidence had suggested a clear division between the northern and southern Altaian populations. For the northern Altaians, their process of ethnogenesis is thought to parallel, in part, that of the Khakass ethnic group. This would have involved the consolidation of previously separate tribes and clans who spoke various Samoyedic or Yeniseian languages on the Yenisei River and Altai-Sayan plateau (Forsyth, 1991, 1992; Potapov, 1962, 1964a).

Contrary to this assumption, northern Altaians possess smaller genetic distances from and similar haplogroup frequencies to Tuvinians. Although Samoyedic, Yeniseian, Mongolic, and Turkic-speaking populations helped to create the current gene pool of Tuva, its population is much larger than that of other south Siberian populations. The smaller populations on the Yenisei were some of the first to be affected by Russian colonization, resulting in the complete cultural assimilation and linguistic annihilation of most of these smaller semi-nomadic groups (Arins, Kotts, etc.) (Forsyth, 1991, 1992). The populations from which northern Altaians were derived are therefore, in effect, already extinct, making them and the consolidated Khakass the closest living descendants.

The third objective was to identify clan or tribal structure among northern Altaian ethnic groups. It was clear that differences persist among some of these groups. In particular, the Kumandin tend to have distinctive NRY profiles, while the Chelkan and Tubalar share a number of similar Y-chromosomes. This pattern is unlike that seen with the mtDNA data, where the Chelkan were the most distinctive population.

Upon closer inspection, the northern Altaian ethnic groups do not necessarily show the same histories. The Chelkan and Kumandin are thought to derive from populations originating in the Yenisei and Altay-Sayan plateau that are now assimilated, although historical evidence also records the movement of Tuvinians into the Altai (Forsyth, 1992; Potapov, 1962). This is not unusual, as several ethnic groups in southern Siberia have such an origin. The Tofalar and Todzhan both separated from Tuvinians some time ago and speak languages very similar to each other. Tofalars call themselves “Tubalars” (Tuba = person; Tubalar = people). A group of Tubalars were thought to have moved from either Tuva or Tofalar territory into the northern Altai in the 19<sup>th</sup> century (Potapov, 1962; Wixman, 1984).

Based on these genetic data, a hypothesized Tuvinian origin for the Tubalar is not unreasonable. The Tubalar do have relatively high diversity estimates, suggesting they derive from a rather large population, or have had considerable interaction with neighboring groups. Yet, there is no clear connection between the Tubalars and Tofalar in the analysis, and therefore, it seems unlikely that the Tubalar of the northern Altai originated from those people.

Of course, there is the question of how these estimates would change if the P (xR1a1a) category was disaggregated into its constituent haplogroups based on high-

resolution SNP data. It has already been shown that the Chelkan and Kumandin Y-chromosomes in the P (xR1a1a) category are actually dissimilar if the high-resolution haplogroups are considered. Based on SNP data (Pakendorf et al., 2006), one Tuvian population is made up of 16.4% Q and 7.3% R1 (xR1a1) haplogroups. These R1 (xR1a1) haplotypes could belong to R1b1b1, as is the case for Kumandins, or some other branch of R1. This ambiguity will need to be clarified through the comparison with available STR data and additional SNP typing. While it may seem unlikely that Kumandins derived from or share common origins with Tuvians, this possibility for Chelkan and Tubalar cannot yet be excluded.

When considering the close association of the Chelkan and Kumandin with the Sel'kups and Dolgan populations, the Samoyedic roots of the Altaian groups become clearer. Kumandins show some affinities with the Khanty and Mansi, as well as the Khakass and Tofalar. This pattern is what we would expect if a considerable proportion of their ancestry came from (or is shared with) historical Samoyedic, Ugric and Yeniseian populations of southern Siberia. The only Yeniseian speaking group still in existence is the Kets, who live along the Yenisei River in the Krasnoyarsk Krai (Popov & Dolgikh, 1964; Vajda, 2001). Less than 1500 individuals have not been assimilated into neighboring groups (Russian Census 2002). One group of Kets participated in a Y chromosome study (Karafet et al., 2002), but it is not clear which haplogroups other Yeniseian speakers would have possessed if they had been tested for the same SNP markers that our populations were. Therefore, it is difficult to know exactly what Yeniseian speakers contributed genetically to these populations in southern Siberia. However, if we assume that they possessed similar genotypes to those observed in the



remaining Ket populations, then we would expect to see high frequencies of haplogroup Q in northern Altaian groups. In this context, we did observe haplogroup Q in the Chelkan and Tubalars, but not in the Kumandins.

The Shor population is sometimes considered to be a northern Altaian ethnic group, as it shares cultural, linguistic, and physical similarities with Chelkan, Tubalars, and Kumandins (Potapov, 1962). Based on its paternal population history, this assertion is not justified. The Shors were thought to derive from Samoyedic, Ugric and Yeniseian speakers, much like the Khakass (Potapov, 1964d). However, based on their Y-chromosome data, this assertion is difficult to confirm, despite the putative ethnographic connection between Ket and Shor. Interestingly, the Aba clan of Shors share connections with the Teleut in the form of epic stories and in cattle-breeding techniques (Potapov, 1964d). Potapov believed this was a consequence of a Teleut element being involved in the formation of the Shor ethnic group (Potapov 1964, 444). From the 17<sup>th</sup> century onwards, there was extensive mixing of Shor and Teleuts groups (Potapov, 1964d). This relationship is not refuted by the Y-chromosome SNP data and, in fact, is supported by the genetic distance and AMOVA analyses. However, it will need to be further explored through haplotype/lineage analysis.

The Altai-kizhi is not particularly close to the other southern Siberian populations. Rather, the closest populations to the Altai-kizhi are the Kyrgyz, Uzbek and Tajik. All are Central Asian populations residing to the west and southwest of the Altai-Sayan region in Kazakhstan and Uzbekistan, around the rivers Syr Darya and Amu Darya. The Kyrgyz and some Uzbeks speak languages of the Kipchak branch of Turkic, the same as that spoken by Altai-kizhi. Other Uzbeks and Tajiks speak Indo-European languages.

All are found in the Central Asian steppe region that has long been dominated by pastoral nomadic communities. This subsistence pattern likely had a greater affect than indigenous Siberians on the genetic composition of southern Altaians.

In conclusion, at least two major gene pools are apparent in southern Siberia, with the boundaries of the steppe and taiga serving as a demarcation between them. To the north are populations like the Khakass, Shor, Khanty, Mansi, Ket, and Nganasan. Conversely, there are many southern Siberian and Central Asians in the other, including groups like the Altai-kizhi, Kyrgyz, Kazakhs, Mongolians, and Uyghurs. Those populations living along this boundary seem to have been influenced by both groups, with the Chelkan and Kumandin having greater affinity with those populations to the north of this boundary (forest belt), and the Tubalars being more similar to the Tuvinians.

Ultimately, just like their maternal ancestry, the northern Altaian groups share more cultural, linguistic, ethnographic and (paternal) genetic affinities with Ugric, Samoyedic and Yeniseian groups. The Y-chromosome variation certainly suggests that the boundary between steppe and taiga persists today. Therefore, most of the Northern Altaians ancestors likely descended from those populations that originated from the hunter-gatherers that historically were known to reside in southern Siberia. The southern Altaians again provide undisputable evidence that they originated from the same general gene pool as most Central Asian populations sharing similar steppe cultures. These data provide even more evidence that the Altai has maintained a “persistent frontier” where genetic, linguistic and cultural admixture continue to influence the populations residing in each region and where distinct cultures, lifestyles, modes of subsistence and ways of life can be found on either side of this boundary.

## Chapter 7: NRY Phylogeography

The phylogeography of haplogroups and the lineages comprising them provide critical information about the possible sources and timing of origins for each haplogroup (Jobling et al., 2004). Understanding these questions is crucial for determining how southern Siberia was populated and how the associations among present day populations were created. In this chapter, I examined each haplogroup individually, and assessed the relationships among lineages through a series of networks. Each of the major haplogroups found in the Altai was discussed in an attempt to parse out the different facets of the populations' histories. These details can be found through examining clues in the genetic data. Such clues may signal particular biological (and presumably) historical events, but will certainly provide a relative chronology that will inform us as to which haplogroups are the oldest and which are the products of more recent events.

Repeat differences that make up STR haplotypes generally were not as diagnostic as mtDNA SNP polymorphisms because of the high mutability of STRs. In Chapter 5, branches for each haplogroup of the mtDNA phylogeny were defined by diagnostic polymorphisms (or polymorphism motifs). This was not possible with the Y-STR haplotypes. Instead, subsets of haplotypes that were closely related grouped together in the network analysis, thereby creating a "haplotype cluster." These clusters were often derived from a single common ancestor and therefore their identification was essential for delineating the history of each haplogroup. Estimates of the variance in STR repeat differences and of the TMRCA of haplogroups and haplotype clusters further provided the data necessary to assess when and where a haplogroup may have originated. To this

end, I completed network analyses, intrapopulation variance estimates and calculated TMRCA's for each haplogroup.

## **7.1 Haplogroup N**

Haplogroup N, which is defined by the NRY marker M231, has been found in populations throughout Northern Eurasia from the Pacific Ocean in the east to the Baltic Sea in the west. Several papers recently examined the distribution of this haplogroup in European and Siberian populations to clarify the emergence and expansion of haplogroup N Y-chromosomes in northern Eurasia (Derenko, Malyarchuk, Denisova et al., 2007; Rootsi et al., 2007). A phylogeographic analysis of our new data and those from published N lineages was performed to understand the origin and phylogenetic relationships of Altaian lineages within this haplogroup.

RM-MJ networks were constructed for haplogroups N and NO\*. A 7-STR profile allowed the inclusion of the greatest number of cultural/ethnic/linguistic categories as well as the broadest geographical area (Figure 7.1). This profile (DYS389I-DYS389b-DYS390-DYS391-DYS392-DYS393-DYS439) was used to assess the phylogeography of N. Network construction included 167 haplotypes representing 587 sampled individuals.

The resulting network consisted of two large clusters, with the rest of the samples scattered diffusely around the center of the network. The light blue circles denote the NO\* (xN,O) haplotypes. These were lineages possessing the M214 marker, but lacking the defining markers for its daughter haplogroups N (M231) or O (M175). These two sister haplogroups account for many of the Y-chromosomes found throughout East and

Southeast Asia (haplogroup O), most of Siberia (haplogroup N1) and some of Northeastern Europe (haplogroup N1c). China and eastern Central Asia are the most likely source for these haplogroups, because most of the NO\* and N1\* (xN1a, N1b, N1c) lineages were found in populations living in those areas. The haplotypes found in those areas also possessed relatively high variances. In fact, southern Siberia is the only location where all N branches are found.

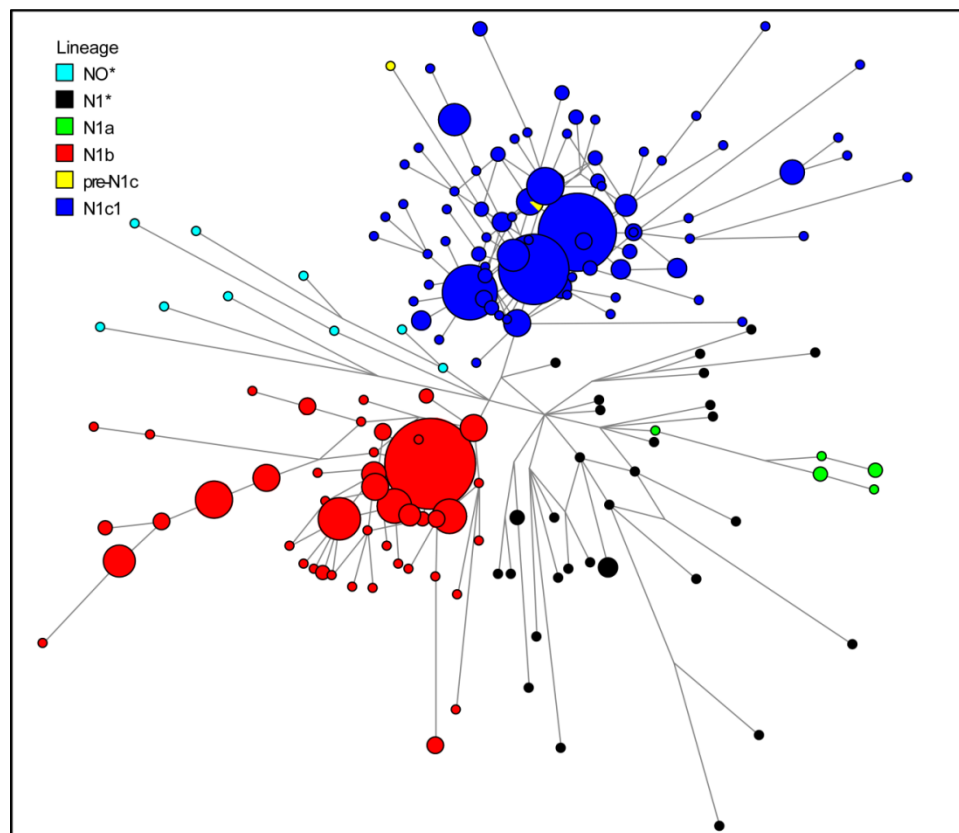


Figure 7.1 RM-MJ network of haplogroup N (7-STRs)

The N1\* lineages (black) had the most diverse STR haplotypes of the N-derived lineages. They appear throughout East Asia, but are concentrated in non-Han populations

in northern and southern China. These precursors to the more derived branches were almost completely absent from Siberia and Europe, indicating that N1 most likely arose somewhere in Central Asia or possibly western China (Derenko, Malyarchuk, Denisova et al., 2007; Rootsi et al., 2007). One such lineage was found in an Altai-kizhi individual from Mendur-Sokkon in the Altai. The current distribution of these Y-chromosomes could be a reflection of a population expansion in East Asia, whereby the Neolithic populations that possessed N Y-chromosomes were displaced by those carrying the O Y-chromosomes that are so abundant in China today.

Of the three derived types of N1, N1a (green) was relatively scarce. It was found in Manchu and Xibe from northeastern and northwestern China, Kazakhs from Kazakhstan, and Buyi from southern China, all at low frequencies (Rootsi et al., 2007; Xue et al., 2006). There was little variation among populations within this haplogroup, suggesting that it had relatively recent emergence.

N1b (red) had one large cluster near the center of the network and a smaller, but more diverse set of lineages from Siberia and Europe. While N1b was characterized in many regions, most lineages originated from Siberia. Unlike N1b, N1c (blue) seemed to have a higher proportion of European representatives and more variation than N1b. Thus, a more detailed analysis of N1b was necessary, as this haplogroup constituted a significant portion of the northern Altaian Y-chromosomes.

#### 7.1.1 Haplogroup N1b

N1b is defined by P43 and occurs in northeastern Europe, European Russia, throughout Siberia, and parts of northern China (Derenko, Malyarchuk, Denisova et al.,

2007; Karafet et al., 2002; Rootsi et al., 2007; Xue et al., 2006). The highest frequency of N1b occurs in the Samoyedic-speaking Nganasans and Nenets of northern Siberia (Tambets et al., 2004). A separate network was created from the data generated from the Altaian samples and those available in the published literature (Figure 7.2). Only seven STRs were found in common among these data sets (same as set as above). In this particular network, 45 haplotypes represented a total of 231 N1b Y-chromosomes.

Two clusters of N1b lineages have previously been distinguished (Figure 7.2). These were defined as Asian (N1b-A) and European (N1b-E) clades (Derenko, Malyarchuk, Denisova et al., 2007; Rootsi et al., 2007). The Asian cluster (red) was considerably more frequent than the European (blue) was and had a wider distribution. These clusters were defined in Rootsi et al. (2007) by 14-13-16-14-19 (DYS19-DYS385b-DYS389b-DYS392-DYS448) for N1b-A and 13-12-18-12-18 for N1b-E. Two additional clusters were identified by Derenko et al. (2007). They were labeled as Asian 1 (N1b-A1) and Asian 2 (N1b-A2) and defined by differences at positions DYS19, DYS391 and DYS439 (Derenko, Malyarchuk, Denisova et al., 2007).

The resolution of the 7-STR profile, however, was not sufficient to fully resolve this classification. This became apparent when examining haplotypes that were previously characterized at a high resolution. Two of the nodes positioned between the largest of N1b-A1 (green) and N1b-A2 (yellow) nodes were a mix of samples that were defined as either N1b-A1 or N1b-A2 in Derenko et al. (2007). This discrepancy probably occurred because DYS19, which was used to define the two clusters by Derenko et al. (2007), was not used in this particular analysis. Populations possessing the European variety were confined to Russia (particularly, northern Russia) and populations from

northwest Siberia (Khanty, Mansi, Komi). The modal haplotype was shared among regions including a significant portion of northwest Siberians and was even found as far away as Turkey (Cinnioglu et al., 2004).

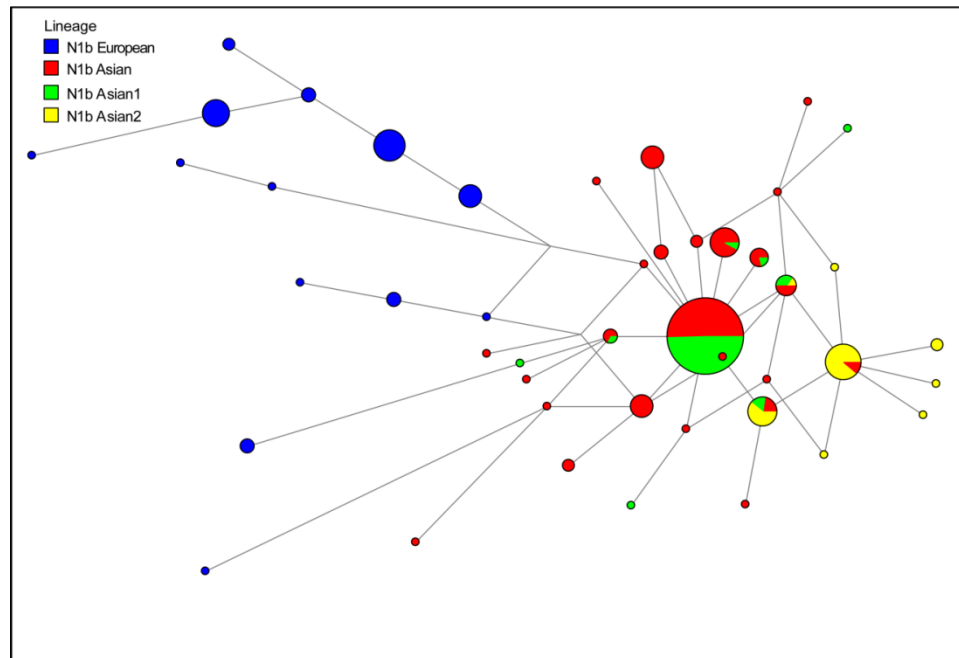


Figure 7.2 RM-MJ network of haplogroup N1b (7-STR)

To attain a higher resolution network, a 10-STR profile was used (Figure 7.3). This profile consisted of the 7-STR profile (DYS389-DYS389b-DYS390-DYS391-DYS392-DYS393-DYS439), plus *DYS19*, *DS437* and *DYS438*. The inclusion of more STR loci in this analysis necessitated the removal of some samples, such that all haplotypes had data for the 10-loci profile. As a result, the following network of N1b lineages contained 201 samples represented by 43 haplotypes.

The first noticeable difference in this 10-loci network was the larger separation between the N1b-A and N1b-E clusters. This separation was due to variation at the



DYS19 locus. Additionally, the distinction between N1b-A1 and N1b-A2 was more apparent. The modal type for N1b-A2 (15-13-16-23-11-14-13-14-10-11) had a three-repeat difference from the modal N1b-A1 haplotype (14-13-16-23-10-14-13-14-10-10). Most of the haplotypes in this network (~68%) were either the modal N1b-A1 type or one-repeat difference from the N1b-A1 motif. N1b-A2 made up the majority of the remaining N1b types (~27%). With the exception of a single Kalmyk sample, the entire N1b-A2 cluster was found in southern Siberia. (The Kalmyk moved from western Mongolia region in the seventeenth century, and therefore can be classified as western Mongolian or south Siberian for all practical purposes (Derenko et al., 2006).) Most of the samples came from Tuva, with six of the samples belonging to Tofalars, who, based on historical records and other genetic data, are closely related with Tuvinians (Derenko et al., 2003; Potapov, 1964e).

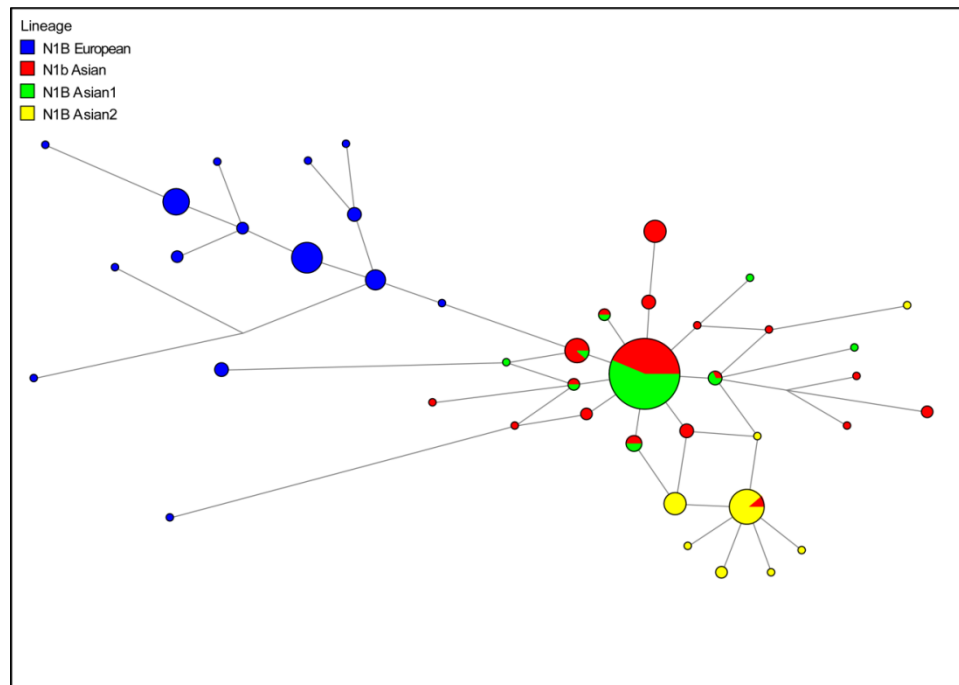


Figure 7.3 RM-MJ network of haplogroup N1b (10-STR)

The two Russians that possessed an N1b-A haplotype originated from the southern portion of the country (Livni and Belgorod) (Balanovsky et al., 2008). One haplotype was shared with a Kalmyk and other belonged to a Kuban Cossack. Only one Russian (from Belgorod) had the modal N1b-A1 haplotype.

Chelkans possessed the modal N1b-A1 haplotype and several lineages with only a single repeat difference from it. Kumandin and Altai-kizhi made up a separate branch, with one or two repeat differences from the modal type. An unspecified Altaian sample (from Rootsi, et al. 2007) shared one of the Kumandin lineages; an Altai-kizhi individual shared the other. The Chelkan and Khakass N1b lineages were relatively diverse and spread out across the network, whereas the Tuvinians, Kumandin and Altai-kizhi haplotypes were confined to fewer branches that were more similar. The Tuvinians also lacked any of the modal haplotype Y-chromosomes. Their lineages were centered on the N1b-A2 modal haplotype described by Derenko et al. (2007). The Shor only possessed the N1b modal haplotype.

It appears that Chelkan and Khakass received their N1b lineages from multiple origins or possibly one single diverse source. The lack of N1b diversity among Kumandin, Shor and Altai-kizhi provided a different pattern, indicating a less diverse source for these lineages. However, the small population sizes for Kumandin and Shor certainly could have played a role in their low diversity. The point about small population sizes also makes the high diversity of Chelkan and Khakass that much more intriguing.

If the Ugric, Yeniseian and Samoyedic-speaking groups did provide the N1b lineages to southern Siberia, then one could expect to see these lineages among their

other descendants (those found to the north). In this case, the Khanty and Mansi do have N1b lineages, but ones that show different haplotype profiles. Asian N1b lineages in the Mansi were largely confined to the modal type (N1b-A1). The Khanty, however, had a more variable set of lineages, one of which was shared with Chelkan, Khakass, and Tuvinian. Another lineage was shared between eight Khanty and one Khakass. Unfortunately, we do not have STR data from Northern Samoyedic groups (Nganasan, Nenets) and, therefore, cannot comment on the variation within these populations or on how they compare to these other groups. We also do not know if they possessed the Asian or European lineages of N1b, and, for this reason, we cannot place them into the Siberian context that is being described here. These data would help to clarify the origin and spread of N1b lineages and could potentially inform us about the genetic nature of historical Samoyedic populations. Despite these limitations, the presence of so many similar N1b haplotypes among Ugric and Samoyedic-speakers and northern Altaians and Khakass leads me to believe that they shared a common paternal source (for this haplogroup).

Finally, a 15-STR profile was used to gain the highest level of resolution with NRY data sets (Figure 7.4). Only a few publications produced data at this level of resolution, although all were used to help refine the relationships that were previously noted using the 7-STR and 10-STR profiles. In this network, 59 haplotypes representing 96 individuals were analyzed via a RM-MJ network. This network included Altaians, Khanty, Russians, Komi and a handful of other individuals from the Volga-Ural, Baltic and southern Siberia regions.

The additional loci helped to resolve the phylogenetic relationships among N1b-A samples to the point where only a single lineage was shared among ethnic groups. One Khakass, one Evenk and two Altaians (one Chelkan and one Altaian of unspecified origin) were possessed this lineage. The new modal type was found only in Khanty.

Many of the samples with the modal N1b-A haplotype in the 10-STR profile lacked the data to make a 15-STR profile. Therefore, those samples were excluded. Of the 73 samples that have the 10-STR profile modal type, only seventeen had the full complement of the 15-STR profile. N1b-A2 was not evident in this network because the majority of samples making up that cluster also lacked the additional five STRs.

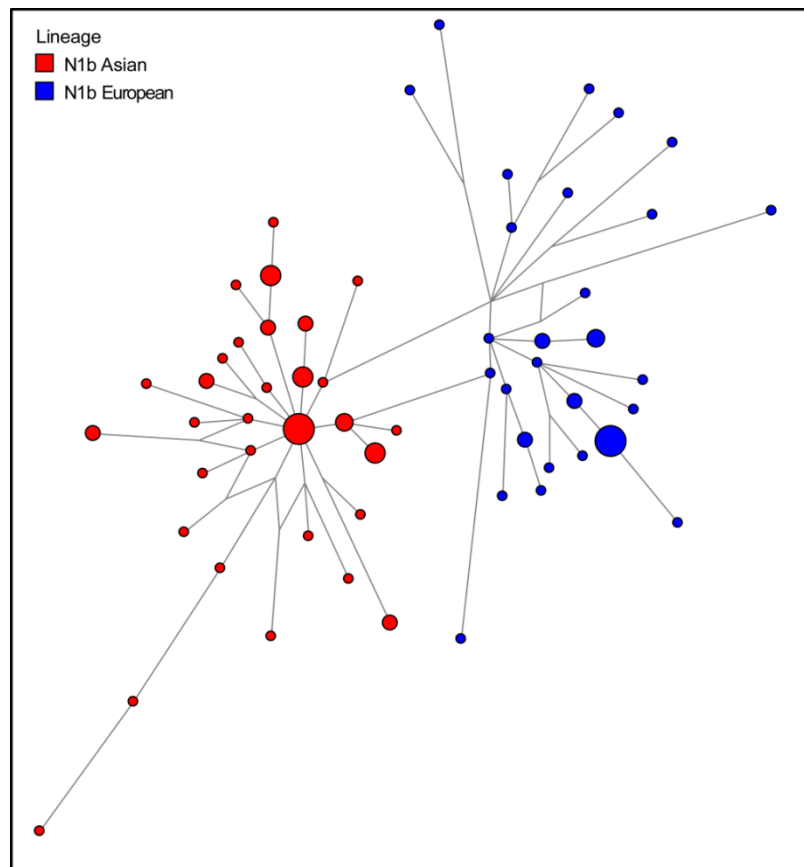


Figure 7.4 RM-MJ network of haplogroup N1b (15-STR)

The median values for all 15 STRs were used as the founder type because there was not a clear founder type between the modal type (which was composed entirely of Khanty) and the most diverse lineage (four ethnic groups represented, but lower overall frequency) (Sengupta et al., 2006). The median haplotype (14-13-16-23-10-14-13-14-10-10-19-15-17-23-12) was one-step away from the most ethnically diverse lineage and two steps away from the modal (Khanty) lineage. Therefore, it seems like a good candidate as a founder haplotype.

On a regional basis, northwest Siberia and the Ural regions of Russia showed the greatest diversity within N1b (Table 7.1). Intrapopulation genetic variance calculations (based on the 7-STR profiles) showed that the Komi (Komi Priluzski, in particular), Mansi, Khanty, and northern Russian populations had the highest levels of variation. Southern Siberians were the next most diverse, although the population level estimates among southern Siberians were quite low. This was also true for Mongolian and Hezhen populations, whose intrapopulation variance was comparable to that of southern Siberian populations, but haplotype diversity was slightly higher. The lowest levels of variation were seen in Evenks (from central Siberia) and Shors (from southern Siberia), each of which having only a single N1b haplotype. Conversely, Turkey had a moderate level of diversity that was slightly higher but similar to that of northwest Siberians.

Overall, N1b had an intrapopulation variance of 0.327, whereas those of N1b-A and N1b-E were 0.113 and 0.183 respectively. These estimates showed that the N1b-E cluster had greater allelic variance than N1b-A, most likely because the populations in which these haplotypes were found are larger in size than their Siberian counterparts, which were affected more by genetic drift and founder events (Rootsi et al., 2007).

Table 7.1 Intrapopulation variances of N1b haplotypes

<b>Region</b>	<b>Population</b>	<b>Number of Samples</b>	<b>Intra-Population Variance</b>
<b>Southern Siberia</b>	<b>All</b>	<b>88</b>	<b>0.113</b>
	Altai-kizhi	4	0.083
	Chelkan	5	0.029
	Kumandin	8	0.031
	Khakass	23	0.067
	Shor	6	0.000
	Tuvinian	28	0.054
	Tofalar	13	0.038
<b>Northwestern Siberia</b>	<b>All</b>	<b>62</b>	<b>0.286</b>
	Khanty	28	0.116
	Mansi	15	0.250
	Komi	18	0.414
	Komi Izhemski	9	0.171
	Komi Priluzski	7	0.544
<b>Central Siberia</b>	Evenk	12	0.000
<b>Mongolia</b>	Mongolian	5	0.071
<b>Northern China</b>	Hezhen	8	0.066
<b>Russia</b>	Northern Russian	24	0.162
<b>Turkey</b>	Turk	15	0.184

It was not a coincidence that the populations with the greatest diversity were the same populations that possessed both N1b-A and N1b-E lineages. Based on the network analysis, the two clusters have probably undergone different population histories. If these developed *in situ*, then we would expect to see intermediate types in addition to those that were present in the clusters. No such haplotypes had been identified to date. Therefore, the clusters should be treated as if they evolved independently, and the founder

haplotypes have been lost. The presence of the two clusters in a single population inflated the allelic variation because the haplotypes were so different from each other when inter-cluster comparisons were made. Separate variance estimates for each of the N1b clusters within the same population showed the overall diversity decreased for those populations when considering each cluster independently. For example, the variance in the Komi population was 0.414, but when the variances for the two clusters were calculated separately, N1b-A was 0.119 and N1b-E was 0.111. The variance in Komi Priluzski decreased to 0.114 for N1b-A when the two N1b-E haplotypes were removed. Similarly, N1b-E decreased to 0.031 in Komi Izhemski once the single N1b-A haplotype was removed. The variance estimates in Mansi were reduced from 0.250 to 0.124 for N1b-E and 0.000 for N1b-A (one haplotype representing nine individuals).

Even once the inflation of variances was controlled for, N1b-E figures were still generally higher than those for N1b-A. They ranged approximately between 0.07 and 0.12, whereas N1b-A variance values were under 0.08. The exception to this pattern was variance in Turkish N1b-A samples, which reached 0.184. Rootsi et al. (2007) tested these Turkish samples for the P43 marker and confirmed that they belong to the N1b-A cluster (Cinnioglu et al., 2004; Rootsi et al., 2007). Several of these haplotypes shared the modal N1b-A haplotype with northern and southern Altaians, Khakass, Tofalars, Khanty, Mansi, Evenk and Shor samples.

The coalescence estimate for the entire haplogroup was about 27.7 ( $\pm$  8.4) kya, using 7-STR profiles and an evolutionary mutation rate. The pedigree mutation rate gave a TMRCA of 10.0 ( $\pm$  3.0) kya. The TMRCA for N1b, however, was 8.8 ( $\pm$  2.1) kya with the evolutionary mutation rate and 3.1 ( $\pm$  0.8) kya with the pedigree rate. These estimates

were just slightly higher than previously published estimates (Derenko, Malyarchuk, Denisova et al., 2007; Rootsi et al., 2007). The two N1b clusters (N1b-A and N1b-E) were relatively similar in age. The European version appeared slightly older than the Asian version, but the standard deviations for these estimates overlapped entirely.

As to the question of which cluster arose first, evidence points to the Asian version being older, lending support for a Siberian source for these Y-chromosomes. The median values for allelic repeats were similar for the N1b-A and N1c branches, further suggesting that the Asian cluster arose first and subsequently gave rise to the European branch (Rootsi et al., 2007). My network analyses also confirmed this supposition, as N1b-A was consistently closer to the root of the network, regardless of STR weight or the presence of diagnostic SNPs. Although the rho statistic for the European cluster was higher, it had a larger standard error associated with it. Consequently, the TMRCAs for each cluster overlapped.

Within my Altaian data set, there were two clusters of N1b Y-chromosomes. One cluster consisted of Chelkan and the other of Kumandin. Using 14-STR profiles, the TMRCAs of these two haplotype clusters were 8.8 ( $\pm$  3.4) kya with the evolutionary rate and 3.2 ( $\pm$  1.2) with the pedigree rate. These estimates were not much different from the TMRCAs calculated for the entire N1b haplogroup. This result suggests that, even though the Chelkan and Kumandin both have N1b Y-chromosomes, these lineages likely did not recently come from the same source. Both haplotype clusters had similar amounts of variation and TMRCAs of roughly equal age. For the Chelkan, it was 1040 ( $\pm$  730) or 370 ( $\pm$  260) years ago, depending on mutation rate used. For the Kumandin, it



was 1290 ( $\pm$  910) or 470 ( $\pm$  330) years ago (again, depending on the mutation rate used). These estimates were also confirmed with BATWING (data not shown).

Two possible scenarios were invoked to explain the current distribution of N1b haplotypes. The first involved the origin of N1b (specifically, N1b-A) in Siberia with expansions into Europe resulting in the N1b-E cluster. Such a migration would have occurred from Siberia or the steppe regions of Kazakhstan and southern Russia/Ukraine to Turkey (as evidenced by the presence of N1b-A there). This most likely occurred through Central Asia, particularly through the steppe region and not through European Russia. There was no evidence that N1b was present in Iran, Afghanistan or Pakistan, thus making a route through what is a mostly Turkic-speaking region between European Russia and the Indo-European speaking regions south of the Caspian and Aral Seas. This scenario would also explain the pattern of variation seen in Russia, as the northern populations were made up exclusively of N1b-E lineages whereas the N1b-A lineages were found only in the south, albeit in very low frequencies. These lineages could be remnants of a migration that ended in Anatolia and Eastern Europe. Whether this distribution can be tied to Turkic language expansions remains to be seen.

The second scenario involves the origin of haplogroup N1b in the Ural region, with a later expansion into southern Siberia (Mirabal et al., 2009). This hypothesis was based solely on the intrapopulation variance calculated for the northwestern populations. According to this scenario, from the time this haplogroup arose in the Urals, the lineages diverged within the populations in that region. One set of these divergent lineages then moved west into Europe, while the other lineages migrated south, thus providing the current distribution of the haplogroup. This scenario does not seem likely, as there were

no intermediate lineages between these two clusters among the populations of the Urals. In addition, if both clusters developed next to one another in these populations, then it is not clear why only one of each moved in different directions. It would seem logical that both clusters would be represented if lineages were contributed from the source populations of this haplogroup.

The first scenario is the most probable based on my network analysis. Populations of northwest Siberia, Komi, Mansi and Khanty, would be the result of an amalgam of western and eastern branches coming together and interacting in northwestern Siberia. This interpretation suggests that these clusters arose independently from each other in east and west and later re-engaged through trade/migration/commerce in the northwestern Siberian region (Pimenoff et al., 2008).

To summarize, N1b was found most frequently among southern Siberians, northwestern Siberians and Russians. Two clusters of N1b lineages showed the same relative amounts of diversity, and therefore, were similar in age. Rho statistics and Bayesian analyses of N1b lineages gave coalescence estimates that would place their origin in either the Neolithic (evolutionary rate) or Bronze Age (pedigree rate). At least two separate lineages diverged from a common ancestor, most likely, similar to the versions found in Siberia. While one hypothesis suggests an origin for these clusters among northwestern Siberians, it seems more likely that the lineages evolved independently and were geographically isolated from each other on either side of the Urals. Later, haplotypes from these clusters came into contact in the populations located between their points of origin (i.e., northwestern Siberia). The relatively young age of this haplogroup is significant as it occurs after the founders of Native Americans went to

New World. The occurrence of N1b among southern Siberians (particularly, Chelkan, Kumandin and Khakass) can thus be viewed as evidence that they derive from the historical indigenous populations of the Yenisei and Altai-Sayan regions of Siberia (the Ugric, Yeniseian and Samoyedic populations of historical record), who also helped to form the Samoyedic populations now found further to the north (Nenets, Ket, Evenk).

### **7.3 Haplogroup R1**

NR1Y haplogroup R1, which is defined by the M173 marker, is ubiquitous throughout Eurasia. Studies first showed contrasting clines of two haplogroups with M173 – R1a and R1 (xR1a) (Semino et al., 2000). These two haplogroups evolved separately and followed different trajectories resulting in higher frequencies of R1a in eastern European populations and higher frequencies of R1 (xR1a) in western European populations. Subsequently, a marker that differentiated most of the R1 (xR1a) Y-chromosomes was discovered, resulting in a change in nomenclature from R1 (xR1a) to R1b. Using the current NR1Y nomenclature, these haplogroups are now known as R1a1a and R1b1b2 and are among the most common NR1Y haplogroups in Eurasia (ISOGG, 2010; Karafet et al., 2008; Underhill et al., 2009).

Coalescence estimates for haplogroup R range from 12 to 21 kya and 35 to 40 kya, depending on the publication (Hammer & Zegura, 2002; Semino et al., 2000). A more recent calculation puts the coalescent time for haplogroup R at 26.8 [19.9 – 34.3] kya and 18.5 [12.5 – 25.7] kya for R1 (Karafet et al., 2008). These dates place the origin of the haplogroup within the Upper Paleolithic of Eurasia and around the time of the LGM. Given the current distribution of R1a1a and R1b1b2, the parallel expansion of

these two sister haplogroups fits nicely with a hypothesis about the founding R1 haplogroup, where it split during the LGM and evolved independently in different glacial refugia (Semino et al., 2000). Once the ice sheets began melting and the climate changed, populations expanded and took these two haplogroups to different parts of Eurasia.

R1a1a seemingly dispersed from southern Ukraine and populated the steppe belts of Eurasia (Passarino et al., 2001; Semino et al., 2000). R1a1a has also been associated with the Kurgan culture in Eastern Europe and Central Asia by some who argue that the current distribution of R1a1a Y-chromosomes was directly the result of expansions of Indo-Iranian language speakers. This expansion was helped along by the domestication of the horse (Passarino et al., 2001; Quintana-Murci, Krausz, Zerjal et al., 2001; Wells et al., 2001). However, Quintana-Murci et al. (2001) noted a relatively higher diversity level of R1a1a in Indian populations and related this to a larger, more diverse population of R1a1a Y-chromosomes moving into the region.

By contrast, R1b1b2 (Hg1 in Rosser et al. 2000) took a different path, spreading throughout Western Europe (Rosser et al., 2000). R1b1b2 has been associated both with Cavalli-Sforza et al.'s first principal component and with the Aurignacian of Europe (Rosser et al., 2000; Semino et al., 2000).

### 7.3.1 Haplogroup R1a1a

R1a1a occurs at its highest frequencies among Central Asian populations, but is also commonly found among populations in Eastern Europe (particularly, the Baltic States). It also appears in India, Pakistan, Nepal, the Xinjiang region of China, Mongolia,

southern and western Russia, with its greatest numbers being found in the Altai, Kyrgyz and Tajik (Wells et al., 2001; Zerjal et al., 2002). Given that Altaian populations also exhibited a high frequency of R1a1a\* lineages, an analysis of current microsatellite variation was conducted to evaluate how the Altaian lineages fit into the broader picture of R1a1a diversity worldwide. STR data were collected from published literature and combined with the R1a1a\* lineages that I characterized in this dissertation. A 7-loci STR profile allowed for the widest comparisons across Eurasia (DYS19-DYS389I-DYS389b-DYS390-DYS391-DYS392-DYS393). Intrapopulation variance was calculated for each population and for each geographic region (Table 7.2).

In addition, networks were generated using the 7-STR profiles. A total of 1,224 R1a1a Y-chromosomes were compared in an RM-MJ network (Figure 7.5). This network consisted of 261 distinct haplotypes, representing populations from southern, northwestern and central Siberia, northern China, Mongolia, Russia, the Baltic region, India, Pakistan, Tibet and Turkey. Unfortunately, the network contained many high-dimensional cycles, forming reticulations throughout the core of the network, because many of the 7-STR profile haplotypes were shared among populations, and consequently, lacked any definition. A second network was created using the “>1 frequency” option in Network v4.5.1.6 to reduce the complexity of the network, although this did not result in cleaner results (data not shown). More STR loci were needed to provide greater resolution to the network.

Table 7.2 Intrapopulation variances for R1a1a haplotypes

<b>Region</b>	<b>Population</b>	<b>Number of Samples</b>	<b>Intra-Population Variance</b>
<b>Southern Siberia</b>	<b>All</b>	<b>226</b>	<b>0.242</b>
	Chelkan	4	0.214
	Tubalar	10	0.271
	Tuvinian	15	0.184
	Khakass	18	0.180
	Teleut	31	0.182
	Shor	23	0.155
	Altai-kizhi	95	0.132
	Altai-kizhi Dulik	60	0.145
	Altai-kizhi Derenko	35	0.113
<b>Central Siberia</b>	<b>All</b>	<b>19</b>	<b>0.225</b>
	Buryat	5	0.200
	Soyot	7	0.034
	Evenk	7	0.306
<b>India</b>	<b>All</b>	<b>114</b>	<b>0.356</b>
	Northern	32	0.339
	Eastern	19	0.309
	Southern	39	0.328
<b>Russia</b>	<b>All</b>	<b>432</b>	<b>0.245</b>
	Northern	33	0.341
	Central	180	0.236
	Southern	98	0.224
<b>Baltic</b>	<b>All</b>	<b>254</b>	<b>0.297</b>
	Finnish	38	0.353
	Eastern Finn	18	0.446
	Western Finn	20	0.267
	Estonian	44	0.394
	Latvian	44	0.259
	Lithuanian	56	0.231
	Karelian	33	0.244
	Swede	39	0.285

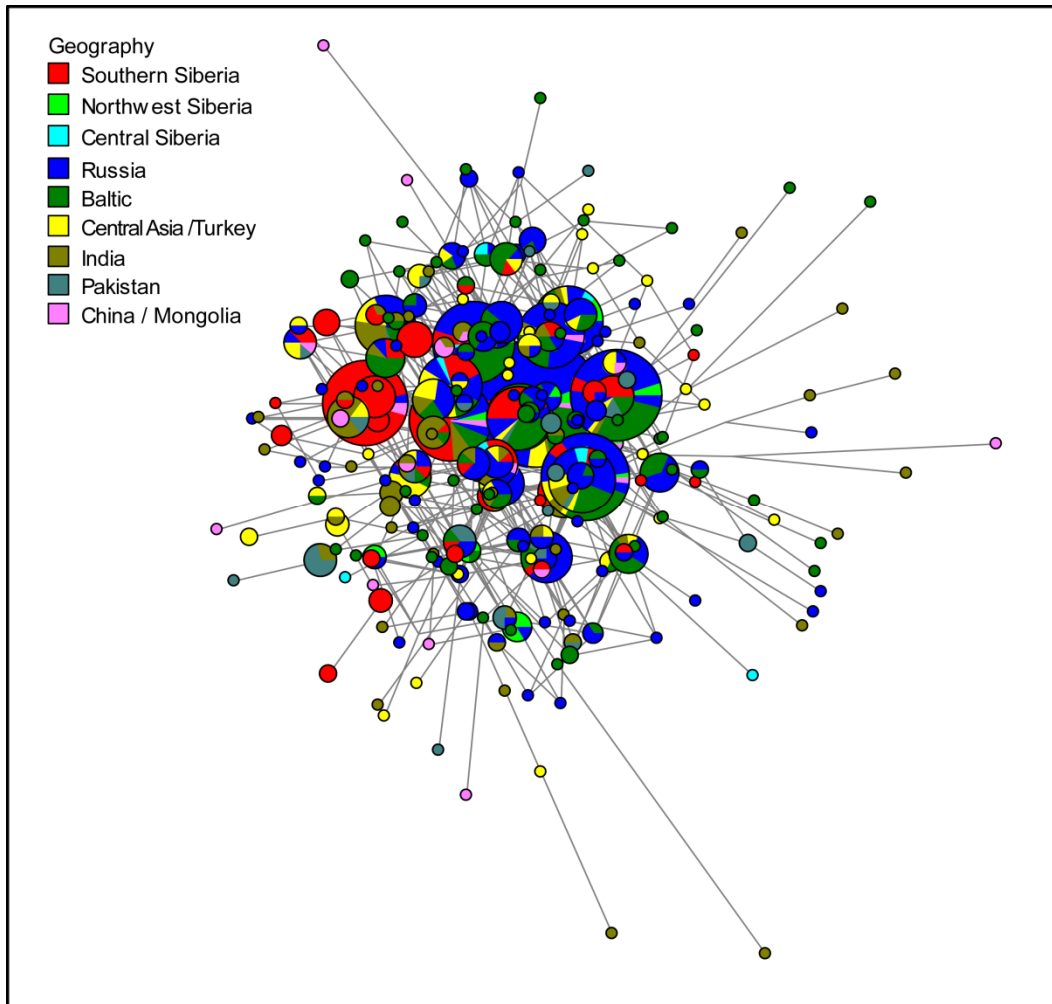


Figure 7.5 RM-MJ network of haplogroup R1a1a (7-STR)

By expanding the set of STRs to a 15-STR profile, the Baltic, Turkish, and Indian samples were removed from the analysis, thus confining the analysis of R1a1a found mostly in southern Siberia, Russia, and Nepal (Figure 7.6). The resolution of this network split many of the nodes that were shared among regions. The only groups having a high degree of haplotype-sharing at this level were the Komi and Russians. For the Altaians, a single Kumandin R1a1a\* matched a Russian from Penzenskaja, and an Altai-kizhi matched a Russian from Tver.

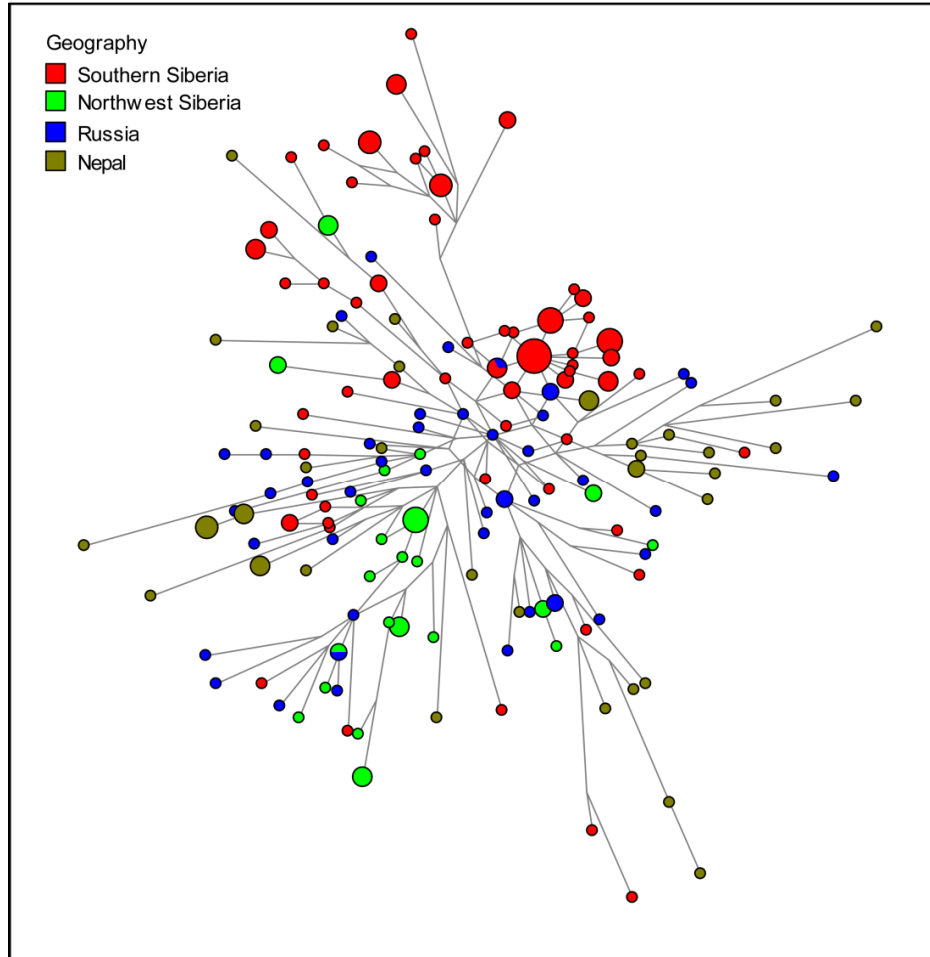


Figure 7.6 RM-MJ network of haplogroup R1a1a (15-STR)

As noted above, Semino et al. (2000) concluded that the two sister haplogroups (R1a1a and R1b1b2) expanded from the Ukraine and Iberian Peninsula, following the LGM. They asserted a Paleolithic origin for the haplogroups, but suggested that the current distribution of R1a1a could be the result of the expansion of the Kurgan culture (Semino et al., 2000). Others have argued for a Kurgan influence more forcefully (Passarino et al., 2001; Quintana-Murci, Krausz, Zerjal et al., 2001; Wells et al., 2001). Nevertheless, in all of these networks, no haplotype cluster showed the same high



frequency/low diversity among regions or ethnic groups, nor was there any identifiable cluster like the Genghis Khan C3\* cluster (see the C3\* section below), as might be expected if a small Kurgan population was responsible for all of the R1a1a lineages in the Altai and southern Siberia.

Other studies focusing on questions of Indian origins viewed these data differently (Kivisild et al., 2003; Sengupta et al., 2006). They argued that the differentiation of R into R1 and R2 occurred in southern or western Asia, given that the presence of R1\* (precursor to R1a1a and R1b1b2) and R2 were both found in India and Pakistan. In addition, Central Asian populations were reported to have lower STR variance estimates, indicating that India, rather than Central Asia, was the most likely source of the R1a1a Y-chromosomes (Kivisild et al., 2003; Sengupta et al., 2006; Zerjal et al., 2002).

My own analysis of STR variance verified these findings (Table 7.2). Sengupta et al. (2006) suggested a second possibility -- that multiple migrations into northwestern India could have increased the variance of R1a1a there. Critical to this question of R1a1a origins is the variance of these Y-chromosomes in Central Asia and Ukraine (the purported source of this haplogroup). Unfortunately, extensive data from Central Asia were not available for direct comparison. The loci used in the Ukraine study also vary substantially from the standard STR data utilized today (Passarino et al., 2001). Therefore, a direct comparison of the Ukrainian data to ours and to those from published literature is not currently possible.

One interesting point about the lineages in southern Siberians and Indians is that the coalescence ages for these lineages were much too old to be explained by a Kurgan

source. Seemingly, R1a1a Y-chromosomes were already in southern Siberia when the Kurgan cultures filtered to the east. In fact, the age of the haplogroup was old enough to suggest that it was present during the re-expansion of human populations in southern Siberia after the LGM. This was likely the case for India as well (Sengupta et al., 2006). It would therefore not be surprising if peoples carrying the Kurgan culture east also carried with them R1a1a Y-chromosomes. The problem, however, is that those Kurgan R1a1a Y-chromosomes might not necessarily be distinguishable from those already present in southern Siberia (or India). Thus, the signals of the original movement into southern Siberia and the movement of Kurgan peoples may have merged with each other, and as a result, are no longer identifiable. As a result, the intrapopulation variances are quite large and display diversity estimates more consistent with the original expansion of the haplogroup. In addition, subsequent migrations and nomadic pastoralist activities across the Eurasian steppe certainly helped to redistribute these lineages across a broad geographic region.

The primary problem with the phylogeography of this haplogroup is the lack of SNP-defined branches (Sengupta et al., 2006; Underhill et al., 2009). The relatively low resolution of individual (or region-specific) clades is obvious in the above-mentioned network analyses. Such markers would help to subdivide the R1a1a lineages into smaller more informative branches, many of which would likely be region specific. These SNPs would also serve as starting points (a zero variance point) where the accumulation of the current STR diversity would allow for estimates of when a particular branch arose and spread.

Since conducting the analysis above, a new study was published on this very issue (Underhill et al., 2009). In this study, over 2,000 R1a1a samples were sequenced for additional SNPs to help differentiate the haplogroup's phylogeny. In the process, a new marker called M458 was discovered. M458 was observed in many populations throughout Central and Eastern Europe. It occurred at its highest frequency in Poland and then decreased in frequency towards southwestern Russia. This marker was also absent east of the Urals, thus providing the first evidence of a difference in European versus Asian R1a1a Y-chromosomes. This new haplogroup (R1a1a7\*) was dated to about 7.9 ( $\pm$  2.6) kya, but the confidence intervals were large enough that it cannot be labeled as part of any particular cultural complex, although it could represent the dispersal of Proto-Indo-European speakers.

Looking at all of the R1a1a\* (x M458) Y-chromosomes, Indian haplotypes had the most diversity followed by those from the Altai and Central Asia (Kyrgyzstan). These estimates corroborated the results of my analysis above. Therefore, the hypothesis that the Kurgan cultures are the source of all R1a1a Y-chromosomes in the Altai is unlikely. Almost as important as finding the M458 marker, Underhill et al. (2009) noted that there was no obvious haplotype cluster associated with this SNP. Therefore, R1a1a7\* STR haplotypes could not be easily differentiated without the use of the M458 marker. Any attempt at identifying R1a1a7 Y-chromosomes without characterizing this SNP is not possible. Thus, studies using only STR data could likely miss the existence of phylogenetic clades due to saturation of mutations at the different STR loci. This point helps to explain the poor resolution of the R1a1a 7-STR network analyses.

Network analysis of the Altaian R1a1a Y-chromosomes was undertaken to assess whether patterns of haplotype clusters could be identified, and if so, when these clusters originated. This analysis uncovered several haplotype clusters among the R1a1a Y-chromosomes (Figure 7.7). Southern Altaians had two distinct clusters. The first cluster ( $\alpha 1$ ) was the largest and included only Altai-kizhi individuals. The TMRCA for this cluster was estimated at 3.1 ( $\pm 1.1$ ) kya with the evolutionary mutation rate and was confirmed using BATWING. A small cluster branched off this main one and included

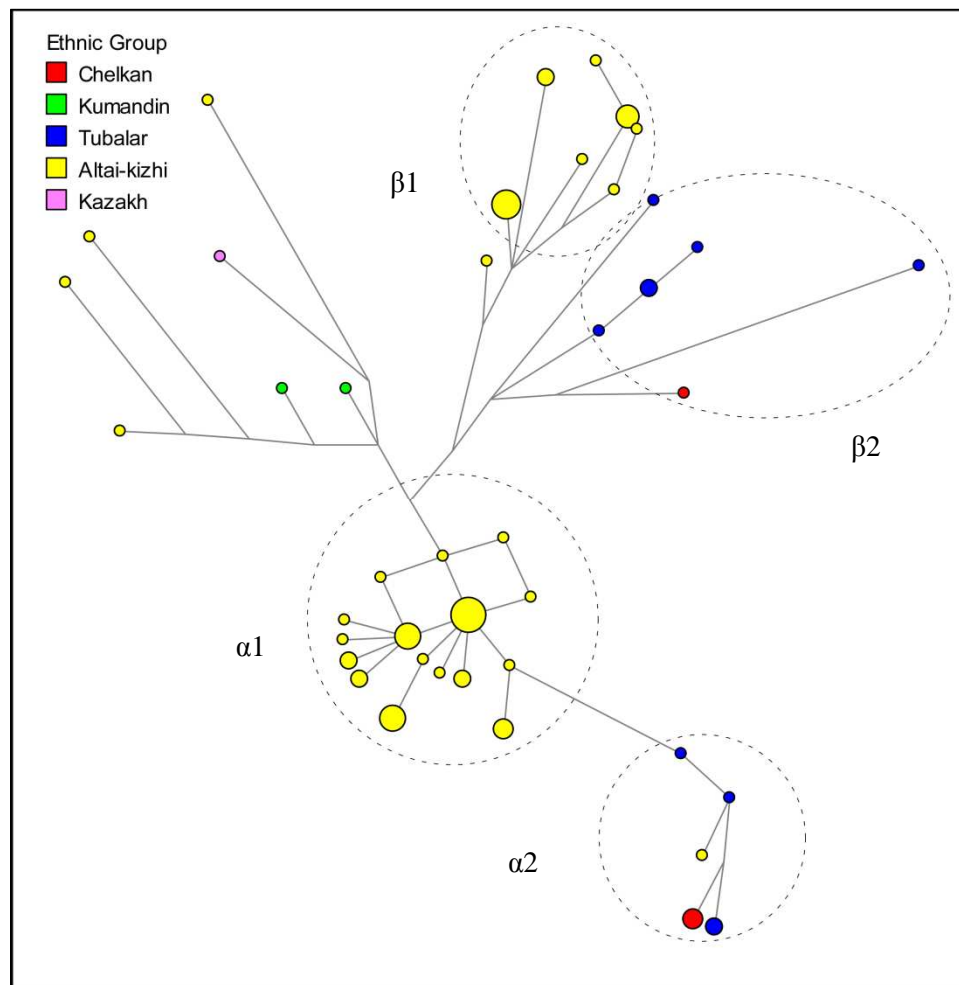


Figure 7.7 RM-MJ network of Altaian haplogroup R1a1a (15-STR)

seven northern Altaians and one southern Altaian ( $\alpha 2$ ). The TMRCA of this cluster was  $3.2 (\pm 1.6)$  kya – essentially the same as the larger cluster.

These dates place the origin of these clusters somewhere in the Eneolithic to the Iron Age in Siberia. These periods are when the Afanasievo and Andronovo cultures were introduced to southern Siberia, which could certainly have brought these lineages from the west. It is also possible that these lineages were indigenous to southern Siberia and began expanding only after the introduction of pastoralism to the Altai by the newly arriving steppe cultures.

The same haplotypes were found in the Kyrgyz (Underhill et al., 2009), but there is some question as to the origin of these people (Golden, 1992). One hypothesis is that they originated in southern Siberia (part of the former Yenisei Kyrgyz). A second hypothesis is that the ethnic group developed out of the Kipchak tribes inhabiting the Central Asian steppe, much like the Kazakhs. A third invokes some combination of the two ideas. The evidence based on 15-STR profiles shows a close association between the Altai-kizhi and Kyrgyz.

Very similar haplotypes were also found in several ancient DNA samples from the Andronovo and Tagar period kurgans in southern Siberia (Keyser et al., 2009). In fact, a number of the haplotypes recovered from these kurgans matched ones from southern Siberian populations (Altai-kizhi, Tuvinian, Shor and Khakass). Because the comparisons were made with only 7-STR profiles, these associations may not be maintained with higher resolution data. Nevertheless, this study provides circumstantial evidence that this haplotype cluster originated from lineages present in southern Siberia during the Bronze and Iron Ages.

The second haplotype cluster ( $\beta 1$ ) for the southern Altaians was dated to 6.4 ( $\pm$  2.4) kya. A second cluster of northern Altaian haplotypes ( $\beta 2$ ) was located near this second southern Altaian cluster, but it was more variable than the southern version. The TMRCA for the northern cluster was 9.6 ( $\pm$  2.8) kya. In both cases, these clusters appeared to have Neolithic origins. The remaining haplotypes did not form distinct clusters.

Based on the network analysis, the R1a1a Y-chromosomes revealed evidence for at least three separate origins. The oldest and hardest to actually identify was represented by the diffuse lineages that did not show any great affinity with other R1a1a lineages. These Y-chromosomes were found in both northern and southern Altaians. The second set of lineages came from two clusters (one of northern Altaians and the other of southern Altaians) that dated to the Neolithic. The final group, which was the largest of the R1a1a clusters, dated to the Eneolithic – Iron Age and was likely associated with cultures originating in the steppe lands. Therefore, not all of the Altaian R1a1a Y-chromosomes can be attested to the arrival of peoples carrying Kurgan cultures, whereas a significant portion of the southern Altaians can.

### 7.3.2 Haplogroup R1b1b1

R1a1a was not the only notable R-derived haplogroup in Altaians. R1b1b1 is a sister haplogroup to R1b1b2, with both haplogroups sharing the P297 polymorphism (Karafet et al., 2008). Unlike R1b1b2, haplogroup R1b1b1 is not found in Europe. Its distribution is centered in Central Asia. M73, the marker defining the R1b1b1 branch, was first published in 2000 and was found in six individuals from “Central Asia/Siberia”

(Underhill et al., 2000). In previous work, Underhill et al (1997) had indicated that they analyzed individuals from various Central Asian groups, including 9 Khorezmian Uzbek, 8 Tajik, 9 Kirghiz, 8 Turkmen, 8 Dungan, 7 Uighur, 9 Kazakhs, and 2 Arabs. However, they did not list the populations from Siberia that they studied. As a result, it is unclear from these publications where specifically the M73 derived samples originated, only that they came from Central Asian and/or Siberian populations.

Studies with a focus on Central Asian, Indian and Siberian populations have sporadically characterized this marker. It was tested in Russia, but only a single Russian sample possessed the marker (Balanovsky et al., 2008). It was also screened for but not found in India or Nepal (Gayden et al., 2007; Regueiro, Cadenas, Gayden, Underhill, & Herrera, 2006; Zerjal et al., 2007). Among populations located south of the Eurasian steppe, it was only found in northern Pakistan, being absent from southern Pakistan, Iran, Turkey and Oman (Regueiro et al., 2006). Sengupta et al. (2006) found M73 in the Hazara populations of Pakistan as well as the Naxi and Uyghur populations in northwestern China. Single representatives were also found in Tu, Mongolian, Han, and Japanese populations (Sengupta et al., 2006). In addition, M73 was found in the Karachay, Megreles and Kabardians in the Caucasus and Tatars and Bashkirs in the circum-Ural region of Russia (Underhill et al., 2009). The highest frequency of M73 was found in a population of Bashkirs in southeastern Bashkirostan, Russia (Myres et al., 2010) and in Balkars of the northwestern Caucasus (Underhill et al., 2009).

Many of the other studies with Central Asians lack sufficient SNP resolution to determine whether M73 exists among sampled populations. M173 (xM17) is found in moderate frequencies (0.08 - 0.13) among Caucasian and Central Asian populations

(Wells et al., 2001). This category would include R1b haplotypes, but based on the resolution of these data, there is no way to differentiate between the R1b1\*-P25, R1b1b1-M73, or R1b1b2-M269 haplogroups. As noted previously, P25 and M269 are present in high frequencies in Western Europe (Cruciani et al., 2002), and thus, M73 may have a different distribution than those markers.

R1 (xM17) makes a significant contribution to several populations. In particular, it is found in Uzbeks, Kurds living in Uzbekistan, Yagnobi of Tajikstan, Armenians and Turkmen (0.21, 0.29, 0.32, 0.36, and 0.37, respectively) (Wells et al., 2001). Central Asian (Kyrgyz, Kazakh, Uyghur, Uzbek, Tajik, Turkmen, Dungan) and Caucasian (Kurd, Georgian, Ossetian, Lezgi, Azeri, and Armenian) populations tested by Zerjal et al. (2002) were also classified as P (xR1a). However, there were insufficient data to determine whether these individuals belonged to Q or R, and, if they did belong to R, whether they also possess M269 or M73 SNPs.

R1b1b1 lineages were analyzed to understand the origin of these Y-chromosomes in Altaian populations, where they were found in moderate frequencies among the Kumandins (Figure 7.8). Only 59 samples containing STR data and confirmed to belong to the M73 haplogroup were available for analysis. Because there were so few data published on M73, I searched the YHRD database using 7-STR profiles of the Altaian haplotypes to help understand its worldwide distribution. Of those M73 haplotypes, one haplotype was shared among several populations, including 4 Kazakhs, 1 Altai-kizhi, 1 Tuvinian, 1 Tibetan, 1 Kalmyk, 1 Turk, 2 Russians and 1 Romanian. A two-step variant haplotype was found among Kumandins. This haplotype was not shared with any other population, but a one-step variant was found in Poland and a two-step variant was found



in Tuva, both of which were different from the modal type. A second Kumandin haplotype matched ones seen in a Kazakh and a Ukrainian.

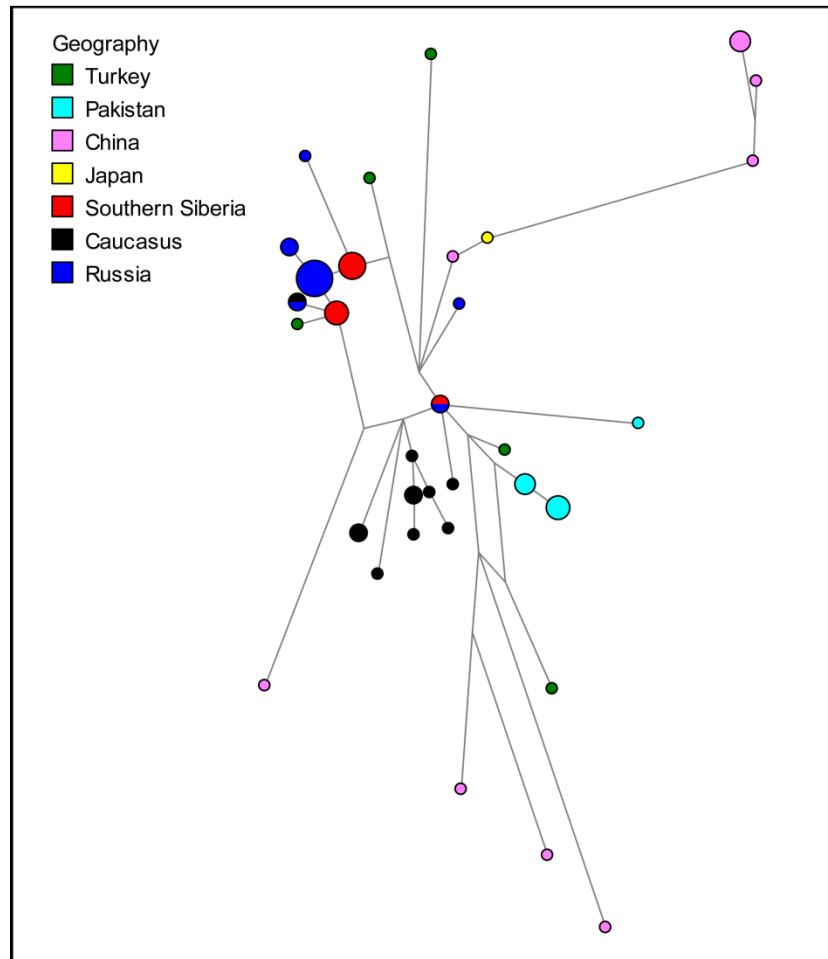


Figure 7.8 RM-MJ network for haplogroup R1b1b1 (8-STR)

The network created using an 8-STR profile provided a TMRCA for the entire haplogroup of  $26.8 (\pm 4.5)$  kya or  $9.7 (\pm 1.6)$  kya, depending on mutation rate used. This network showed two separate clusters of R1b1b1 lineages. One cluster was noted by Myers et al. (2010) and consisted of relatively short repeat scores at one locus (DYS390 = 19). This cluster is of particular interest because most of the Kumandin lineages fell

into this set of haplotypes. Along with the Kumandins were Y-chromosomes from Bashkirs and Tatars. The TMRCA for this cluster was 4530 ( $\pm$  2290) years ago with the evolutionary rate and 1630 ( $\pm$  830) years ago with the pedigree rate. In either case, the TMRCA indicates a recent historical common ancestor for these lineages. When the 14-STR profiles were used with only samples from my data set, this cluster dated to 3880 ( $\pm$  1770) years ago, or 1398 ( $\pm$  640) years ago depending on the mutation rate.

Not surprisingly, the variance was greatest among the Turkish population, which has two lineages with affinities to the southern Siberian and Kazakh lineages and two other more distinctive haplotypes (Table 7.3). The variance of Uyghur lineages (although there were only 3) was also relatively high. The Kumandin and Hazara had moderate variance levels (0.262 and 0.145, respectively) and were consistent with a larger sample of closely related haplotypes, plus an outlier. The Kazakh haplotypes were identical, and thus, no variation was present. These estimates generally are indicative of a Central Asian origin for this haplogroup.

Table 7.3 Intrapopulation variance for R1b1b1 haplotypes

<b>Region</b>	<b>Population</b>	<b>Number of Samples</b>	<b>Intra-Population Variance</b>
<b>Southern Siberia</b>	Kumandin	6	0.262
	Teleut*	11	0.127
	Altaian Kazakh	3	0.000
<b>Central Asia</b>	Uyghur	3	0.762
	Hazara	8	0.145
<b>China</b>	Naxi	4	0.071
<b>Turkey</b>	Turk	4	1.524
<b>Caucasus</b>	Balkar	9	0.099
<b>Circum-Urals</b>	Bashkir	10	0.057

\*Note: Teleuts were not tested for M73 (Kharkov et al., 2009). These are R1b (xM269) Y-chromosomes that have haplotypes very similar to R1b1b1 haplotypes. One or more of these could lack the M73 marker.

The TMRCA estimate for the entire haplogroup was similar to that estimated for R1a1a. Given the distribution of these lineages among populations located in or historically associated with Central Asia, it seems likely that this lineage moved across the Eurasian steppe after the LGM, possibly along with R1a1a. This also means that the hypothesis claiming all R1b Y-chromosomes expanded out of the Iberian refugium cannot explain the current distributions of all R1b lineages.

### **7.3 Haplogroup Q1a3\***

Haplogroup Q plays a critical role in understanding the population histories of groups in Central Asia, Siberia and the Americas. M242, the marker defining this haplogroup, was first described by Seielstad et al. (2003), who studied its distribution in Central Asian populations. Prior to the discovery of this marker, M3 was the only Q-derived marker tested in population studies. M3 was characterized in the majority of Native American Y-chromosomes and was tested widely to find associations between New and Old World populations (Bolnick, Bolnick, & Smith, 2006; Bortolini et al., 2002; Bortolini et al., 2003; Karafet et al., 1999; Lell et al., 2002; Malhi et al., 2008; Zegura et al., 2004).

Additionally, NRY markers 92R7 or M45 were tested to identify haplogroup P Y-chromosomes (of which M3 is a derived branch). Microsatellite analysis of haplogroup P provided evidence of at least two haplotype clusters with different distributions in Siberia, Mongolia and the Americas (Lell et al., 2002). These data were used to infer the presence of multiple migrations into the New World from Siberia. Some of those M45 lineages (“M45a”) in Siberia and the Americas almost certainly possess the M242

marker, even though it was not tested in that study. Thus, characterizing the distribution of M242 and M242-derived Y-chromosomes is essential for understanding how the New World was inhabited and from where the ancestral American populations may have originated. Despite this evidence, few studies test for the presence of M242, opting instead for a “P (xR1a1a)” category.

The discovery of M242 allowed for several of the previously defined markers (M25, M120, and M3) to be placed in their proper phylogenetic positions. Several studies continued to explore the Q-derived Y-chromosomes, providing a framework of major Q branches for which the complex nature of the haplogroup is becoming better understood (Sharma, Rai, Bhat, Bhanwer, & Bamezai, 2007; Shen et al., 2004). These branches were located mostly in Central Asia, India, and Pakistan. One branch (M356) was found in the Middle East. Another branch was defined by M346. It was first published in Karafet et al. (2008), and was found to be ancestral to the M3 marker that is ubiquitous in the Americas. The current distribution of M346 is unknown. This dissertation marks the first time that M346 was characterized in Siberian populations, providing a direct link between the Q lineages in the Old World with those of the Americas.

The complete distribution of M346 is unknown. Unfortunately, to complicate matters further, we do not even have full knowledge of where haplogroup Q Y-chromosomes are located or the diversity of Q lineages within populations, let alone the distribution of this newly found marker. To help clarify this issue, I analyzed Q derived STR data obtained through published literature in addition to those generated in this dissertation.

Three RM-MJ networks were created to investigate the relationships between haplogroup Q lineages. The first network incorporated 122 5-STR profile (DYS19-DYS390-DYS391-DYS392-DYS393) haplotypes from 510 samples (Figure 7.9). This network included Siberians, Central Asians, Indian, Pakistani, Turks and Native Americans. Because so few STRs were used in these studies, many of the southern Siberian lineages matched those found in the Americas. There was also sharing between Central Asian and American haplotypes, but this was infrequent. The two largest nodes of the 5-STR network were made up of Q1a3\* samples from the Altai and Q and Q1a3a samples from the Americas. No distinct haplotype clusters were associated exclusively with any subbranch of haplogroup Q or with any geographic region. Ultimately, the resolution was far too low to make any conclusive statements about the phylogeography of this haplogroup.

The second network consisted of 329 samples represented by 174 8-STR profile haplotypes (DYS19-DYS389I-DYS389b-DYS390-DYS391-DYS392-DYS393-DYS439) (Figure 7.10). There was considerably more haplotypic resolution in this network, with very little haplotype sharing among regions. The only haplotype sharing occurred at two nodes. In the first instance, two Altai-kizhi individuals shared a Y-chromosome haplotype with an Apache. At the second node, two Altai-kizhi shared a haplotype with one Tanana individual. Despite the lack of overall haplotype sharing, the cluster of southern Siberians was close to many of the American haplotypes, in some cases with only a single repeat difference between them.

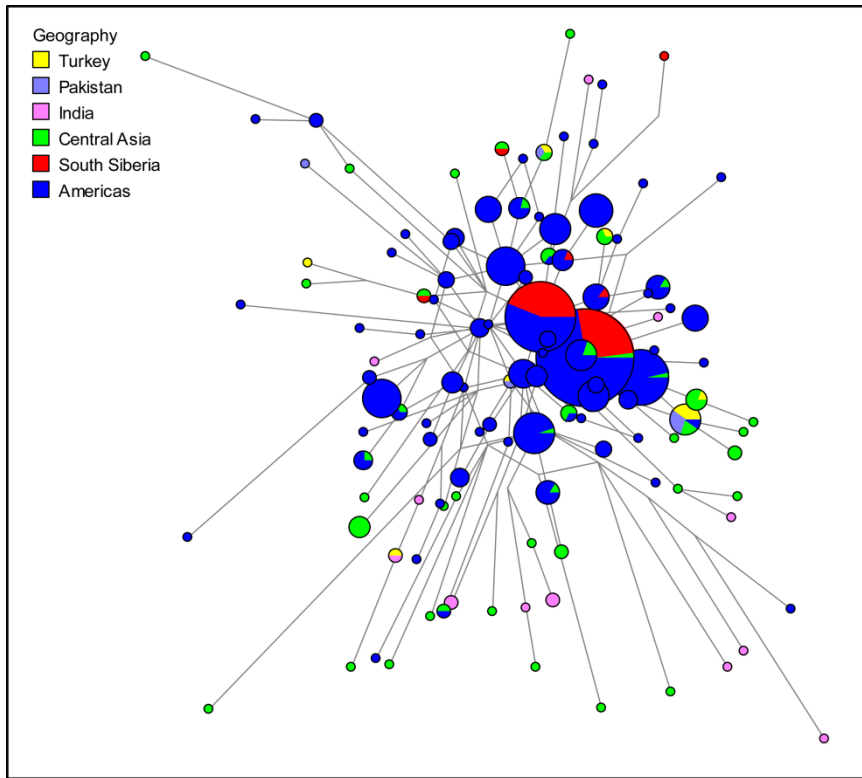


Figure 7.9 RM-MJ network of haplogroup Q (5-STR)

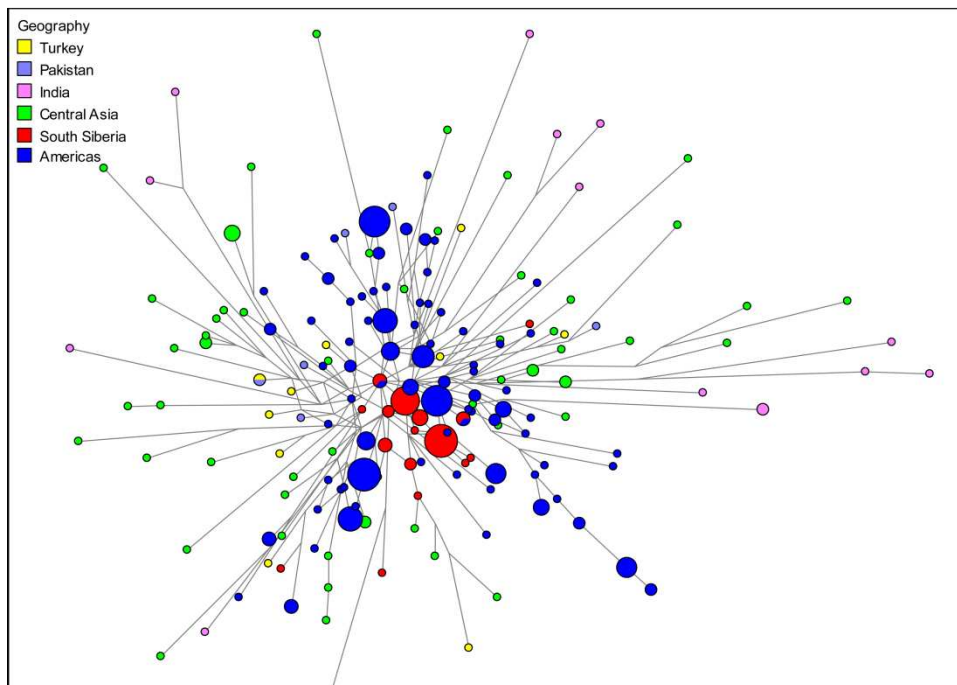


Figure 7.10 RM-MJ network of haplogroup Q (8-STR)

Haplogroup Q in southern Siberia was intriguing. The northern and southern Altaians did not share a single lineage, and the haplotypes found in the southern Altaians were more similar to those of Tuvinians. Affinities between southern Siberian and at least some of the American Q lineages were also noted in these networks. This relationship was confirmed with the presence of the M346 marker in the Americas (Bailliet et al., 2009). Q-M242 has been described in American populations (Bortolini et al. 2003; Malhi et al. 2008), but it is still not clear whether all Q haplotypes that lack the M3 marker belong to the M346 branch. Regardless, M346 is a good candidate for the “M45a” cluster described by Lell et al. (2002).

Bortolini et al. (2003) showed three different modal haplotypes for each of the Mongolians, Chipewyan and Amerindian groups they analyzed. Although those of the Chipewyan appeared to be unique, the Q lineage for Mongolians matched one characterized in the Altai-kizhi, while the modal haplotype for Amerindians was found among Chelkan and Tubalar Q1a3\* lineages. This is not to suggest that northern Altaians are the source population for Amerindians, but rather they share a common paternal ancestry. M346 is also found in India, although we have no haplotypes for comparison (Sengupta et al., 2006). Thus, the geographic location of ancestral Native Americans cannot be conclusively determined. However, these results confirm that at least some portion of the ancestral populations that went into the formation of Altaians also played a primary role in the formation of Native Americans.

TMRCAs for the entire haplogroup were calculated from rho statistics. With the evolutionary mutation rate, TMRCAs were estimated to be around 19 kya. The TMRCA was between 6.2 and 7.8 kya with the pedigree mutation rate. Clearly, in this case, the

pedigree rate provided unrealistically recent TMRCA. These Y-chromosomes were found in Siberia, Central Asia and throughout the Americas, and yet the pedigree rate TMRCA was only 7.8 kya. This estimate using the pedigree rate is not consistent with non-genetic evidence for the peopling of the New World.

Intrapopulation variances were calculated from the 8-STR profiles (Table 7.4). By far the greatest variances occurred in the Q haplotypes from Central Asia, particularly among the Uzbek and Dungan. Pakistan, Turkey and the Americas all had greater diversity than southern Siberian populations. This trend indicated that the source of Q Y-chromosomes likely was Central Asia. Given the TMRCA mentioned above, its origin occurred around the LGM.

Table 7.4 Intrapopulation variances for Q haplotypes

<b>Region</b>	<b>Population</b>	<b>Number of Samples</b>	<b>Intra-Population Variance</b>
<b>Southern Siberia</b>	<b>All</b>	<b>54</b>	<b>0.280</b>
	Altai-kizhi	20	0.080
	Chelkan	15	0.116
	Tubalar	11	0.094
	Tuvinian	6	0.067
<b>Central Asia</b>	<b>All</b>	<b>53</b>	<b>0.844</b>
	Uzbek	19	0.839
	Pamiri	13	0.811
	Yadhava	4	0.750
	Kazakh	3	0.238
	Dungan	3	1.095
<b>Pakistan</b>	<b>All</b>	<b>6</b>	<b>0.481</b>
<b>Turkey</b>	Turk	10	0.517
<b>Americas</b>	<b>All</b>	<b>185</b>	<b>0.447</b>

In the Altai, the northern and southern Altaians possessed lineages from two separate haplotype clusters. The TMRCA for these lineages was 14.5 kya ( $\pm 3.8$  kya)



using 14-STR profiles and an evolutionary mutation rate, while the TMRCA was 5.2 kya ( $\pm 1.4$  kya) with the pedigree rate. The variation in the southern Altaians was essentially twice that found in the northern Altaians. As a result, the TMRCA for northern Altaian Q1a3\* Y-chromosomes was 2.5 ( $\pm 1.1$ ) or 0.89 ( $\pm 0.41$ ) kya, depending on the mutation rate. The southern Altaian Q1a3\* Y-chromosomes had TMRCA of 5.0 ( $\pm 1.9$ ) or 1.8 ( $\pm 0.69$ ) kya. If the evolutionary rate is used (because the pedigree rate gave unrealistically recent dates for the whole haplogroup), then the southern Altaian common ancestor likely existed in the Neolithic, whereas the common ancestor for northern Altaian Q1a3\* lineages dated to the Bronze or Iron Age. It is also possible that the southern Altaians acquired a number of more distantly related lineages through gene flow with other populations. On the other hand, the northern Altaian lineages were so similar that it is highly likely that one common ancestor was the source for all of these haplotypes. BATWING estimates mostly matched those obtained using rho statistics, except for the southern Altaians, in which more recent estimates were attained. The much older TMRCA estimates for all the Altaian Q1a3\* lineages and the separate clusters in the network analysis provide enough evidence to show that Q1a3\* in Altaians did not originate from a single, recent, common source.

#### **7.4 Haplogroup C3\* and C3c**

Haplogroup C has a wide distribution, being found in Central Asia, Southeast Asia, Oceania, Northern Asia and the Americas (Derenko, Malyarchuk, Wozniak et al., 2007; Hammer et al., 2001; Karafet et al., 2002; Karafet et al., 2001; Kayser et al., 2003; Malhi et al., 2008; Malyarchuk, Derenko, Denisova et al., 2010; Sengupta et al., 2006;

Underhill et al., 2001; Xue et al., 2006; Zegura et al., 2004; Zhong et al., 2010). Given its high haplotypic diversity and its ancient split in the Y chromosome phylogeny, it is considered one of the first haplogroups to become successfully established outside of Africa (Underhill et al., 2001). The geographic distribution of haplogroup C also coincides with the mitochondrial macrohaplogroup M. Both mtDNA macrohaplogroup M and NRY haplogroup C were often used to infer at least one of the migration routes out of Africa, which traced along a southern route (Kivisild et al., 2003; Metspalu et al., 2004; Underhill et al., 2001).

Because haplogroup C is rather old, sufficient time has passed allowing for the differentiation of C branches throughout the world. C3 is a branch of particular importance for Asian and American population histories. It is defined by NRY marker M217 and is found throughout Central Asia, Mongolia, Siberia, northern China and parts of the Americas (Karafet et al., 2002; Malhi et al., 2008; Malyarchuk, Derenko, Denisova et al., 2010; Xue et al., 2006; Zegura et al., 2004; Zhong et al., 2010). C3b, which is defined by marker P39, is the only non-M45-derived haplogroup indigenous to the Americas (Zegura et al., 2004). It was found in higher frequencies among Na-Dene speakers but was also observed at low frequencies among some Amerind populations. Its presence in the New World has been cited as evidence for multiple migrations from Northern Asia, and particularly as a second migration that involved Na-Dene speakers (Karafet et al., 2002; Lell et al., 2002; Schurr & Sherry, 2004).

Even though C3b has not been found in Asia, the ancestral form (C3\*) has. Other branches of C3 include C3a, C3d and C3e, which have been found in India, Pakistan, southern Siberia and northern China (Sengupta et al., 2006; Zhong et al., 2010), as well

as C3f whose geographical location has not been reported (Karafet et al., 2008). The final derived branch identified within the C3 phylogeny is C3c. C3\* and C3c were found in the Altai, and therefore, they are the focus for the rest of this section. Accordingly, two networks were generated for haplogroup C3\* and C3c to visualize the geographic spread of these Y-chromosomes and to show the relative amounts of diversity among these haplotypes.

Among all of the C3 lineages, two have been the focus of considerable interest (Xue et al., 2005; Zerjal et al., 2003). The first belongs to C3\* and is famously associated with Genghis Khan, founder of the Mongol Empire during the thirteenth century, and his male descendants (Zerjal et al., 2003). This lineage has the derived state of marker M217, but ancestral states for all currently known derived branches. Its haplotype was defined using 15 STRs with the following motif: 13-16-25-10-11-13-14-12-11-11-11-12-14-10-10 (DYS389I-DYS389b-DYS390-DYS391-DYS 392-DYS 393-DYS388-DYS425-DYS426- DYS434- DYS435- DYS436- DYS437- DYS438-DYS439). The remarkable characteristic of this lineage is its widespread distribution, which extends from the Pacific Ocean to Central Asia and the Middle East. Zerjal et al. (2003) concluded it was nearly impossible for the distribution of this lineage to occur merely by chance. The lack of additional lineages from other haplogroups found in Mongolia today (especially when considering the diversity of current Mongolian populations) and the TMRCA for the haplotype cluster (estimated at  $1.0 \pm 0.3$  kya) arguably pointed to one scenario. Zerjal et al. (2003) invoked “social selection” to explain the unusual geographical provenience by associating this and related haplotypes with Genghis Khan and his descendants. Thus, the current distribution of the lineage and

its related haplotypes are the direct consequence of the spread of the Mongol Empire (and in this case, the Y-chromosome of Genghis Khan and his male descendents).

The second lineage that was the focus of additional scrutiny is the so-called “Manchu modal haplotype” (Xue et al., 2005). This set of lineages is on an M217/M48/M77/M86 background, which defines haplogroup C3c1. C3c1 has a more restricted distribution than C3\*. It is found among Mongolians, northern Chinese minorities, and some Siberian populations (Malyarchuk, Derenko, Denisova et al., 2010; Pakendorf et al., 2006; Pakendorf et al., 2007; Xue et al., 2005; Xue et al., 2006; Zhong et al., 2010). Much like the Genghis Khan star cluster, this lineage was discovered after a large number of samples were characterized and found to share this haplotype. The STR profile is 13-13-16-24-9-11-13-12-11-11-11-12-14-10-11 (DYS388-DYS389I-DYS389b-DYS390-DYS391-DYS392-DYS393-DYS425-DYS426-DYS434-DYS435-DYS436-DYS437-DYS438-DYS439). The authors of the study propose a similar scenario as Zerjal et al. (2003) for the distribution of the “Manchu haplotype” (Xue et al., 2005). Because the lineage has a coalescence age of about 500 years, it was suggested that the nobility of the Qing Dynasty are responsible for the lineage’s expansion. Some 61 additional STR loci were characterized to further differentiate a haplotype cluster associated with the modal type.

Among the Altaian samples analyzed for this dissertation, only the Altaian Kazakhs and Altai-kizhi possessed the modal haplotype from the Genghis Khan star cluster (Figure 6.2). The great majority of these Y-chromosomes in my data set came from the Altaian Kazakhs. Outside of Mongolia where this haplotype cluster occurs in over one third of the male population, it is most frequently found in Altaian Kazakhs

(~8%), followed by Altai-kizhi (3%), Kalmyks, Tuvinians and Buryats (2%) (Derenko, Malyarchuk, Wozniak et al., 2007; Zerjal et al., 2003).

This haplotype cluster was identified in the 14-STR RM-MJ network and had a TMRCA of 1550 ( $\pm$  390) years ago using the evolutionary rate and 560 ( $\pm$  230) years ago with the pedigree rate. The age of this haplotype cluster was consistent with the expansion of the Mongol Empire when the pedigree rate is used, but the TMRCA date could coincide with the migrations of nomadic pastoralists in northern China in the early first millennium CE if the evolutionary rate is used.

Most Altai-kizhi, however, had C3\* lineages that were distinctive from the Genghis Khan modal cluster, but still had affinities with Mongolian populations. One haplotype in particular has a deletion in the AZFc region, which includes the STR locus DYS448. These were restricted to Altai-kizhi in my sample set, but have also been reported in Kalmyk (formerly inhabitants of western Mongolia), Kyrgyz, and Tajiks (Balaesque et al., 2008). The haplotypes with the DYS448 deletion were calculated estimated to have a coalescence date of 2,900 ( $\pm$  766) years using the pedigree mutation rate. The TMRCA of this haplotype cluster in Altai-kizhi dated to 860 ( $\pm$  390) years ago using the evolutionary rate and 310 ( $\pm$  140) years ago using the pedigree rate.

(BATWING estimates mirror these results.) An exact match was found between the Altai-kizhi and Kyrgyz, with single-step variants present in Teleut, Kalmyk, Mongolian and Tajik. These findings point to a relatively recent common ancestor among these lineages, and therefore, these populations.

Forty-seven Altaian Kazakhs and five Altai-kizhi belonged to C3c1. While none belonged to the Manchu haplotype cluster identified by Xue et al. (2005), all formed

another cluster that was characterized by duplications of the DYS19 STR locus. Only three men with Y-chromosomes from this haplotype cluster did not show duplications at DYS19. However, the manner in which these genotypes were attained can only identify duplications when the second locus has a different repeat score than the first. Therefore, if a locus is duplicated and the repeat lengths are the same in both locations, then it will not be evident in the STR genotypes. Given the similarities at all other STR loci, it is assumed that these three men do have duplications at DYS19, but that the duplicated versions are the same length as the original.

Other C3c1 Y-chromosomes with duplications at DYS19 have been found (Balaesque et al., 2009). They were identified in Kalmyk, Mongolians, Kazakhs from Kazakhstan, Kyrgyz and northern Chinese populations. Using the pedigree rate, these lineages were dated to 1780 ( $\pm$  630) years ago (Balaesque et al., 2009). The C3c cluster from my data set gave a TMRCA of 1360 ( $\pm$  610) years ago with the pedigree mutation rate and 3780 ( $\pm$  1680) years ago using the evolutionary rate. (Again, estimates from BATWING matched these results).

Phylogeographic data have suggested a common origin of all haplogroup C Y-chromosomes in southeastern Asia, where many of its subhaplogroups are located. C\* was found in India and Southeast Asia, C1 in Japan, C2 in Melanesia and Polynesia, C4 in Australia, C5 in southern Asia (India and Pakistan) and C6 in New Guinea (Hudjashov et al., 2007; ISOGG, 2010; Karafet et al., 2008; Karafet et al., 2001; Kayser et al., 2006; Kivisild et al., 2003; Scheinfeldt et al., 2006; Sengupta et al., 2006; Underhill et al., 2001; Zhong et al., 2010). Haplogroup C has not yet been found in Africa. Therefore, it is assumed to have appeared during or after the migration(s) into the Asian continent.

Zhong et al. noted that the distribution of haplogroup C branches generally do not overlap, suggesting long-term isolation among these branches (Zhong et al., 2010, 429).

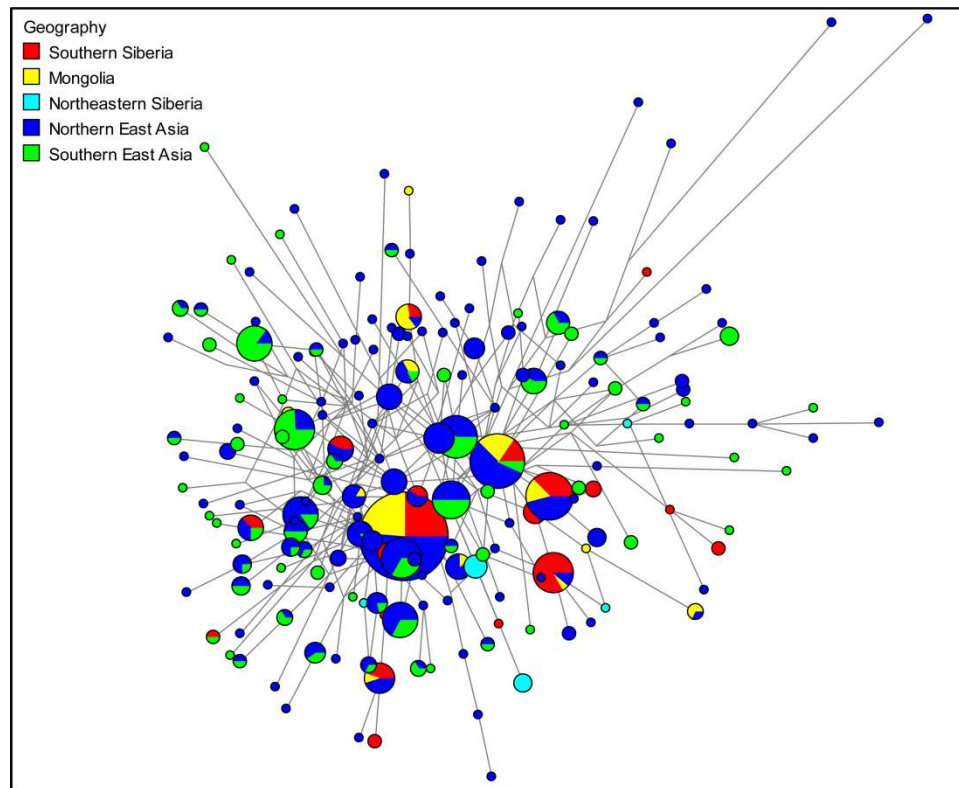


Figure 7.11 RM-MJ network for haplogroup C3\* (7-STR)

The TMRCA for the entire haplogroup has been estimated to be around 50 kya (Underhill et al., 2001). Based on STR diversity, I dated C3 to around 24.1 ( $\pm$  7.0) kya, which is similar to previous estimates (Zhong et al., 2010). The C3\* network was created using a 7-STR profile (Figure 7.11). Of the five geographic regions that have C3\* lineages (there was not a high-resolution data set for Central Asian populations), the southern East Asian populations displayed the greatest haplotypic diversity. Some of these lineages were shared with northern East Asian populations, but the northern East Asian populations shared nearly all of the lineages from Mongolia and southern Siberia.

The TMRCA for this haplogroup was estimated at 24.1 ( $\pm$  7.0) kya with the evolutionary mutation rate and 8.9 ( $\pm$  2.5) kya with the pedigree rate. Remarkably, despite the great diversity in southern East Asia, the Genghis Khan Haplotype cluster was not found there. Given the intrapopulation variance estimates, C3\* likely emerged from East Asia before the LGM, but it is not clear whether this occurred in the north or south of East Asia (Table 7.5).

Table 7.5 Intrapopulation variance for C3\* and C3c haplotypes

	<b>Region</b>	<b>Number of Samples</b>	<b>Intra-Population Variance</b>
<b>Haplogroup C3*</b>	Southern East Asia	155	0.402
	Northern East Asia	348	0.435
	Mongolia	46	0.234
	Southern Siberia	87	0.234
<b>Haplogroup C3c</b>	Northern East Asia	54	0.163
	Mongolia	59	0.159
	Southern Siberia	93	0.091
	Central Siberia	53	0.046

Haplogroup C3c1 provided a much clear pattern of haplotype diversity (Figure 7.12). All of the C3c Y-chromosomes came from Altaic speakers, with Mongolian and Northern Chinese populations having the greatest intrapopulation variance (Table 7.5). These similarities also suggested a more recent common ancestry for C3c compared to C3\*. The TMRCA for this haplogroup was 4.9 ( $\pm$  2.2) kya using the evolutionary rate and 1.8 ( $\pm$  0.8) kya using the pedigree rate. The geographic distribution of this haplogroup indicates that it likely was spread along with Altaic language expansions from northern China and Mongolia, probably with the adoption of nomadic pastoralism.



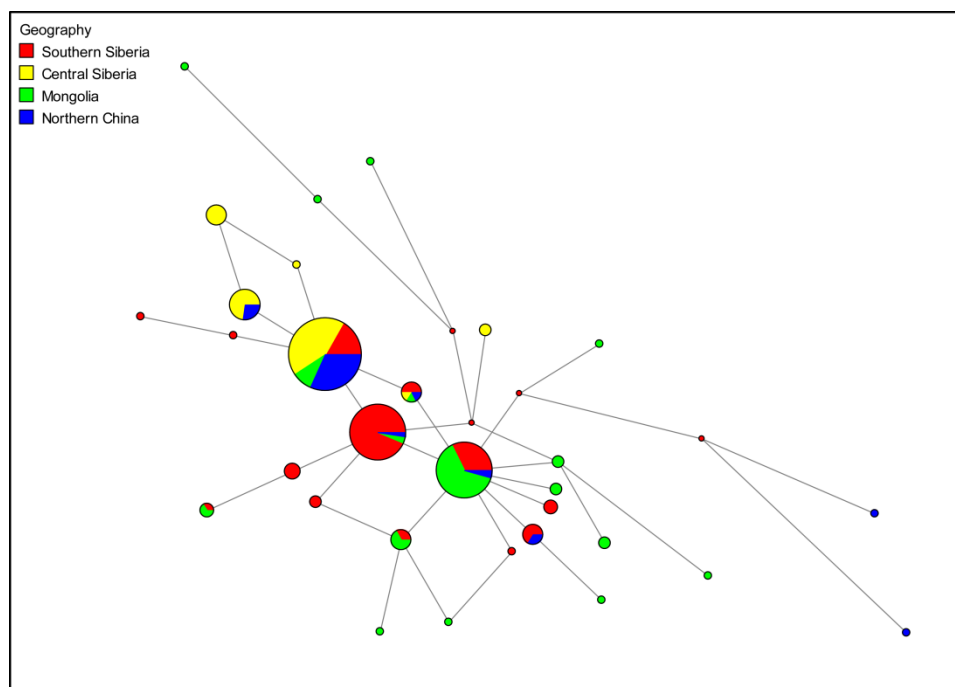


Figure 7.12 RM-MJ network of haplogroup C3c1 (9-STR)

## 7.5 Chapter Conclusions

The purpose of this chapter was to explore the phylogeny of the Y-chromosome in relation to the haplogroups that make up a significant portion of the Altaian paternal gene pool. By using phylogeographic methods, it was possible to ascertain the origins and source areas for each of these haplogroups. The high-resolution re-analysis also provided the information to determine when haplotype clusters originated and how populations were related given their membership among these smaller, more similar lineages. The limitation to this approach was the variable level of haplotypic resolution for the comparative data sets. While the data I generated included 17 Y-STR loci, most publications used only 7 – 10 STRs. Larger numbers of STRs are important for distinguishing the small differences that may exist between individuals or populations. Many haplotypes were shared when only considering seven STRs, but once fifteen were

used, very little sharing occurred. This level of resolution is critical for attaining useful divergence times between haplotype clusters within haplogroups. The greater the number of samples and loci used in the analysis, the lower the standard errors that are associated with each TMRCA. In addition, greater SNP resolution is also necessary for the proper placement of Y-chromosome lineages in the phylogeny. In several cases, the STR diversity was not sufficient to distinguish between several branches of a haplogroup, especially when only seven STRs were characterized.

The three haplogroups making up most of the northern Altaian Y-chromosome gene pools were Q1a3\*, R1a1a and N1b. N1b was found in modern populations that speak Samoyedic and Ugric languages. Similar haplotypes were found among the northern Altaians, Khakass and Shor of southern Siberia. The TMRCA placed the origin of this haplogroup in the Neolithic period, where it likely originated in southern or western Siberia. It almost certainly was present in the indigenous populations living along the Yenisei River, which later provided contributions to the modern populations in the northern portions of southern Siberia and in the Samoyedic-speaking peoples that moved north over the last few centuries.

Q1a3\* was found in both northern and southern Altaian populations, yet no haplotype sharing occurred between them. The TMRCA of all the Altaian lineages showed an ancient separation between these ethnic groups. Therefore, the co-occurrence of this haplogroup is not due to a common recent ancestry or from admixture. Rather the southern Altaian versions were more similar to other southern Siberian and Central Asian Q lineages. These likely represent a relatively recent origin, probably from Central Asia, where its diversity is greatest. The northern Altaian Q1a3\* provided a TMRCA that

dated to the Bronze or Iron Ages of southern Siberia, but their Q lineages were unique in that no other haplotypes in Siberia were found to be similar to them. In fact, it could be argued that the American Q lineages are the most similar. It appears that these lineages may represent a variety of Q Y-chromosomes that were more abundant in Siberia at one time, but have been replaced by other lineages in the recent historical past. If this is the case, then the northern Altaian Q Y-chromosomes may provide the best opportunity for finding a specific link between the New and Old World paternal gene pools.

The origin of R1a1a is of particular interest for many because it is so common throughout much of Eurasia. Initially, it was believed that this haplogroup originated in Central Asia, but given the STR diversity and locations of Y-chromosomes ancestral to R1a1a, India or southwestern Asia now seems a more likely source.

R1a1a was found in both southern and northern Altaian populations but, in this instance, it appears that least three separate waves of R1a1a Y-chromosomes may have influenced Altaian populations. The oldest would be the many dissimilar Y-chromosomes that have could have been in these regions for thousands of years. However, since they do not form distinctive clusters, it is nearly impossible to prove when they arrived, or where from where they came. Both northern and southern Altaians also had individuals belonging to one of two haplotype clusters that dated to the Neolithic. Once again, no haplotype sharing occurred between northern and southern Altaians for the Neolithic haplotype clusters. The third set of R1a1a lineages also form two distinctive clusters that date to the Eneolithic or Bronze Ages. If any R1a1a lineages could be associated with the Kurgan culture, then these would be the ones.

In addition to R1a1a, the Kumandin also possessed R1b1b1 Y-chromosomes. These were not typical for southern Siberian populations, but have been found at low frequencies among the Teleut. Based on the STR data, it appears that this haplogroup originated in Central Asia, possibly near the Ural Mountains. This is also the likely source for the Y-chromosomes that were found in the Altai. The arrival of R1b1b1 in the Altai probably coincided with or just preceded the migrations of the Afanasievo culture.

The last haplogroups, C3\* and C3c, certainly played a larger role in southern Altaian paternal gene pools in the Altai-kizhi and especially the Altaian Kazakhs. These lineages were probably the most recent to be incorporated into the Altaian populations. They were close or exact matches with C3\* and C3c lineages in Mongolia, Northern China and Central Asia, with several belonging to the Genghis Khan Haplotype cluster. Depending on the mutation rate used, however, the spread of C3 could coincide with westward expansions of Altaic peoples out of northeastern China. Regardless, it seems likely that these lineages were directly the result of Mongolic expansions into the Altai in recent years.

The initial modern human inhabitants of Siberia likely had haplogroups Q and R1, which first differentiated in Central Asia, and haplogroup C, which likely differentiated in Northern China. Q and C are found in the Americas, so it is logical that these two haplogroups were present in Siberia before the human settlement of the New World. Haplogroup N1b (and probably N1c) emerged in Siberia during the Neolithic after which an influx of new R1a1a Y-chromosomes entered from the steppe lands during the Eneolithic and Bronze Ages. Of the ten ancient Y-chromosomes characterized from southern Siberia, nine belong to R1a1a, while one belonged to C3 (x3c) (Keyser et al.,

2009). All of these samples came from Andronovo or Tagar kurgan burials. Certainly, R1a1a is associated with the people of these cultures. What is not yet known, however, are the NRY profiles of indigenous Neolithic or Bronze Age Siberians. It is likely that they possessed the N1b and Q lineages that were predecessors of those found in the northern Altai today. Following these expansions, C3\* was reintroduced to Siberia, probably during the Iron Age, with the emergence of the Xiongnu and other nomadic pastoralist tribes, but also again in the thirteenth century with the expansions of Genghis Khan and his Mongol empire.

The paternal genetic history of the Altai is remarkable in having retained a detailed molecular (historical) record of human occupation there. It informs us of its vast and complicated story, where peoples have interacted and cultural identities maintained. The details of NRY variation differ from the maternal histories already discussed for these populations, yet the overall pattern is consistent. Like the mtDNA, it is only with phylogeographic methods that a relative chronology can be gleaned from the Y-chromosome lineages. This chronology itself provides a crucial piece of the puzzle as to when Altaian Y-chromosomes expanded and where they originated.

## Conclusions

At the beginning of this dissertation, a series of objectives were laid out that, if achieved, would provide critical information that would allow us to better understand the genetic relationships and population histories of Altaian peoples. The first of these was to attain a high-resolution data set for the complementary maternal and paternal genetic systems. This goal was accomplished by using PCR-RFLP methods, control region sequencing and, even in some cases, complete mitochondrial genome re-sequencing, to obtain the highest resolution mtDNA data set possible for Altaian populations. For the NRY, almost 90 biallelic markers were characterized to place the Altaian Y-chromosomes in the most specific haplogroup possible. In addition, 17 Y-STRs were genotyped to provide high-resolution paternal haplotypes. To date, this is the highest resolution data set of any study on Siberian populations, in which both Y-chromosome and mtDNA were characterized.

By characterizing mtDNA and NRY genetic variation of Altaian populations at this level, differences between the northern and southern Altaians were detected. These differences were apparent with the Y-chromosome data, although one northern Altaian ethnic group (the Tubalars) was not significantly different from the Altai-kizhi (a southern Altaian ethnic group). In fact, the diversity among northern Altaian ethnic groups was often greater than the diversity between northern and southern Altaian populations. With the NRY data, the Kumandins were clearly an outlier from the Chelkan and Tubalar, which shared similar Q1a3\* and R1a1a Y-chromosomes. When considering the mtDNA data, the Chelkan were a distinctive northern Altaian ethnic group, having high frequencies of mtDNA haplogroups D5c2, F1a, F1b, F2a and N9a.

Overall, however, the mtDNA variation was much greater than that obtained with the NRY genetic system. The effects of social structure were probably responsible for this pattern of variation. In all cases, Altaian populations are patriarchal societies, where clan/süök membership is determined by following the father's line. Furthermore, these populations are patrilocal, thereby allowing closely related male relatives to congregate in the same general locations. These patterns remain today, despite the collectivization practices by the Russian government in the last century.

The differences found among the Northern Altaians can be explained in a several ways. First, while they appear to share a general common ancestry, the northern Altaian ethnic groups meandered down different paths through the process of genetic drift, only to emerge more recently as distinctive populations. All of them likely have genetic contributions from the former hunter-fisher groups that lived along the Yenisei River and in the Altai/Sayan region, which we know consisted of a complex and diverse set of culture groups (including Samoyedic, Yeniseian, Ugric and Turkic speakers). These same groups helped to form the foundation of all southern Siberian populations. In addition, there was clear evidence for the recent genetic influence of Mongolian-speaking populations, with this influence being more significant in the southern Altaians and Buryats.

The phylogeographic analyses of mtDNA and NRY haplogroups provided a temporal context for inferring how these populations formed over time. The mtDNA data provided evidence that haplogroups C and D (and also probably A and U5) likely made up the most ancient maternal lineages among Siberian populations, with C and D having TMRCAs in the Paleolithic. Similarly, there were several NRY haplogroups that also

occurred in ancient Siberians. These include some C3, Q1a3 and R1a1a Y-chromosomes. All three of these NRY haplogroups likely originated around or just before the LGM in the area of modern-day southern Siberia, eastern Central Asia and/or Northern China and Mongolia.

The TMRCA for many of the remaining mtDNAs and Y-chromosomes fell into one of three general time periods – the Neolithic, the Bronze/Iron Age and the recent past. The Neolithic period witnessed an explosion of new mtDNA and NRY types. A variety of branches of haplogroups C, D, U4 and U5 clearly dated to this time. NRY haplogroups N1b and N1c also arose during the Neolithic, and two Altaian R1a1a Y-STR haplotype clusters date to this period. Given the general increase in population sizes throughout Siberia at this time, it is not surprising that so many branches have TMRCA originating in the Neolithic.

The Bronze/Iron Ages were distinctive from the other time periods in the increase in frequency of West Eurasian lineages. For the mtDNA, greater frequencies of U4, U5, H and J were found in ancient DNA samples from Central Asia and southern Siberia. This is expected, given the archaeological and historical evidence collected from Eurasian sites. The presence of the Afanasievo and Andronovo peoples clearly played a significant role on the redistribution of culture traditions throughout the steppe. Furthermore, it can be argued that genetic components accompanied the cultural intrusions. For the NRY, a high frequency of R1a1a in southern Altaians could be associated with these Kurgan cultures moving into the area from the west. Y-STR haplotype clusters were also found in haplogroups C3 and C3c, which could indicate population movement from Northern China/Mongolia westward.



Finally, several mtDNA haplotypes and NRY haplotype clusters (that had very little, if any, variation) were found among the Altaians. Many of these were singleton members of a haplogroup. For example, one Chelkan possessed an NRY haplogroup L Y-chromosome and one Tubalar that had an E1b1b1c Y-chromosome. For the mtDNA, there was a single representative of haplogroup W. A recent arrival of these lineages into the Altai-Sayan region can explain their low levels of diversity.

In addition, several men possessed Y-STR haplotypes that belong to the Genghis Khan haplotype cluster. While the presence of some of the Altaian C3\* and C3c Y-chromosomes likely originated in the Iron Age, the Genghis Khan C3\* haplotypes showed the influence of Mongol political dominance in the southern Altai, which lasted for roughly 500 years. The Mongol influence could also explain the current distribution of low diversity mtDNA B haplotypes in southern Siberians.

Another objective of the dissertation was to determine what, if any, role the Altai had played in the peopling of the New World. The results presented here support a common origin of Altaian and Native American paternal lineages. Haplogroup Q is found in both northern and southern Altaian populations. Although the haplogroup specific to Native Americans (Q1a3a) is not found among Altaians, the M346 marker (designating the Q1a3\* haplogroup) is. Q1a3a occurs on a Q1a3 background, thus making Altaian Q Y-chromosomes the closest related type. Furthermore, the Altaian Q1a3 and Native American Q1a3 and Q1a3a haplotypes are related to each other. The modal haplotypes for the Altaian Q1a3 are similar to those for the American Q1a3 and Q1a3a Y chromosomes (Zegura et al., 2004).

The other haplogroup that so far has been found in Native Americans is C3b. Although C3b is not found in Altaians (nor anywhere else in the Old World), Zegura et al. (2004) found the closest C3\* haplotypes to those in the Americas among the Selkups, Kets and Altaians. Hence, these data further support a relationship between the paternal ancestral populations of Altaians and Native Americans. The lack of R1a1a and N-derived lineages among Native Americans likely indicates the colonization of the New World occurred prior to the emergence of these haplogroups or their movement into Siberia 8-15 kya. A single migration of dominant Q and minor C3 Y-chromosomes followed by differentiation via isolation and genetic drift is sufficient to explain the NRY diversity in the indigenous populations of the Americas.

Regarding the mtDNA data, Altaians possess all five of the haplogroups that are generally found among Native Americans (A, B, C, D and X). The problem, however, is when examined at a refined level (subhaplogroup and haplotype), there is not much evidence for a recent origin of the groups. Haplogroups A, B and X have very little diversity in Altaians, and these mtDNA lineages belong to subhaplogroups that are not found in the Americas. While, more variation is present in C and D mtDNAs, only one of the C mtDNA haplotype belongs to the C1 subhaplogroup. This subhaplogroup was found in Siberia and the New World, but the Altaian haplotype belongs to a branch not found in the Americas.

To conclude, the persistence of genetic differences between the northern and southern populations represents a “persistent frontier,” one that existed as far back as the Eneolithic and Bronze Ages. The differences in culture, language, subsistence strategies and economies have been maintained. Thus, today, cultural phenotypic and genetic

differences between northern and southern Altaians are evident. The northern Altaians have a strong affinity with populations that share a common ancestry to hunter-fisher peoples living in the Yenisei River region, while the southern Altaians show greater resemblance to the peoples living on the steppe. These differences have been preserved despite the redistribution of lineages through long-range migrations, nomadic pastoralism, and recent collectivization practices. Between maintaining this boundary and by being influenced by the effects of genetic drift, the current gene pools of Altaians were created and reflect their shared genetic relationships and history.

Appendix 1: Comparative Datasets for mtDNA and NRY Analyses

<b>mtDNA (HVS1)</b>			
<b>#</b>	<b>Population</b>	<b>Number</b>	<b>Reference</b>
1	Chelkan	91	This study
2	Kumandin	52	This study
3	Tubalar	71	This study
4	Altai-kizhi	90	Derenko et al., 2007
5	Shor	82	Derenko et al., 2007
6	Telenghit	71	Derenko et al., 2007
7	Teleut	53	Derenko et al., 2007
8	Khakass	57	Derenko et al., 2007
9	Soyot	30	Derenko et al., 2003
10	Tuvinian	105	Derenko et al., 2007
11	Todzhan	48	Derenko et al., 2003
12	Tofalar	58	Derenko et al., 2003
13	Tofalar	45	Starikovskaya et al., 2005
14	Tubalar-Chelkan	72	Starikovskaya et al., 2005
15	Tuvinian	97	Starikovskaya et al., 2005
16	Tuvinian	59	Pakendorf et al., 2006
17	Siberian Tatar	218	Naumova et al., 2008
18	Altai-kizhi	276	This study
19	Altaians (Telenghits, Altai-kizhi, Chelkan, Tubalar, Maimalar)	110	Derenko et al., 2003
20	Buryat	295	Derenko et al., 2007
21	Khanty	106	Pimenoff et la. 2008
22	Mansi	63	Pimenoff et al., 2008
23	Khanty	46	Naumova et al., 2009
24	Ket	38	Derbeneva et al., 2002a
25	Nganasan	39	Volodko et al., 2008
26	Yukaghir – Lower Kolyma	82	Volodko et al., 2008
27	Yukaghir – Upper Kolyma	18	Volodko et al., 2008
28	Yukaghir – Upper Anadyr	32	Volodko et al., 2008
29	Yukaghir	22	Pakendorf et al., 2006
30	Koryak – Alutor	56	Schurr et al., 1999
31	Koryak – Karagin	37	Schurr et al., 1999
32	Koryak – Palan	54	Schurr et al., 1999
33	Itel'men	46	Schurr et al., 1999
34	Chukchi	182	Volodko et al., 2008
35	Eskimo – Chaplin	50	Volodko et al., 2008
36	Eskimo – Naukan	39	Volodko et al., 2008
37	Eskimo – Sireniki	37	Volodko et al., 2008
38	Central Even	22	Pakendorf et al., 2007

<b>mtDNA (HVS1)</b>			
<b>#</b>	<b>Population</b>	<b>Number</b>	<b>Reference</b>
39	Western Evenk	23	Pakendorf et al., 2007
40	Eastern Evenk	45	Derenko et al., 2007
41	Western Evenk	73	Derenko et al., 2007
42	Western Evenk	39	Pakendorf et al., 2006
43	Iengra Evenk	19	Pakendorf et al., 2007
44	Evenk	72	Starikovskaya et al., 2005
45	Yakut	83	Puzyrev et al., 2003
46	Yakut	178	Pakendorf et al., 2006
47	Yakut	36	Derenko et al., 2007
48	Yakut-speaking Evenk	32	Pakendorf et al., 2006
49	Negidal	46	Starikovskaya et al., 2005
50	Nivki	51	Starikovskaya et al., 2005
51	Orok	61	Kong et al., 2003
52	Udegey	46	Starikovskaya et al., 2005
53	Ulchi	79	Starikovskaya et al., 2005
54	Crimean Tatar	20	Comas et al., 2004
55	Dungan	16	Comas et al., 2004
56	Highland Kyrgyz	47	Comas et al., 1998
57	Lowland Kyrgyz	48	Comas et al., 1998
58	Kyrgyz	20	Comas et al., 2004
59	Kazakh	55	Comas et al., 1998
60	Kazakh	50	Chaix et al., 2007
61	Kara-Kalpak – On Tort Uruw	53	Chaix et al., 2007
62	Kara-Kalpak – Qongirat	55	Chaix et al., 2007
63	Tajik	20	Comas et al., 2004
64	Tajik	44	Derenko et al., 2007
65	Turkmen	20	Comas et al., 2004
66	Turkmen	51	Chaix et al., 2007
67	Uyghur	71	Comas et al., 1998
68	Uzbek	20	Comas et al., 2004
69	Uzbek	40	Chaix et al., 2007
70	Khoreman Uzbek	20	Comas et al., 2004
71	Kazakh (Xinjiang)	60	Yao et al., 2004
72	Uzbek (Xinjiang)	70	Yao et al., 2004
73	Uyghur (Xinjiang)	51	Yao et al., 2004
74	Hui	51	Yao et al., 2004
75	Mongolian	102	Kolman et al., 1996
76	Mongolian	47	Derenko et al., 2007
77	Mongolian (Xinjiang)	59	Yao et al., 2004
78	Mongolian (Inner Mongolia)	48	Kong et al., 2003
79	Daur	45	Kong et al., 2003
80	Ewenki	47	Kong et al., 2003

<b>mtDNA (HVS1)</b>			
<b>#</b>	<b>Population</b>	<b>Number</b>	<b>Reference</b>
81	Oroqen	44	Kong et al., 2003
82	Korean	48	Kong et al., 2003
83	Han Chinese	250	Yao et al., 2002
84	Kalmyk	110	Derenko et al., 2007
85	Khamnigan	99	Derenko et al., 2007
Total Number – 85 populations; 5613 Individuals			

<b>NR1 (Biallelic Markers)</b>			
<b>#</b>	<b>Population</b>	<b>Number</b>	<b>Reference</b>
1	Chelkan	25	This study
2	Kumandin	17	This study
3	Tubalar	27	This study
4	Altai-kizhi	120	This study
5	Altaian Kazakh	119	This study
6	Altai-kizhi	92	Derenko et al., 2006
7	Teleut	47	Derenko et al., 2006
8	Teleut	35	Kharkov et al., 2009
9	Khakass	53	Derenko et al., 2006
10	Shor	51	Derenko et al., 2006
11	Todzhan	36	Derenko et al., 2006
12	Soyot	34	Derenko et al., 2006
13	Buryat	238	Derenko et al., 2006
14	Kalmyk	68	Derenko et al., 2006
15	Evenk	50	Derenko et al., 2006
16	Tofalar	32	Derenko et al., 2006
17	Tuvinian	113	Derenko et al., 2006
18	Mongolian	47	Derenko et al., 2006
19	Mongolian	65	Zerjal et al., 2002
20	Kyrgyz	21	Zerjal et al., 2002
21	Dungan	41	Zerjal et al., 2002
22	Uyghur	33	Zerjal et al., 2002
23	Kazakh	38	Zerjal et al., 2002
24	Uzbek	28	Zerjal et al., 2002
25	Tajik	22	Zerjal et al., 2002
26	Turkmen	21	Zerjal et al., 2002
27	Uyghur	31	Xue et al., 2006
28	Uyghur	39	Xue et al., 2006
29	Xibe	41	Xue et al., 2006
30	Mongolian	65	Xue et al., 2006
31	Daur	39	Xue et al., 2006

<b>NRV (Biallelic Markers)</b>			
<b>#</b>	<b>Population</b>	<b>Number</b>	<b>Reference</b>
32	Ewenki	26	Xue et al., 2006
33	Hezhe	45	Xue et al., 2006
34	Hui	35	Xue et al., 2006
35	Manchu	35	Xue et al., 2006
36	Oroqen	31	Xue et al., 2006
37	Tibetan	35	Xue et al., 2006
43	Nganasan	38	Karafet et al., 2002; Tambets et al., 2004
44	Nenet	148	Karafet et al., 2002; Tambets et al., 2004
45	Sel'kup	131	Karafet et al., 2002; Tambets et al., 2004
46	Ket	48	Karafet et al., 2002; Tambets et al., 2004
47	Dolgan	67	Karafet et al., 2002; Tambets et al., 2004
38	Khalkh	85	Katoh et al., 2005
39	Uriankha	60	Katoh et al., 2005
40	Zakhchin	60	Katoh et al., 2005
41	Khoton	40	Katoh et al., 2005
42	Manchu	101	Katoh et al., 2005
48	Even	41	Karafet et al., 1999
49	Mari	111	Tambets et al., 2004
50	Udmurt	87	Tambets et al., 2004
51	Manchu	52	Karafet et al., 2001
52	Chinese Evenk	41	Karafet et al., 2001
53	Oroqen	23	Karafet et al., 2001
54	Uyghur	68	Karafet et al., 2001
55	Mongolian	147	Karafet et al., 2001
56	Siberian Evenk	95	Karafet et al., 2001
57	Buryat	81	Karafet et al., 2001
58	Kazakh	30	Karafet et al., 2001
59	Uzbek	54	Karafet et al., 2001
60	Kyrgyz	13	Karafet et al., 2001
61	Tibetan	75	Karafet et al., 2001
62	Eastern Yugur	45	Zhou et al., 2008
63	Western Yugur	52	Zhou et al., 2008
64	Western Even	22	Pakendorf et al., 2007
65	Central Even	24	Pakendorf et al., 2007
66	Stony Tunguska Evenk	40	Pakendorf et al., 2006
67	Yakut-speaking Evenk	33	Pakendorf et al., 2006
68	Yakut	184	Pakendorf et al., 2006

<b>NRV (Biallelic Markers)</b>			
<b>#</b>	<b>Population</b>	<b>Number</b>	<b>Reference</b>
69	Yukaghir	13	Pakendorf et al., 2006
70	Khanty	28	Pimenoff et al., 2008
71	Mansi	25	Pimenoff et al., 2008
72	Khanty	27	Mirabal et al., 2009
73	Komi Izhemski	54	Mirabal et al., 2009
74	Komi Priluzski	49	Mirabal et al., 2009
75	Northern Russian	382	Balanovsky et al., 2009
76	Central Russian	359	Balanovsky et al., 2009
77	Southern Russian	484	Balanovsky et al., 2009
Total – 77 Populations; 5412 Individuals			

<b>NRV (STRs)</b>			
<b>#</b>	<b>Population</b>	<b>Number</b>	<b>Reference</b>
1	Chelkan	25	This study
2	Kumandin	17	This study
3	Tubalar	27	This study
4	Altai-kizhi	120	This study
5	Altaian Kazakh	119	This study
6	Teleut	35	Kharkov et al., 2009
7	Khanty	28	Pimenoff et al., 2008
8	Mansi	25	Pimenoff et al., 2008
9	Stony Tunguska Evenk	40	Pakendorf et al., 2006
10	Iengra Evenk	9	Pakendorf et al., 2006
11	Central Even	24	Pakendorf et al., 2007
12	Western Even	22	Pakendorf et al., 2007
13	Yukaghir	13	Pakendorf et al., 2006
14	Yakut	18	Sengupta et al., 2006
15	Uyghur	31	Xue et al., 2006
16	Uyghur	39	Xue et al., 2006
17	Mongolian	65	Xue et al., 2006
18	Turk	523	Cinnoglu et al.,
19	Hazara	25	Sengupta et al., 2006
20	Madjar	45	Biro et al., 2009
Total – 20 Populations; 1250 Individuals			



**Appendix 2: Altaian Y-STR Data**

Haplotype #	Haplogroup	DYS19	DYS385	DYS389I	DYS389b	DYS390	DYS391	DYS392	DYS393	DYS437	DYS438	DYS439	DYS448	DYS456	DYS458	DYS635	Y GATA H4	Chelkan	Kumandin	Tubalar	Altai-kizhi	Altaian Kazakh
1	C3*	15	12-15	13	16	24	9	11	13	14	11	11	null	16	16	21	11				2	
2	C3*	15	12-13	13	16	25	10	11	13	14	10	10	22	15	18	21	11				3	
3	C3*	16	12-15	13	16	24	9	11	13	14	11	11	null	16	16	21	11				6	
4	C3*	15	12-15	13	16	24	9	11	13	14	12	11	null	16	16	21	11				1	
5	C3*	16	12-15	11	16	24	9	11	13	14	11	11	null	16	16	21	11				1	
6	C3*	15	13-13	13	16	25	10	11	13	14	10	10	22	15	18	21	11				1	
7	C3*	16	12-15	13	16	24	9	11	13	14	11	11	null	16	16	21	11				1	
8	C3*	16	12-16	13	16	24	9	11	13	14	11	11	null	16	16	21	11				1	
9	C3*	16	12-15	13	16	24	9	11	13	14	11	11	null	15	16	21	11				1	
10	C3*	15	12-15	13	16	24	9	11	13	14	11	11	null	16	16	21	11				1	
11	C3*	16	12-15	13	16	24	9	11	13	14	11	11	null	16	16	20	11				1	
12	C3*	16	12-13	13	16	25	10	11	13	14	10	10	23	15	18	21	11					3
13	C3*	17	12-13	13	16	25	10	11	13	14	10	10	22	15	18	21	11				2	
14	C3*	16	12-13	13	16	25	10	11	13	14	10	10	22	15	18	21	11				7	
15	C3*	16	12-13	13	16	25	10	11	13	14	10	10	22	15	17	21	11				1	
16	C3*	16	12-12	13	16	25	10	11	13	14	10	10	22	15	18	21	11				4	
17	C3*	15	12-14	14	16	24	10	11	13	14	11	10	20	17	17	21	11				1	
18	C3*	16	12-13	13	16	26	10	11	13	14	10	10	22	15	19	21	11				1	
19	C3*	16	12-13	13	16	25	10	11	13	14	10	10	21	15	18	21	11				1	
20	C3*	15	12-14	14	16	24	10	11	13	14	11	10	null	18	17	21	11				1	
21	C3*	15	12-14	14	16	24	10	11	13	14	11	10	null	17	17	21	11				1	
22	C3*	16	12-13	13	17	25	10	11	13	14	10	10	23	16	18	21	11				1	
23	C3*	16	12-13	13	16	25	10	11	13	14	10	10	23	15	17	22	11				1	
24	C3c	16	12-12	14	17	23	9	11	13	14	10	11	20	15	19	23	10				1	1
25	C3c	15-17	12-12	14	17	24	9	11	13	14	10	11	20	15	18	22	10				1	
26	C3c	15-17	12-12	14	17	24	9	11	13	14	10	11	20	15	18	23	10				1	
27	C3c	15-16	12-12	14	17	23	9	11	13	14	10	11	20	15	18	23	10				1	
28	C3c	15-17	12-12	14	17	24	9	11	13	15	10	11	20	15	18	23	10				1	
29	C3c	15-16	12-12	15	17	24	9	11	13	14	10	11	20	15	18	24	10					1
30	C3c	15-16	12-12	13	17	24	9	11	13	14	10	12	20	15	19	24	10					2
31	C3c	15-16	12-12	14	17	24	9	11	13	14	10	12	20	15	19	24	10					1
32	C3c	15-17	12-12	14	17	23	9	11	13	14	10	11	20	15	19	23	10					1
34	C3c	15-17	12-12	13	17	24	9	11	13	14	10	11	20	15	19	24	10					2
35	C3c	15-17	12-12	13	17	24	9	11	13	14	10	11	20	15	18	24	10					1
36	C3c	16-17	12-12	14	17	24	9	11	13	14	10	11	20	15	18	24	10					1
37	C3c	15-17	12-12	13	30	24	9	11	13	14	10	11	20	15	19	24	10					16
38	C3c	15-17	12-12	13	17	24	9	11	13	14	10	11	20	15	18	25	10					1
39	C3c	17-18	12-12	14	16	24	9	11	13	14	10	11	21	15	18	23	10					1

Haplotype #	Haplogroup	DYS19	DYS385	DYS389I	DYS389b	DYS390	DYS391	DYS392	DYS393	DYS437	DYS438	DYS439	DYS448	DYS456	DYS458	DYS635	Y GATA H4	Chelkan	Kumandin	Tubalar	Altai-kizhi	Altai Kazakh	
40	C3c	15	12-12	14	17	24	9	11	13	14	10	11	20	15	18	24	10					1	
41	C3c	15-17	12-12	14	17	24	9	11	13	14	10	11	20	15	18	24	10					3	
42	C3c	15-17	12-12	13	17	24	9	11	13	14	10	12	20	15	19	24	10					1	
43	C3c	15-17	12-12	13	17	24	9	11	13	14	10	11	20	15	19	25	10					3	
44	C3c	15-17	12-12	13	17	24	9	11	13	14	10	11	20	15	18	24	10					5	
45	C3c	15-17	12-12	13	17	24	9	11	13	14	10	12	20	15	18	23	10					1	
46	C3c	15-17	12-12	13	17	24	9	11	13	14	10	11	X	15	19	24	10					1	
47	C3c	15-17	12-12	14	17	24	9	11	13	14	10	11	20	15	20	24	10					1	
48	C3c	15-17	12-12	13	16	25	9	11	13	14	10	11	20	15	19	24	10					1	
49	C3c	15-17	12-12	13	18	24	9	11	13	14	10	11	20	15	19	24	10					1	
50	D3a	15	11-11	14	16	25	11	7	13	14	11	13	18	15	17	21	11					1	
51	D3a	15	11-11	14	16	26	11	7	13	14	11	12	19	15	17	21	11					2	
52	D3a	15	11-11	14	16	25	11	7	13	14	11	11	19	15	17	21	11					2	
53	D3a	15	11-11	14	16	25	11	7	13	14	11	12	19	15	17	21	11					1	
54	E1b1b1c	13	17-18	13	18	25	9	11	14	14	10	12	20	15	17	21	11			1			
55	G1	13	13-17	14	15	23	10	12	13	16	10	13	22	16	15	20	11					3	
56	G1	13	13-17	14	15	23	11	12	13	16	10	12	22	16	15	20	11					1	
57	G2	15	16-16	13	16	22	10	10	13	16	10	11	21	16	17	21	11					2	
58	I2a	16	14-16	13	20	24	11	11	13	15	10	13	20	15	18	25	11			1			
59	J2a	14	15-16	13	16	23	10	11	12	15	9	11	21	15	18	23	11				3	1	
60	J2a	14	13-16	13	16	24	11	11	12	15	9	11	19	15	20	21	11					4	
61	L	15	9-16	13	17	24	10	14	12	16	10	11	19	15	17	22	11	1					
62	NO*(xN1,O)	14	12-12	13	17	23	11	14	13	14	10	10	19	15	17	24	11					1	
63	N1*	17	11-13	14	17	24	10	14	14	15	9*	11	20	17	17	23	11		1	3			
64	N1b*	14	12-13	13	16	23	10	14	12	14	10	10	19	15	17	23	12					1	
65	N1b*	15	12-13	13	16	23	10	14	13	14	10	10	19	15	17	24	12	2					
66	N1b*	14	12-13	13	16	23	10	14	13	14	10	10	19	15	17	24	13	1					
67	N1b*	14	12-13	13	16	23	10	14	13	14	10	10	19	15	17	24	12	1					
68	N1b*	14	12-12	13	16	23	11	14	13	14	10	10	19	15	17	24	12	1					
69	N1b*	14	12-12	13	16	23	10	15	12	14	10	10	19	15	17	24	12						
70	N1b*	14	12-12	13	16	23	10	15	12	14	10	10	19	15	17	24	13						
71	N1b*	14	12-12	13	16	23	10	14	12	14	10	10	19	15	17	24	12						
72	N1b*	14	11-12	13	16	23	10	15	12	14	10	10	19	15	17	24	12						
73	N1c(xN1c1)	14	11-12	13	15	24	11	15	13	14	10	11	20	16	18	21	13					1	
74	N1c1	14	11-11	13	16	23	11	14	14	14	10	10	19	14	17	22	12					1	
75	N1c1	15	12-13	14	16	23	11	14	14	14	10	11	20	15	18	23	11					1	
76	O3a3c*	15	13-18	12	17	23	10	13	12	15	10	12	19	15	17	19	12					1 19	

Haplotype #	Haplogroup	DYS19	DYS385	DYS389I	DYS389b	DYS390	DYS391	DYS392	DYS393	DYS437	DYS438	DYS439	DYS448	DYS456	DYS458	DYS635	Y GATA H4	Chelkan	Kumandin	Tubalar	Altai-kizhi	Altaiian Kazakh
77	O3a3c*	15	13-18	12	16	23	10	13	12	15	10	12	19	15	17	20	12					1
78	O3a3c*	15	13-18	12	17	22	10	13	12	15	10	12	19	15	17	19	12					1
79	O3a3c*	15	13-17	12	17	23	10	13	12	15	10	12	19	15	17	19	12					3
80	O3a3c*	13	12-15	12	16	23	10	12	12	15	11	11	19	15	17	19	11					3
81	O3a3c*	15	14-18	12	17	23	10	13	12	15	10	12	19	15	17	19	12					1
82	O3a3c*	15	13-18	12	17	23	10	13	12	15	10	13	19	15	17	19	12					1
83	O3a3c*	15	13-18	13	17	23	10	13	12	15	10	12	19	15	17	19	12					1
84	O3a3c*	15	13-18	12	17	23	10	13	12	15	10	12	19	15	17	20	12					1
85	O3a3c1*	14	12-17	12	15	23	10	15	12	15	11	13	20	15	17	20	12			1	1	
86	Q1a2	13	13-16	13	15	24	10	15	13	14	11	13	22	16	17	24	11			1		
87	Q1a3*	13	15-19	14	17	24	10	14	13	14	11	12	19	17	16	22	11	1				
88	Q1a3*	13	15-18	14	19	23	10	15	14	13	11	12	19	15	16	22	11					1
89	Q1a3*	13	14-14	13	18	23	10	14	13	13	11	12	19	15	16	22	11					8
90	Q1a3*	13	15-16	13	18	23	10	14	13	13	11	12	19	15	18	22	11					4
91	Q1a3*	13	15-16	13	19	23	10	14	13	13	11	11	18	15	18	22	11					2
92	Q1a3*	13	15-16	13	17	23	10	14	13	13	11	12	19	15	18	22	11					2
93	Q1a3*	13	15-16	13	17	23	9	14	13	13	11	12	19	15	17	22	11					1
94	Q1a3*	13	14-14	13	18	23	10	14	13	13	11	12	19	14	16	22	11					1
95	Q1a3*	13	15-16	14	18	23	10	14	13	13	11	12	19	15	18	22	11					1
96	Q1a3*	13	15-19	14	17	23	10	14	13	14	11	12	19	17	16	22	11	2				
97	Q1a3*	10	15-19	14	17	24	10	14	13	14	11	13	19	17	16	22	11	1				
98	Q1a3*	13	15-19	14	17	25	10	14	13	14	11	13	19	17	16	22	11	1				
99	Q1a3*	13	15-19	14	17	24	10	14	13	14	11	13	19	18	16	22	10	1				
100	Q1a3*	13	15-19	14	17	24	10	14	13	14	11	13	19	17	16	22	11	6		1		
101	Q1a3*	13	15-16	14	17	24	10	14	13	14	11	13	19	16	16	22	11	1		1		
102	Q1a3*	13	15-18	14	17	24	10	14	13	14	11	13	19	16	16	22	11	1		4		
103	Q1a3*	13	15-18	14	17	24	10	14	13	14	11	13	19	17	16	22	11	1				
104	Q1a3*	13	15-16	13	17	23	10	14	12	13	11	11	20	15	16	22	11					1
105	Q1a3*	13	15-18	15	17	24	10	14	13	14	11	13	19	16	16	22	12			1		
106	Q1a3*	13	15-17	14	17	24	10	14	13	14	11	12	19	16	16	22	11			2		
107	R1a1a*	17	11-17	14	17	26	11	11	13	14	11	11	19	15	15	23	11					2
108	R1a1a*	16	11-14	14	18	25	11	11	13	14	11	10	21	16	15	23	10					5
109	R1a1a*	15	11-15	12	18	25	11	11	13	14	11	10	21	15	15	23	12					1
110	R1a1a*	16	11-17	14	17	26	11	11	13	14	11	11	19	15	15	23	11					4
111	R1a1a*	16	11-18	14	17	25	11	11	13	14	11	11	19	15	15	23	10					2
112	R1a1a*	16	11-17	13	17	26	11	11	13	15	11	11	19	15	15	23	11					1
113	R1a1a*	16	11-14	14	18	25	11	11	13	14	11	10	21	16	15	23	12					9
114	R1a1a*	16	11-14	15	18	25	11	11	13	14	11	10	21	17	15	23	12					2

Haplotype #	Haplogroup	DYS19	DYS385	DYS389I	DYS389b	DYS390	DYS391	DYS392	DYS393	DYS437	DYS438	DYS439	DYS448	DYS456	DYS458	DYS635	Y GATA H4	Chelkan	Kumandin	Tubalar	Altai-kizhi	Altai Kazakh	
115	R1a1a*	16	11-14	14	18	26	10	11	13	14	11	10	21	16	15	23	12				3		
116	R1a1a*	16	11-14	14	18	25	11	11	13	14	12	10	21	16	15	23	12				1		
117	R1a1a*	16	11-14	14	18	25	11	11	13	14	11	10	21	16	15	24	12				2		
118	R1a1a*	16	11-14	14	17	25	11	11	13	14	11	10	21	15	15	23	12				1		
119	R1a1a*	16	11-14	14	17	25	11	11	13	14	11	10	21	17	15	23	12				1		
120	R1a1a*	16	11-14	14	18	25	12	11	13	14	11	10	21	17	15	23	12				2		
121	R1a1a*	16	11-14	14	18	25	11	11	13	14	11	10	21	17	15	23	12				5		
122	R1a1a*	16	11-14	14	18	25	11	11	13	14	11	10	21	15	15	23	12				1		
123	R1a1a*	16	11-17	14	18	26	10	11	13	14	11	11	19	15	15	23	11				3		
124	R1a1a*	16	11-14	14	18	25	11	11	13	14	11	10	21	17	15	23	11				1		
125	R1a1a*	16	11-17	14	18	26	11	11	13	14	11	12	19	15	15	23	11				1		
126	R1a1a*	16	11-17	14	18	26	11	11	13	14	11	11	19	15	15	23	11				1		
127	R1a1a*	16	11-17	14	18	26	10	11	13	14	11	12	19	15	15	23	11				1		
128	R1a1a*	16	11-14	14	18	25	11	11	13	14	11	10	21	16	15	23	11				1		
129	R1a1a*	17	11-13	12	18	25	10	11	13	14	11	10	20	16	15	23	13				1		
130	R1a1a*	16	11-17	14	17	26	11	11	13	14	11	11	19	15	16	23	11				1		
131	R1a1a*	16	11-18	14	18	26	10	11	13	14	11	11	19	15	15	23	11				1		
132	R1a1a*	16	11-17	14	17	26	11	11	13	14	11	11	20	15	15	23	11				1		
133	R1a1a*	16	11-14	14	17	25	11	11	13	14	11	10	21	16	15	23	12				1		
134	R1a1a*	16	11-14	14	18	25	10	11	13	14	11	10	21	16	15	23	12				1		
135	R1a1a*	16	12-14	14	17	24	10	11	13	14	11	10	20	17	14	23	12				1		
136	R1a1a*	16	14-15	14	19	25	11	11	13	14	11	10	21	17	15	23	12				1		
137	R1a1a*	15	11-15	13	19	25	10	11	13	14	11	10	20	15	16	23	13				1		
138	R1a1a*	16	12-14	14	18	24	9	11	13	14	11	10	20	18	14	23	12	3					
139	R1a1a*	16	11-14	14	17	25	11	11	13	14	11	11	20	16	17	23	13	1					
140	R1a1a*	15	11-14	13	17	25	11	11	13	14	11	10	20	16	15	23	11		1				
141	R1a1a*	16	11-14	13	16	23	11	11	13	14	11	10	20	15	15	23	12			1		1	
142	R1a1a*	15	11-15	13	17	24	10	11	13	14	11	10	20	16	15	23	12		1				
143	R1a1a*	16	11-14	14	17	24	11	11	13	14	11	11	20	16	16	23	14				2		
144	R1a1a*	16	12-14	14	18	24	10	11	13	14	11	10	20	17	14	23	12				1		
145	R1a1a*	16	11-14	14	17	24	11	11	13	14	11	11	20	16	16	23	13				1		
146	R1a1a*	16	12-14	14	19	24	9	11	13	14	11	10	20	17	14	23	12				2		
147	R1a1a*	16	11-14	13	17	25	10	11	13	14	11	11	20	17	16	23	11				1		
148	R1a1a*	16	11-14	13	17	25	11	11	14	14	11	12	21	16	17	23	14				1		
149	R1a1a*	16	12-14	14	18	24	10	11	13	14	11	10	20	16	14	23	12				1		
150	R1a1a*	16	11-14	14	17	25	11	11	13	14	11	11	20	16	16	23	14				1		
151	R1b1b1	14	13-13	14	16	19	11	13	13	15	10	13	19	15	17	24	11				1	1	
152	R1b1b1	14	13-13	14	16	19	11	13	13	14	10	13	19	15	17	24	10						2

Haplotype #	Haplogroup	DYS19	DYS385	DYS389I	DYS389b	DYS390	DYS391	DYS392	DYS393	DYS437	DYS438	DYS439	DYS448	DYS456	DYS458	DYS635	Y GATA H4	Chelkan	Kumandin	Tubalar	Altai-kizhi	Altaiian Kazakh
153	R1b1b1	14	13-13	13	16	19	10	13	13	15	10	13	19	15	18	24	11	2				
154	R1b1b1	14	13-13	13	16	19	10	13	13	15	10	13	19	15	17	24	11	3				
155	R1b1b1	14	13-16	13	17	22	11	13	13	15	10	13	20	15	16	23	11	1				
156	T	15	13-15	13	16	23	11	13	13	14	9	11	18	17	15	21	11					1

## Bibliography

- 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.
- Achilli, A., Rengo, C., Battaglia, V., Pala, M., Olivieri, A., Fornarino, S., et al. (2005). Saami and Berbers--an unexpected mitochondrial DNA link. *Am J Hum Genet*, 76(5), 883-886.
- Agulnik, A. I., Zharkikh, A., Boettger-Tong, H., Bourgeron, T., McElreavey, K., & Bishop, C. E. (1998). Evolution of the DAZ gene family suggests that Y-linked DAZ plays little, or a limited, role in spermatogenesis but underlines a recent African origin for human populations. *Hum Mol Genet*, 7(9), 1371-1377.
- Albu, M., Min, X. J., Hickey, D., & Golding, B. (2008). Uncorrected nucleotide bias in mtDNA can mimic the effects of positive Darwinian selection. *Mol Biol Evol*, 25(12), 2521-2524.
- Amo, T., & Brand, M. D. (2007). Were inefficient mitochondrial haplogroups selected during migrations of modern humans? A test using modular kinetic analysis of coupling in mitochondria from hybrid cell lines. *Biochem J*, 404(2), 345-351.
- 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.
- Anthony, D. W. (2007). *The horse, the wheel, and language: how Bronze-Age riders from the Eurasian steppes shaped the modern world*. Princeton, N.J.: Princeton University Press.
- Atkinson, Q. D., Gray, R. D., & Drummond, A. J. (2008). mtDNA variation predicts population size in humans and reveals a major Southern Asian chapter in human prehistory. *Mol Biol Evol*, 25(2), 468-474.
- Avise, J. C., Arnold, J., Ball, R. M., Bermingham, E., Lamb, T., Neigel, J. E., et al. (1987). Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Systematics*, 18, 489-522.
- Awadalla, P., Eyre-Walker, A., & Smith, J. M. (1999). Linkage disequilibrium and recombination in hominid mitochondrial DNA. *Science*, 286(5449), 2524-2525.
- Bailliet, G., Ramallo, V., Muzzio, M., García, A., Santos, M. R., Alfaro, E. L., et al. (2009). Brief communication: Restricted geographic distribution for Y-Q\*

- paragroup in South America. *American Journal of Physical Anthropology*, 140(3), 578-582.
- Balanovsky, O., Rootsi, S., Pshenichnov, A., Kivisild, T., Churnosov, M., Evseeva, I., et al. (2008). Two sources of the Russian patrilineal heritage in their Eurasian context. *Am J Hum Genet*, 82(1), 236-250.
- Balaresque, P., Bowden, G. R., Parkin, E. J., Omran, G. A., Heyer, E., Quintana-Murci, L., et al. (2008). Dynamic nature of the proximal AZFc region of the human Y chromosome: multiple independent deletion and duplication events revealed by microsatellite analysis. *Hum Mutat*, 29(10), 1171-1180.
- Balaresque, P., Parkin, E. J., Roewer, L., Carvalho-Silva, D. R., Mitchell, R. J., van Oorschot, R. A., et al. (2009). Genomic complexity of the Y-STR DYS19: inversions, deletions and founder lineages carrying duplications. *Int J Legal Med*, 123(1), 15-23.
- Balloux, F., Handley, L. J., Jombart, T., Liu, H., & Manica, A. (2009). Climate shaped the worldwide distribution of human mitochondrial DNA sequence variation. *Proc Biol Sci*, 276(1672), 3447-3455.
- Bandelt, H. J., Forster, P., & Rohl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol*, 16(1), 37-48.
- Bandelt, H. J., & Parson, W. (2008). Consistent treatment of length variants in the human mtDNA control region: a reappraisal. *Int J Legal Med*, 122(1), 11-21.
- Bandelt, H. J., Quintana-Murci, L., Salas, A., & Macaulay, V. (2002). The fingerprint of phantom mutations in mitochondrial DNA data. *Am J Hum Genet*, 71(5), 1150-1160.
- Barfield, T. J. (1989). *The perilous frontier: nomadic empires and China*. Cambridge, Mass., USA: Basil Blackwell.
- Barrell, B. G., Bankier, A. T., & Drouin, J. (1979). A different genetic code in human mitochondria. *Nature*, 282(5735), 189-194.
- Bazaliiskii, V. I. (2010). Mesolithic and Neolithic mortuary complexes in the Baikal region. In A. W. Weber, M. A. Katzenberg & T. G. Schurr (Eds.), *Prehistoric Hunters-Gathers of the Baikal Region, Siberia* (pp. 51-86). Philadelphia: University of Pennsylvania Press.
- Bazin, E., Glemin, S., & Galtier, N. (2006). Population size does not influence mitochondrial genetic diversity in animals. *Science*, 312(5773), 570-572.

- Bermisheva, M. A., Kutuev, I. A., Spitsyn, V. A., Villems, R., Batyrova, A. Z., Korshunova, T., et al. (2005). [Analysis of mitochondrial DNA variation in the population of Oroks]. *Genetika*, 41(1), 78-84.
- Betti, L., Balloux, F., Amos, W., Hanihara, T., & Manica, A. (2009). Distance from Africa, not climate, explains within-population phenotypic diversity in humans. *Proc Biol Sci*, 276(1658), 809-814.
- Bielawski, J. P., & Yang, Z. (2001). Positive and Negative Selection in the DAZ Gene Family. *Mol Biol Evol*, 18(4), 523-529.
- Bobrick, B. (1992). *East of the Sun: the epic conquest and tragic history of Siberia*. New York: Poseidon Press.
- Bobrov, V. V. (1988). On the Problem of Interethnic Relations in South Siberia in the Third and Early Second Millennium B.C. *Arctic Anthropology*, 25(2), 30-46.
- Bokovenko, N. A. (1995a). History of studies and the main problems in the archaeology of southern Siberia during the Scythian period. In J. Davis-Kimball, V. A. Bashilov & L. T. Yablonsky (Eds.), *Nomads of the Eurasian steppes in the Early Iron Age* (pp. 255-261). Berkeley, CA: Zinat Press.
- Bokovenko, N. A. (1995b). The Tagar Culture in the Minusinsk Basin. In J. Davis-Kimball, V. A. Bashilov & L. T. Yablonsky (Eds.), *Nomads of the Eurasian steppes in the Early Iron Age* (pp. 299-314). Berkeley, CA: Zinat Press.
- Bokovenko, N. A. (1995c). Tuva during the Scythian period. In J. Davis-Kimball, V. A. Bashilov & L. T. Yablonsky (Eds.), *Nomads of the Eurasian steppes in the Early Iron Age* (pp. 265-281). Berkeley, CA: Zinat Press.
- Bolnick, D. A., Bolnick, D. I., & Smith, D. G. (2006). Asymmetric male and female genetic histories among Native Americans from Eastern North America. *Mol Biol Evol*, 23(11), 2161-2174.
- Bortolini, M. C., Salzano, F. M., Bau, C. H., Layrisse, Z., Petzl-Erler, M. L., Tsuneto, L. T., et al. (2002). Y-chromosome biallelic polymorphisms and Native American population structure. *Ann Hum Genet*, 66(4), 255-259.
- Bortolini, M. C., Salzano, F. M., Thomas, M. G., Stuart, S., Nasanen, S. P., Bau, C. H., et al. (2003). Y-chromosome evidence for differing ancient demographic histories in the Americas. *Am J Hum Genet*, 73(3), 524-539.
- Bowcock, A. M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J. R., & Cavalli-Sforza, L. L. (1994). High resolution of human evolutionary trees with polymorphic microsatellites. *Nature*, 368(6470), 455-457.



- Brace, C. L. (2005). *"Race" is a four-letter word: the genesis of the concept*. New York: Oxford University Press.
- Brandstätter, A., Sängler, T., Lutz-Bonengel, S., Parson, W., Béraud-Colomb, E., Wen, B., et al. (2005). Phantom mutation hotspots in human mitochondrial DNA. *Electrophoresis*, 26(18), 3414-3429.
- Brown, W. M. (1980). Polymorphism in mitochondrial DNA of humans as revealed by restriction endonuclease analysis. *Proc Natl Acad Sci U S A*, 77(6), 3605-3609.
- Brown, W. M., George, M., Jr., & Wilson, A. C. (1979). Rapid evolution of animal mitochondrial DNA. *Proc Natl Acad Sci U S A*, 76(4), 1967-1971.
- Cann, R. L., Stoneking, M., & Wilson, A. C. (1987). Mitochondrial DNA and human evolution. *Nature*, 325(6099), 31-36.
- Chaix, R., Austerlitz, F., Khegay, T., Jacquesson, S., Hammer, M. F., Heyer, E., et al. (2004). The genetic or mythical ancestry of descent groups: lessons from the Y chromosome. *Am J Hum Genet*, 75(6), 1113-1116.
- Chaix, R., Quintana-Murci, L., Hegay, T., Hammer, M. F., Mobasher, Z., Austerlitz, F., et al. (2007). From social to genetic structures in central Asia. *Curr Biol*, 17(1), 43-48.
- Chandrasekar, A., Kumar, S., Sreenath, J., Sarkar, B. N., Urade, B. P., Mallick, S., et al. (2009). Updating phylogeny of mitochondrial DNA macrohaplogroup M in India: dispersal of modern human in South Asian corridor. *PLoS One*, 4(10), e7447.
- Chang, X., Wang, Z., Hao, P., Li, Y. Y., & Li, Y. X. (2010). Exploring mitochondrial evolution and metabolism organization principles by comparative analysis of metabolic networks. *Genomics*, 95(6), 339-344.
- Changchun, Y., Li, X., Xiaolei, Z., Hui, Z., & Hong, Z. (2006). Genetic analysis on Tuoba Xianbei remains excavated from Qilang Mountain Cemetery in Qahar Right Wing Middle Banner of Inner Mongolia. *FEBS Lett*, 580(26), 6242-6246.
- Chard, C. S. (1958). An Outline of the Prehistory of Siberia Part 1. The Pre-Metal Periods. *Southwestern Journal of Anthropology*, 14(1), 1-33.
- Charlesworth, B., & Charlesworth, D. (2000). The degeneration of Y chromosomes. *Philos Trans R Soc Lond B Biol Sci*, 355(1403), 1563-1572.
- Chen, Y. S., Olckers, A., Schurr, T. G., Kogelnik, A. M., Huoponen, K., & Wallace, D. C. (2000). mtDNA variation in the South African Kung and Khwe and their genetic relationships to other African populations. *Am J Hum Genet*, 66(4), 1362-1383.

- Chernykh, E. N. (2008). Formation of the Eurasian "steppe belt" of stockbreeding cultures: viewed through the prism of archaeometallurgy and radiocarbon dating. *Archaeology, Ethnology and Anthropology of Eurasia*, 35(3), 36-53.
- Chikisheva, T. A. (2008). The origins of the early nomadic populations of Tuva: craniometrical evidence. *Archaeology, Ethnology and Anthropology of Eurasia*, 36(4), 120-139.
- Chinnery, P. F., Thorburn, D. R., Samuels, D. C., White, S. L., Dahl, H. M., Turnbull, D. M., et al. (2000). The inheritance of mitochondrial DNA heteroplasmy: random drift, selection or both? *Trends Genet*, 16(11), 500-505.
- Chlachula, J. (2001). Pleistocene climate change, natural environments and palaeolithic occupation of the Altai area, west-central Siberia. *Quaternary International*, 80-81, 131-167.
- Cinnioglu, C., King, R., Kivisild, T., Kalfoglu, E., Atasoy, S., Cavalleri, G. L., et al. (2004). Excavating Y-chromosome haplotype strata in Anatolia. *Hum Genet*, 114(2), 127-148.
- Cleaves, F. W. (1982). *The secret history of the Mongols* (pp. 277). Cambridge, MA: Harvard University Press.
- Collins, D. N. (1991). Subjugation and settlement in seventeenth-century Siberia. In A. Wood (Ed.), *The history of Siberia: from Russian conquest to revolution* (pp. 37-56). London ; New York: Routledge.
- Comas, D., Calafell, F., Mateu, E., Perez-Lezaun, A., Bosch, E., Martinez-Arias, R., et al. (1998). Trading genes along the silk road: mtDNA sequences and the origin of central Asian populations. *Am J Hum Genet*, 63(6), 1824-1838.
- Comas, D., Plaza, S., Wells, R. S., Yuldaseva, N., Lao, O., Calafell, F., et al. (2004). Admixture, migrations, and dispersals in Central Asia: evidence from maternal DNA lineages. *Eur J Hum Genet*, 12(6), 495-504.
- Comrie, B. (1981). *The languages of the Soviet Union*. Cambridge [Eng.] ; New York: Cambridge University Press.
- Conquest, R. (1986). *The harvest of sorrow: Soviet collectivization and the terror-famine*. New York: Oxford University Press.
- Consolazio, C. F., Johnson, R. E., & Pecora, L. J. (1963). *Physiological measurements of metabolic functions in man*. New York: McGraw Hill.
- Cox, M. P. (2006). Minimal hierarchical analysis of global human Y-chromosome SNP diversity by PCR-RFLP. *Anthropol Sci*, 114(1), 69-74.

- Crawford, M. H. (2007). Genetic structure of circumpolar populations: a synthesis. *Am J Hum Biol*, 19(2), 203-217.
- Crawford, M. H., Williams, J. T., & Duggirala, R. (1997). Genetic structure of the indigenous populations of Siberia. *Am J Phys Anthropol*, 104(2), 177-192.
- Cropp, S. J., & Boinski, S. (2000). The Central American squirrel monkey (*Saimiri oerstedii*): introduced hybrid or endemic species? *Mol Phylogenet Evol*, 16(3), 350-365.
- Cropp, S. J., Larson, A., & Cheverud, J. M. (1999). Historical biogeography of tamarins, genus *Saguinus*: the molecular phylogenetic evidence. *Am J Phys Anthropol*, 108(1), 65-89.
- Cruciani, F., Santolamazza, P., Shen, P., Macaulay, V., Moral, P., Olckers, A., et al. (2002). A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. *Am J Hum Genet*, 70(5), 1197-1214.
- de Knijff, P. (2000). Messages through bottlenecks: on the combined use of slow and fast evolving polymorphic markers on the human Y chromosome. *Am J of Hum Genet*, 67(5), 1055-1061.
- Debets, G. F. (1962). The origin of the Kirgiz people in the light of physical anthropological findings. In H. N. Michael (Ed.), *Studies in Siberian Ethnogenesis* (Vol. Artic Institute of North America Anthropology of the North: Translations from Russian Sources, pp. 129-143). Toronto: University of Toronto Press.
- Delghandi, M., Utsi, E., & Krauss, S. (1998). Saami mitochondrial DNA reveals deep maternal lineage clusters. *Hum Hered*, 48(2), 108-114.
- Derbeneva, O. A., Starikovskaia, E. B., Volod'ko, N. V., Wallace, D. C., & Sukernik, R. I. (2002). [Mitochondrial DNA variation in Kets and Nganasans and the early peoples of Northern Eurasia]. *Genetika*, 38(11), 1554-1560.
- Derbeneva, O. A., Starikovskaya, E. B., Wallace, D. C., & Sukernik, R. I. (2002). Traces of early Eurasians in the Mansi of northwest Siberia revealed by mitochondrial DNA analysis. *Am J Hum Genet*, 70(4), 1009-1014.
- Derbeneva, O. A., Sukernik, R. I., Volodko, N. V., Hosseini, S. H., Lott, M. T., & Wallace, D. C. (2002). Analysis of mitochondrial DNA diversity in the Aleuts of the Commander Islands and its implications for the genetic history of Beringia. *Am J Hum Genet*, 71(2), 415-421.
- Derenko, M. V., Denisova, G. A., Malyarchuk, B. A., Dambueva, I. K., Luzina, F. A., Lotosh, E. A., et al. (2001). [Structure of the gene pool of ethnic groups from the

- Altai-Sayan region from data on mitochondrial polymorphism]. *Genetika*, 37(10), 1402-1410.
- Derenko, M. V., Grzybowski, T., Malyarchuk, B. A., Czarny, J., Miscicka-Sliwka, D., & Zakharov, I. A. (2001). The presence of mitochondrial haplogroup X in Altaians from South Siberia. *Am J Hum Genet*, 69(1), 237-241.
- Derenko, M. V., Grzybowski, T., Malyarchuk, B. A., Dambueva, I. K., Denisova, G. A., Czarny, J., et al. (2003). Diversity of mitochondrial DNA lineages in South Siberia. *Ann Hum Genet*, 67(Pt 5), 391-411.
- Derenko, M. V., Maliarchuk, B. A., Denisova, G. A., Dambueva, I. K., Kakpakov, V. T., Dorzhu, C. M., et al. (2002). [Molecular genetic differentiation of ethnic populations in Southern and Eastern Siberia based on mitochondrial DNA polymorphism]. *Genetika*, 38(10), 1409-1416.
- Derenko, M. V., Maliarchuk, B. A., Denisova, G. A., M., D. C., Karamchakova, O. N., Luzina, F. A., et al. (2002). [Polymorphism of the Y-chromosome diallelic loci in the ethnic populations of the Altai-Sayan region]. *Genetika*, 38(3), 393-399.
- Derenko, M. V., Maliarchuk, B. A., & Solovenchuk, L. L. (1996). [The "Indian deletion" in the hypervariable segment II of mitochondrial DNA is absent in representatives of the native population of Northeastern Asia]. *Genetika*, 32(6), 854-855.
- Derenko, M. V., Maliarchuk, B. A., & Zakharov, I. A. (2002). [Origin of caucasoid-specific mitochondrial DNA lineages in the ethnic populations of the Altai-Sayan region]. *Genetika*, 38(9), 1292-1297.
- Derenko, M. V., Malyarchuk, B., Denisova, G., Wozniak, M., Grzybowski, T., Dambueva, I., et al. (2007). Y-chromosome haplogroup N dispersals from south Siberia to Europe. *J Hum Genet*, 52(9), 763-770.
- Derenko, M. V., Malyarchuk, B., Denisova, G. A., Wozniak, M., Dambueva, I., Dorzhu, C., et al. (2006). Contrasting patterns of Y-chromosome variation in South Siberian populations from Baikal and Altai-Sayan regions. *Hum Genet*, 118(5), 591-604.
- Derenko, M. V., Malyarchuk, B., Grzybowski, T., Denisova, G., Dambueva, I., Perkova, M., et al. (2007). Phylogeographic analysis of mitochondrial DNA in northern Asian populations. *Am J Hum Genet*, 81(5), 1025-1041.
- Derenko, M. V., Malyarchuk, B. A., Dambueva, I. K., Shaikhaev, G. O., Dorzhu, C. M., Nimaev, D. D., et al. (2000). Mitochondrial DNA variation in two South Siberian Aboriginal populations: implications for the genetic history of North Asia. *Hum Biol*, 72(6), 945-973.

- Derenko, M. V., Malyarchuk, B. A., Wozniak, M., Denisova, G. A., Dambueva, I. K., Dorzhu, C. M., et al. (2007). [Distribution of the male lineages of Genghis Khan's descendants in northern Eurasian populations]. *Genetika*, 43(3), 422-426.
- Di Cosmo, N. (1994). Ancient Inner Asian Nomads: Their Economic Basis and Its Significance in Chinese History. *The Journal of Asian Studies*, 53(4), 1092-1126.
- Dolukhanov, P. M., Shukurov, A. M., Tarasov, P. E., & Zaitseva, G. I. (2002). Colonization of Northern Eurasia by modern humans: radiocarbon chronology and environment. *Journal of Archaeological Science*, 29(6), 593-606.
- Dorit, R. L., Akashi, H., & Gilbert, W. (1995). Absence of polymorphism at the ZFY locus on the human Y chromosome. *Science*, 268(5214), 1183-1185.
- Dulik, M. C., Gokcumen, O., Zhadanov, S. I., Osipova, L., & Schurr, T. G. (2007). A phylogeographic analysis of haplogroup D5 and its implications for the peopling of East Asia. *Am J Phys Anthropol*, 132(S44), 102.
- Dulik, M. C., Zhadanov, S. I., Osipova, L., & Schurr, T. G. (2006). Mitochondrial DNA variation in northern Altaian ethnic groups. *Am J Phys Anthropol*, 129(S42), 85.
- Dupanloup, I., Schneider, S., & Excoffier, L. (2002). A simulated annealing approach to define the genetic structure of populations. *Mol Ecol*, 11(12), 2571-2581.
- Elson, J. L., Andrews, R. M., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M., & Howell, N. (2001). Analysis of European mtDNAs for recombination. *Am J Hum Genet*, 68(1), 145-153.
- Elson, J. L., & Lightowlers, R. N. (2006). Mitochondrial DNA clonality in the dock: can surveillance swing the case? *Trends Genet*, 22(11), 603-607.
- Elson, J. L., Turnbull, D. M., & Howell, N. (2004). Comparative genomics and the evolution of human mitochondrial DNA: assessing the effects of selection. *Am J Hum Genet*, 74(2), 229-238.
- Elson, J. L., Turnbull, D. M., & Taylor, R. W. (2007). Testing the adaptive selection of human mtDNA haplogroups: an experimental bioenergetics approach. *Biochem J*, 404(2), e3-5.
- Endicott, P., & Ho, S. Y. (2008). A Bayesian evaluation of human mitochondrial substitution rates. *Am J Hum Genet*, 82(4), 895-902.
- Endicott, P., Ho, S. Y., Metspalu, M., & Stringer, C. (2009). Evaluating the mitochondrial timescale of human evolution. *Trends Ecol Evol*, 24(9), 515-521.
- Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theor Popul Biol*, 3(1), 87-112.

- Excoffier, L., Laval, G., & Schneider, S. (2005). Arlequin (version 3.0): An integrated software package for population genetics data analysis. *Evol Bioinform Online*, 1, 47-50.
- Excoffier, L., & Lischer, H. E. L. (2010). Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, 10, 564-567.
- Excoffier, L., Smouse, P. E., & Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, 131(2), 479-491.
- Eyre-Walker, A. (2006). Evolution. Size does not matter for mitochondrial DNA. *Science*, 312(5773), 537-538.
- Eyre-Walker, A., & Awadalla, P. (2001). Does human mtDNA recombine? *J Mol Evol*, 53(4-5), 430-435.
- Eyre-Walker, A., Smith, N. H., & Smith, J. M. (1999). How clonal are human mitochondria? *Proc Biol Sci*, 266(1418), 477-483.
- Fagundes, N. J., Kanitz, R., Eckert, R., Valls, A. C., Bogo, M. R., Salzano, F. M., et al. (2008). 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.
- Ferris, S. D., Brown, W. M., Davidson, W. S., & Wilson, A. C. (1981). Extensive polymorphism in the mitochondrial DNA of apes. *Proc Natl Acad Sci U S A*, 78(10), 6319-6323.
- Ferris, S. D., Wilson, A. C., & Brown, W. M. (1981). Evolutionary tree for apes and humans based on cleavage maps of mitochondrial DNA. *Proc Natl Acad Sci U S A*, 78(4), 2432-2436.
- Finnila, S., Hassinen, I. E., Ala-Kokko, L., & Majamaa, K. (2000). Phylogenetic network of the mtDNA haplogroup U in Northern Finland based on sequence analysis of the complete coding region by conformation-sensitive gel electrophoresis. *Am J Hum Genet*, 66(3), 1017-1026.
- Fisher, R. A. (1930). *The genetical theory of natural selection*. Oxford: The Clarendon press.
- Foresta, C., Moro, E., & Ferlin, A. (2001). Y chromosome microdeletions and alterations of spermatogenesis. *Endocr Rev*, 22(2), 226-239.

- 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.
- Forster, P., Rohl, A., Lunnemann, P., Brinkmann, C., Zerjal, T., Tyler-Smith, C., et al. (2000). A short tandem repeat-based phylogeny for the human Y chromosome. *Am J Hum Genet*, 67(1), 182-196.
- Forsyth, J. (1991). The Siberian native peoples before and after the Russian conquest. In A. Wood (Ed.), *The history of Siberia: from Russian conquest to revolution* (pp. 69-91). London: Routledge.
- Forsyth, J. (1992). *A history of the peoples of Siberia: Russia's north Asian colony, 1581-1990*. Cambridge, England: Cambridge University Press.
- Friedlaender, J. S., Friedlaender, F. R., Hodgson, J. A., Stoltz, M., Koki, G., Horvat, G., et al. (2007). Melanesian mtDNA complexity. *PLoS ONE*, 2(2), e248.
- Friedlaender, J. S., Schurr, T., Gentz, F., Koki, G., Friedlaender, F., Horvat, G., et al. (2005). Expanding Southwest Pacific mitochondrial haplogroups P and Q. *Mol Biol Evol*, 22(6), 1506-1517.
- Fu, Y. X. (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics*, 147(2), 915-925.
- Gagneux, P., Wills, C., Gerloff, U., Tautz, D., Morin, P. A., Boesch, C., et al. (1999). Mitochondrial sequences show diverse evolutionary histories of African hominoids. *Proc Natl Acad Sci U S A*, 96(9), 5077-5082.
- Gayden, T., Cadenas, A. M., Regueiro, M., Singh, N. B., Zhivotovsky, L. A., Underhill, P. A., et al. (2007). The Himalayas as a directional barrier to gene flow. *Am J Hum Genet*, 80(5), 884-894.
- Gerber, A. S., Loggins, R., Kumar, S., & Dowling, T. E. (2001). Does nonneutral evolution shape observed patterns of DNA variation in animal mitochondrial genomes? *Annu Rev Genet*, 35, 539-566.
- Gerrard, D. T., & Filatov, D. A. (2005). Positive and negative selection on mammalian Y chromosomes. *Mol Biol Evol*, 22(6), 1423-1432.
- Giles, R. E., Blanc, H., Cann, H. M., & Wallace, D. C. (1980). Maternal inheritance of human mitochondrial DNA. *Proc Natl Acad Sci U S A*, 77(11), 6715-6719.
- Gimbutas, M. A. e., & Hencken, H. O. N. (1956). *The prehistory of eastern Europe*. Cambridge, Mass.: Peabody Museum.

- Goebel, T. (1999). Pleistocene human colonization of Siberia and peopling of the Americas: An ecological approach. *Evolutionary Anthropology: Issues, News, and Reviews*, 8(6), 208-227.
- Gokcumen, O., Dulik, M. C., Pai, A. A., Zhadanov, S. I., Rubinstein, S., Osipova, L. P., et al. (2008). Genetic variation in the enigmatic Altaian Kazakhs of South-Central Russia: insights into Turkic population history. *Am J Phys Anthropol*, 136(3), 278-293.
- Golden, P. B. (1992). *An introduction to the history of the Turkic peoples: ethnogenesis and state-formation in medieval and early modern Eurasia and the Middle East*. Wiesbaden: Otto Harrassowitz.
- Goldstein, D. B., Ruiz Linares, A., Cavalli-Sforza, L. L., & Feldman, M. W. (1995). An evaluation of genetic distances for use with microsatellite loci. *Genetics*, 139(1), 463-471.
- Goldstein, D. B., Zhivotovsky, L. A., Nayar, K., Linares, A. R., Cavalli-Sforza, L. L., & Feldman, M. W. (1996). Statistical properties of the variation at linked microsatellite loci: implications for the history of human Y chromosomes. *Mol Biol Evol*, 13(9), 1213-1218.
- Golubenko, M. V., Puzyrev, V. P., Saliukov, V. B., Kucher, A. N., & Sanchat, N. O. (2001). [Analysis of distribution of "mongoloid" haplogroups of mitochondrial DNA among indigenous population of the Tuva Republic]. *Genetika*, 37(6), 831-839.
- Gould, S. J. (2002). *The structure of evolutionary theory*. Cambridge, Mass.: Belknap Press of Harvard University Press.
- Graves, J. A. M., Ferguson-Smith, M. A., McLaren, A., Mittwoch, U., Renfree, M. B., & Burgoyne, P. (1995). The evolution of mammalian sex chromosomes and the origin of sex determining genes [and discussion]. *Philosophical Transactions: Biological Sciences*, 350(1333), 305-312.
- Gray, M. W., Burger, G., & Lang, B. F. (1999). Mitochondrial evolution. *Science*, 283(5407), 1476-1481.
- Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., et al. (2010). A draft sequence of the Neandertal genome. *Science*, 328(5979), 710-722.
- Greenberg, B. D., Newbold, J. E., & Sugino, A. (1983). Intraspecific nucleotide sequence variability surrounding the origin of replication in human mitochondrial DNA. *Gene*, 21(1-2), 33-49.
- Grousset, R. (1970). *The Empire of the Steppes: a History of Central Asia* (N. Walford, Trans.). New Brunswick, N.J.: Rutgers University Press.



- Gryaznov, M. P. (1969). *The Ancient Civilization of Southern Siberia* (J. Hogarth, Trans.). New York: Cowles Book Company, Inc.
- Gusmao, L., Butler, J. M., Carracedo, A., Gill, P., Kayser, M., Mayr, W. R., et al. (2006). DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis. *Forensic Sci Int*, 157(2-3), 187-197.
- Hagelberg, E. (2003). Recombination or mutation rate heterogeneity? Implications for Mitochondrial Eve. *Trends Genet*, 19(2), 84-90.
- Hagelberg, E., Goldman, N., Lio, P., Whelan, S., Schiefenhover, W., Clegg, J. B., et al. (1999). Evidence for mitochondrial DNA recombination in a human population of island Melanesia. *Proc R Soc Lond B*, 266, 485-492.
- Hagelberg, E., Goldman, N., Lio, P., Whelan, S., Schiefenhover, W., Clegg, J. B., et al. (2000). Evidence for mitochondrial DNA recombination in a human population of island Melanesia: correction. *Proc R Soc Lond B*, 267, 1595-1596.
- Halemba, A. (2003). Contemporary religious life in the Republic of Altai: the interaction of Buddhism and Shamanism. *Sibirica*, 3(2), 165-182.
- Hammer, M. F. (1995). A recent common ancestry for human Y chromosomes. *Nature*, 378(6555), 376-378.
- Hammer, M. F., Blackmer, F., Garrigan, D., Nachman, M. W., & Wilder, J. A. (2003). Human population structure and its effects on sampling Y chromosome sequence variation. *Genetics*, 164(4), 1495-1509.
- Hammer, M. F., Karafet, T., Rasanayagam, A., Wood, E. T., Altheide, T. K., Jenkins, T., et al. (1998). Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. *Mol Biol Evol*, 15(4), 427-441.
- Hammer, M. F., Karafet, T. M., Redd, A. J., Jarjanazi, H., Santachiara-Benerecetti, S., Soodyall, H., et al. (2001). Hierarchical patterns of global human Y-chromosome diversity. *Mol Biol Evol*, 18(7), 1189-1203.
- Hammer, M. F., Spurdle, A. B., Karafet, T., Bonner, M. R., Wood, E. T., Novelletto, A., et al. (1997). The geographic distribution of human Y chromosome variation. *Genetics*, 145(3), 787-805.
- Hammer, M. F., & Zegura, S. (2002). The human Y chromosome haplogroup tree: nomenclature and phylogeography of its major divisions. *Annu Rev Anthropol*, 31, 303-321.

- Hammer, M. F., & Zegura, S. L. (1996). The role of the Y chromosome in human evolutionary studies. *Evolutionary Anthropology: Issues, News, and Reviews*, 5(4), 116-134.
- Harpending, H., & Rogers, A. (2000). Genetic perspectives on human origins and differentiation. *Annu Rev Genomics Hum Genet*, 1, 361-385.
- Hasegawa, M., Di Rienzo, A., Kocher, T. D., & Wilson, A. C. (1993). Toward a more accurate time scale for the human mitochondrial DNA tree. *J Mol Evol*, 37(4), 347-354.
- Hasegawa, M., & Horai, S. (1991). Time of the deepest root for polymorphism in human mitochondrial DNA. *J Mol Evol*, 32(1), 37-42.
- Hasegawa, M., Kishino, H., Hayasaka, K., & Horai, S. (1990). Mitochondrial DNA evolution in primates: transition rate has been extremely low in the lemur. *J Mol Evol*, 31(2), 113-121.
- Hasegawa, M., Kishino, H., & Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*, 22(2), 160-174.
- Hemphill, B. E., & Mallory, J. P. (2004). Horse-mounted invaders from the Russo-Kazakh steppe or agricultural colonists from western Central Asia? A craniometric investigation of the Bronze Age settlement of Xinjiang. *Am J Phys Anthropol*, 124(3), 199-222.
- Henn, B. M., Gignoux, C. R., Feldman, M. W., & Mountain, J. L. (2009). Characterizing the time dependency of human mitochondrial DNA mutation rate estimates. *Mol Biol Evol*, 26(1), 217-230.
- Herodotus, Waterfield, R., & Dewald, C. (1998). *The histories*. New York: Oxford University Press.
- 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., & Machado, C. A. (2003). The study of structured populations--new hope for a difficult and divided science. *Nat Rev Genet*, 4(7), 535-543.
- Heyer, E., Balaesque, P., Jobling, M. A., Quintana-Murci, L., Chaix, R., Segurel, L., et al. (2009). Genetic diversity and the emergence of ethnic groups in Central Asia. *BMC Genet*, 10(1), 49.

- Heyer, E., Puymirat, J., Dieltjes, P., Bakker, E., & de Knijff, P. (1997). Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. *Hum Mol Genet*, 6(5), 799-803.
- Heyer, E., Zietkiewicz, E., Rochowski, A., Yotova, V., Puymirat, J., & Labuda, D. (2001). Phylogenetic and familial estimates of mitochondrial substitution rates: study of control region mutations in deep-rooting pedigrees. *Am J Hum Genet*, 69(5), 1113-1126.
- Hillis, D. M. (1987). Molecular versus morphological approaches to systematics. *Annual Review of Ecology and Systematics*, 18, 23-42.
- Hixson, J. E., & Brown, W. M. (1986). A comparison of the small ribosomal RNA genes from the mitochondrial DNA of the great apes and humans: sequence, structure, evolution, and phylogenetic implications. *Mol Biol Evol*, 3(1), 1-18.
- Ho, S. Y., & Larson, G. (2006). Molecular clocks: when times are a-changin'. *Trends in Genetics*, 22(2), 79-83.
- Ho, S. Y., Phillips, M. J., Cooper, A., & Drummond, A. J. (2005). Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Mol Biol Evol*, 22(7), 1561-1568.
- Hofmann, S., Jaksch, M., Bezold, R., Mertens, S., Aholt, S., Paprotta, A., et al. (1997). Population genetics and disease susceptibility: characterization of central European haplogroups by mtDNA gene mutations, correlation with D loop variants and association with disease. *Hum Mol Genet*, 6(11), 1835-1846.
- Horai, S., Hayasaka, K., Kondo, R., Tsugane, K., & Takahata, N. (1995). Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. *Proc Natl Acad Sci U S A*, 92(2), 532-536.
- Horovitz, I., & Meyer, A. (1995). Systematics of New World monkeys (Platyrrhini, Primates) based on 16S mitochondrial DNA sequences: a comparative analysis of different weighting methods in cladistic analysis. *Mol Phylogenet Evol*, 4(4), 448-456.
- Howell, N., Elson, J. L., Howell, C., & Turnbull, D. M. (2007). Relative rates of evolution in the coding and control regions of African mtDNAs. *Mol Biol Evol*, 24(10), 2213-2221.
- Howell, N., Elson, J. L., Turnbull, D. M., & Herrnstadt, C. (2004). African Haplogroup L mtDNA sequences show violations of clock-like evolution. *Mol Biol Evol*, 21(10), 1843-1854.
- Howell, N., Kubacka, I., & Mackey, D. A. (1996). How rapidly does the human mitochondrial genome evolve? *Am J Hum Genet*, 59(3), 501-509.

- 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.
- Hudjashov, G., Kivisild, T., Underhill, P. A., Endicott, P., Sanchez, J. J., Lin, A. A., et al. (2007). Revealing the prehistoric settlement of Australia by Y chromosome and mtDNA analysis. *Proc Natl Acad Sci U S A*, 104(21), 8726-8730.
- Hughes, A. L. (2008). Near neutrality: leading edge of the neutral theory of molecular evolution. *Ann N Y Acad Sci*, 1133, 162-179.
- Hughes, J. F., Skaletsky, H., Pyntikova, T., Graves, T. A., van Daalen, S. K. M., Minx, P. J., et al. (2010). Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature*, 463(7280), 536-539.
- Hunley, K. L., Healy, M. E., & Long, J. C. (2009). The global pattern of gene identity variation reveals a history of long-range migrations, bottlenecks, and local mate exchange: implications for biological race. *Am J Phys Anthropol*, 139(1), 35-46.
- Hurles, M. E., Veitia, R., Arroyo, E., Armenteros, M., Bertranpetit, J., Perez-Lezaun, A., et al. (1999). Recent male-mediated gene flow over a linguistic barrier in Iberia, suggested by analysis of a Y-chromosomal DNA polymorphism. *Am J Hum Genet*, 65(5), 1437-1448.
- Hurst, G. D., & Jiggins, F. M. (2005). Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts. *Proc Biol Sci*, 272(1572), 1525-1534.
- Ingman, M., & Gyllensten, U. (2007a). Rate variation between mitochondrial domains and adaptive evolution in humans. *Hum Mol Genet*, 16(19), 2281-2287.
- Ingman, M., & Gyllensten, U. (2007b). A recent genetic link between Sami and the Volga-Ural region of Russia. *Eur J Hum Genet*, 15(1), 115-120.
- Ingman, M., Kaessmann, H., Paabo, S., & Gyllensten, U. (2000). Mitochondrial genome variation and the origin of modern humans. *Nature*, 408(6813), 708-713.
- Innan, H., & Nordborg, M. (2002). Recombination or mutational hot spots in human mtDNA? *Mol Biol Evol*, 19(7), 1122-1127.
- Irwin, D. M., Kocher, T. D., & Wilson, A. C. (1991). Evolution of the cytochrome b gene of mammals. *J Mol Evol*, 32(2), 128-144.
- ISOGG. (2010). Y-DNA Haplogroup Tree 2010, Version: 5.31. Retrieved 04 November, 2010

- Jettmar, K. (1950). The Karasuk Culture and its south-eastern affinities. *Bulletin of The Museum of Far Eastern Antiquities*, 22, 83-126.
- Jettmar, K. (1951). The Altai before the Turks. *Bulletin of The Museum of Far Eastern Antiquities*, 23, 135-260.
- Jin, H. J., Kim, K. C., & Kim, W. (2010). Genetic diversity of two haploid markers in the Udegey population from southeastern Siberia. *Am J Phys Anthropol*, 142(2), 303-313.
- Jobling, M. A., Hurles, M., & Tyler-Smith, C. (2004). *Human evolutionary genetics: origins, peoples & disease*. New York: Garland Science.
- Jobling, M. A., & Tyler-Smith, C. (1995). Fathers and sons: the Y chromosome and human evolution. *Trends Genet*, 11(11), 449-456.
- Jobling, M. A., & Tyler-Smith, C. (2000). New uses for new haplotypes the human Y chromosome, disease and selection. *Trends Genet*, 16(8), 356-362.
- Jobling, M. A., & Tyler-Smith, C. (2003). The human Y chromosome: an evolutionary marker comes of age. *Nat Rev Genet*, 4(8), 598-612.
- Jobling, M. A., Williams, G. A., Schiebel, G. A., Pandya, G. A., McElreavey, G. A., Salas, G. A., et al. (1998). A selective difference between human Y-chromosomal DNA haplotypes. *Curr Biol*, 8(25), 1391-1394.
- Johnson, M. J., Wallace, D. C., Ferris, S. D., Rattazzi, M. C., & Cavalli-Sforza, L. L. (1983). Radiation of human mitochondria DNA types analyzed by restriction endonuclease cleavage patterns. *J Mol Evol*, 19(3-4), 255-271.
- Jordana, X., Galtés, I., Turbat, T., Batsukh, D., García, C., Isidro, A., et al. (2009). The warriors of the steppes: osteological evidence of warfare and violence from Pazyryk tumuli in the Mongolian Altai. *Journal of Archaeological Science*, 36(7), 1319-1327.
- Jorde, L. B., & Bamshad, M. (2000). Questioning evidence for recombination in human mitochondrial DNA. *Science*, 288(5473), 1931.
- Karafet, T. M., Mendez, F. L., Meilerman, M. B., Underhill, P. A., Zegura, S. L., & Hammer, M. F. (2008). New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res*, 18(5), 830-838.
- Karafet, T. M., Osipova, L. P., Gubina, M. A., Posukh, O. L., Zegura, S. L., & Hammer, M. F. (2002). High levels of Y-chromosome differentiation among native Siberian populations and the genetic signature of a boreal hunter-gatherer way of life. *Hum Biol*, 74(6), 761-789.

- Karafet, T. M., & Sukernik, R. I. (1978). [Genetic structure of an isolated group of the native population of northern Siberia, the Nganasani (Tavgi) of the Taimyr. III. A family analysis of blood groups]. *Genetika*, 14(3), 527-531.
- Karafet, T. M., Xu, L., Du, R., Wang, W., Feng, S., Wells, R. S., et al. (2001). Paternal population history of East Asia: sources, patterns, and microevolutionary processes. *Am J Hum Genet*, 69(3), 615-628.
- Karafet, T. M., Zegura, S. L., Posukh, O., Osipova, L., Bergen, A., Long, J., et al. (1999). Ancestral Asian source(s) of new world Y-chromosome founder haplotypes. *Am J Hum Genet*, 64(3), 817-831.
- Karafet, T. M., Zegura, S. L., Vuturo-Brady, J., Posukh, O., Osipova, L., Wiebe, V., et al. (1997). Y chromosome markers and Trans-Bering Strait dispersals. *Am J Phys Anthropol*, 102(3), 301-314.
- Kayser, M., Brauer, S., Cordaux, R., Casto, A., Lao, O., Zhivotovsky, L. A., et al. (2006). Melanesian and Asian origins of Polynesians: mtDNA and Y chromosome gradients across the Pacific. *Mol Biol Evol*, 23(11), 2234-2244.
- Kayser, M., Brauer, S., Weiss, G., Schiefenhover, W., Underhill, P., Shen, P., et al. (2003). Reduced Y-chromosome, but not mitochondrial DNA, diversity in human populations from West New Guinea. *Am J Hum Genet*, 72(2), 281-302.
- Kayser, M., Krawczak, M., Excoffier, L., Deltjes, P., Corach, D., Pascali, V., et al. (2001). An extensive analysis of Y-chromosomal microsatellite haplotypes in globally dispersed human populations. *Am J Hum Genet*, 68(4), 990-1018.
- Kayser, M., Roewer, L., Hedman, M., Henke, L., Henke, J., Brauer, S., et al. (2000). Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. *Am J Hum Genet*, 66(5), 1580-1588.
- Keyser-Tracqui, C., Crubezy, E., & Ludes, B. (2003). Nuclear and mitochondrial DNA analysis of a 2,000-year-old necropolis in the Egyin Gol Valley of Mongolia. *Am J Hum Genet*, 73(2), 247-260.
- Keyser, C., Bouakaze, C., Crubezy, E., Nikolaev, V. G., Montagnon, D., Reis, T., et al. (2009). Ancient DNA provides new insights into the history of south Siberian Kurgan people. *Hum Genet*, 126(3), 395-410.
- Khar'kov, V. N., Stepanov, V. A., Medvedev, O. F., Spiridonova, M. G., Maksimova, N. R., Nogovitsyna, A. N., et al. (2008). [The origin of Yakuts: analysis of Y-chromosome haplotypes]. *Mol Biol (Mosk)*, 42(2), 226-237.

- Kharkov, V. N., Medvedeva, O. F., Luzina, F. A., Kolbasko, A. V., Gafarov, N. I., Puzyrev, V. P., et al. (2009). [Comparative characteristics of the gene pool of Teleuts inferred from Y-chromosomal marker data]. *Genetika*, 45(8), 1132-1142.
- Khodzhayov, T. K. (2008). Cranial characteristics of the Saka populations of the eastern Pamirs. *Archaeol Ethnol Anthropol Eurasia*, 35(3), 143-156.
- Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, 217(5129), 624-626.
- Kimura, M. (1977). Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature*, 267(5608), 275-276.
- Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge, England: Cambridge University Press.
- Kingman, J. F. C. (1982). The coalescent. *Stoch Proc Appl*, 13, 235-248.
- Kivisild, T., Bamshad, M. J., Kaldma, K., Metspalu, M., Metspalu, E., Reidla, M., et al. (1999). Deep common ancestry of Indian and western-Eurasian mitochondrial DNA lineages. *Curr Biol*, 9(22), 1331-1334.
- Kivisild, T., Rootsi, S., Metspalu, M., Mastana, S., Kaldma, K., Parik, J., et al. (2003). The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. *Am J Hum Genet*, 72(2), 313-332.
- 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.
- Kivisild, T., Tolk, H. V., Parik, J., Wang, Y., Papiha, S. S., Bandelt, H. J., et al. (2002). The emerging limbs and twigs of the East Asian mtDNA tree. *Mol Biol Evol*, 19(10), 1737-1751.
- Kivisild, T., & Villems, R. (2000). Questioning evidence for recombination in human mitochondrial DNA. *Science*, 288(5473), 1931a.
- Knowles, L. L., & Maddison, W. P. (2002). Statistical phylogeography. *Mol Ecol*, 11(12), 2623-2635.
- Kocher, T. D., Thomas, W. K., Meyer, A., Edwards, S. V., Paabo, S., Villablanca, F. X., et al. (1989). Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers. *Proc Natl Acad Sci U S A*, 86(16), 6196-6200.
- 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.

- Kondo, R., Horai, S., Satta, Y., & Takahata, N. (1993). Evolution of hominoid mitochondrial DNA with special reference to the silent substitution rate over the genome. *J Mol Evol*, 36(6), 517-531.
- Kong, Q. P., Yao, Y. G., Liu, M., Shen, S. P., Chen, C., Zhu, C. L., et al. (2003). Mitochondrial DNA sequence polymorphisms of five ethnic populations from northern China. *Hum Genet*, 113(5), 391-405.
- Kong, Q. P., Yao, Y. G., Sun, C., Bandelt, H. J., Zhu, C. L., & Zhang, Y. P. (2003). Phylogeny of east Asian mitochondrial DNA lineages inferred from complete sequences. *Am J Hum Genet*, 73(3), 671-676.
- Kovtun, I. V. (2008). The bear image in western Siberian art of the 2nd millennium BC and its relevance for delimiting the eastern periphery of the Samus culture. *Archaeology, Ethnology and Anthropology of Eurasia*, 35(3), 97-104.
- Kozintsev, A. G. (2000). On the biological affinities and origin of the Northern Pontic Scythians. *Archaeol Ethnol Anthropol Eurasia*, 3(3), 145-152.
- Kozintsev, A. G. (2007). Scythians of the North Pontic region: Between-group cranial variation, affinities, and origins. *Archaeology, Ethnology and Anthropology of Eurasia*, 32(1), 143-157.
- Kozintsev, A. G. (2008). The "Mediterraneans" of southern Siberia and Kazakhstan, Indo-European migrations, and the origin of the Scythians: a multivariate craniometric analysis. *Archaeology, Ethnology and Anthropology of Eurasia*, 36(4), 140-144.
- Kozintsev, A. G. (2009). Craniometric evidence of the early caucasoid migrations to Siberia and eastern Central Asia, with reference to the Indo-European problem. *Archaeology, Ethnology and Anthropology of Eurasia*, 37(4), 125-136.
- Krause, J., Briggs, A. W., Kircher, M., Maricic, T., Zwyns, N., Derevianko, A., et al. (2010). A complete mtDNA genome of an early modern human from Kostenki, Russia. *Curr Biol*, 20(3), 231-236.
- Krause, J., Fu, Q., Good, J. M., Viola, B., Shunkov, M. V., Derevianko, A. P., et al. (2010). The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature*, 464(7290), 894-897.
- Krause, J., Orlando, L., Serre, D., Viola, B., Prufer, K., Richards, M. P., et al. (2007). Neanderthals in central Asia and Siberia. *Nature*, 449(7164), 902-904.
- Krausz, C., Quintana-Murci, L., Meyts, E. R.-D., Jorgensen, N., Jobling, M. A., Rosser, Z. H., et al. (2001). Identification of a Y chromosome haplogroup associated with reduced sperm counts. *Hum. Mol. Genet.*, 10(18), 1873-1877.



- Kreitman, M. (2000). Methods to detect selection in populations with applications to the human. *Annual Review of Genomics and Human Genetics*, 1(1), 539-559.
- Kryazhimskiy, S., & Plotkin, J. B. (2008). The population genetics of dN/dS. *PLoS Genet*, 4(12), e1000304.
- Kumar, S., Hedrick, P., Dowling, T., & Stoneking, M. (2000). Questioning evidence for recombination in human mitochondrial DNA. *Science*, 288(5473), 1931.
- Kuroki, Y., Toyoda, A., Noguchi, H., Taylor, T. D., Itoh, T., Kim, D. S., et al. (2006). Comparative analysis of chimpanzee and human Y chromosomes unveils complex evolutionary pathway. *Nat Genet*, 38(2), 158-167.
- Kuzmin, Y. V., & Keates, S. G. (2005). Dates are not just data: Paleolithic settlement patterns in Siberia derived from radiocarbon records. *American Antiquity*, 70(4), 773-789.
- Kuzmina, E. E., & Mair, V. H. (2008). *The prehistory of the Silk Road*. Philadelphia: University of Pennsylvania Press.
- Lahermo, P., Sajantila, A., Sistonen, P., Lukka, M., Aula, P., Peltonen, L., et al. (1996). The genetic relationship between the Finns and the Finnish Saami (Lapps): analysis of nuclear DNA and mtDNA. *Am J Hum Genet*, 58(6), 1309-1322.
- Lahn, B. T., Pearson, N. M., & Jégalian, K. (2001). The human Y chromosome, in the light of evolution. *Nat Rev Genet*, 2(3), 207-216.
- Lappalainen, T., Laitinen, V., Salmela, E., Andersen, P., Huoponen, K., Savontaus, M. L., et al. (2008). Migration waves to the Baltic Sea region. *Ann Hum Genet*, 72(Pt 3), 337-348.
- Larichev, V., Khol'ushkin, U., & Laricheva, I. (1987). Lower and middle Paleolithic of northern Asia: achievements, problems, and perspectives. *Journal of World Prehistory*, 1(4), 415-464.
- Legrand, S., & Bokovenko, N. (2006). The emergence of the Scythians: Bronze Age to Iron Age in South Siberia. *Antiquity*, 80, 843-879.
- Lell, J. T., Brown, M. D., Schurr, T. G., Sukernik, R. I., Starikovskaya, Y. B., Torroni, A., et al. (1997). Y chromosome polymorphisms in Native American and Siberian populations: identification of Native American Y chromosome haplotypes. *Hum Genet*, 100(5-6), 536-543.
- Lell, J. T., Sukernik, R. I., Starikovskaya, Y. B., Su, B., Jin, L., Schurr, T. G., et al. (2002). The dual origin and Siberian affinities of Native American Y chromosomes. *Am J Hum Genet*, 70(1), 192-206.

- Leonard, W. R., Sorensen, M. V., Galloway, V. A., Spencer, G. J., Mosher, M. J., Osipova, L., et al. (2002). Climatic influences on basal metabolic rates among circumpolar populations. *Am J Hum Biol*, 14(5), 609-620.
- Levin, M. G. (1964). The anthropological types of Siberia. In M. G. Levin & L. P. Potapov (Eds.), *The Peoples of Siberia* (pp. 99-104). Chicago: The University of Chicago Press.
- Levin, M. G., & Potapov, L. P. (1964). *The peoples of Siberia* (S. P. Dunn, Trans.). Chicago: University of Chicago Press.
- Li, C., Li, H., Cui, Y., Xie, C., Cai, D., Li, W., et al. (2010). Evidence that a West-East admixed population lived in the Tarim Basin as early as the early Bronze Age. *BMC Biol*, 8, 15.
- Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., et al. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319(5866), 1100-1104.
- Lin, S. J., Tanaka, K., Leonard, W., Gerelsaikhhan, T., Dashnyam, B., Nyamkhishig, S., et al. (1994). A Y-associated allele is shared among a few ethnic groups of Asia. *Jpn J Hum Genet*, 39(3), 299-304.
- Loogvali, E.-L., Kivisild, T., Margus, T. n., & Villems, R. (2009). Explaining the imperfection of the molecular clock of Hominid mitochondria. *PLoS ONE*, 4(12), e8260.
- Macaulay, V., Richards, M., Hickey, E., Vega, E., Cruciani, F., Guida, V., et al. (1999). The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. *Am J Hum Genet*, 64(1), 232-249.
- Macaulay, V., Richards, M. B., Forster, P., Bendall, K. E., Watson, E., Sykes, B., et al. (1997). mtDNA mutation rates--no need to panic. *Am J Hum Genet*, 61(4), 983-990.
- Malhi, R. S., Cybulski, J. S., Tito, R. Y., Johnson, J., Harry, H., & Dan, C. (2010). Brief communication: mitochondrial haplotype C4c confirmed as a founding genome in the Americas. *Am J Phys Anthropol*, 141(3), 494-497.
- Malhi, R. S., Gonzalez-Oliver, A., Schroeder, K. B., Kemp, B. M., Greenberg, J. A., Dobrowski, S. Z., et al. (2008). Distribution of Y chromosomes among Native North Americans: A study of Athapaskan population history. *Am J Phys Anthropol*, 137(4), 412-424.
- Mallory, J. P. (1976). The chronology of the early Kurgan tradition *Journal of Indo-European Studies*, 4(4), 257-294.

- Mallory, J. P. (1977). The chronology of the early Kurgan tradition 2. *Journal of Indo-European Studies*, 5(4), 339-368.
- Mallory, J. P. (1989). *In search of the Indo-Europeans: language, archaeology and myth*. London: Thames and Hudson.
- Mallory, J. P., & Mair, V. H. (2000). *The Tarim mummies: ancient China and the mystery of the earliest peoples from the West*. New York: Thames & Hudson.
- Malyarchuk, B. A. (2004). [Differentiation of the mitochondrial subhaplogroup U4 in the populations of Eastern Europe, Ural, and Western Siberia: implication to the genetic history of the Uralic populations]. *Genetika*, 40(11), 1549-1556.
- Malyarchuk, B. A., Derenko, M., Denisova, G., Wozniak, M., Grzybowski, T., Dambueva, I., et al. (2010). Phylogeography of the Y-chromosome haplogroup C in northern Eurasia. *Ann Hum Genet*, . *In press*.
- Malyarchuk, B. A., Derenko, M., Grzybowski, T., Perkova, M., Rogalla, U., Vanecek, T., et al. (2010). The peopling of Europe from the mitochondrial haplogroup U5 perspective. *PLoS ONE*, 5(4), e10285.
- Malyarchuk, B. A., Derenko, M. B., Balmysheva, N. P., Lapinskii, A. G., & Solovenchuk, L. L. (1994). [Restriction polymorphism of the main noncoding region of mitochondrial DNA in native and migrant inhabitants of Northeast Asia]. *Genetika*, 30(4), 542-545.
- Malyarchuk, B. A., & Derenko, M. V. (1995). [Polymorphism of mitochondrial DNA region V in native and migrant inhabitants of Northeast Asia]. *Genetika*, 31(9), 1308-1313.
- Malyarchuk, B. A., & Derenko, M. V. (1999). Molecular instability of the mitochondrial haplogroup T sequences at nucleotide positions 16292 and 16296. *Ann Hum Genet*, 63(6), 489-497.
- Malyarchuk, B. A., Derenko, M. V., & Solovenchuk, L. L. (1994). [Restriction types of main noncoding segments of mitochondrial DNA in aboriginal and migrant inhabitants of Northeast Asia]. *Genetika*, 30(6), 851-857.
- Malyarchuk, B. A., Grzybowski, T., Derenko, M., Perkova, M., Vanecek, T., Lazur, J., et al. (2008). Mitochondrial DNA phylogeny in Eastern and Western Slavs. *Mol Biol Evol*, 25(8), 1651-1658.
- Malyarchuk, B. A., Grzybowski, T., Derenko, M. V., Czarny, J., Wozniak, M., & Miscicka-Sliwka, D. (2002). Mitochondrial DNA variability in Poles and Russians. *Ann Hum Genet*, 66(4), 261-283.

- Malyarchuk, B. A., Lapinskii, A. G., Balmysheva, N. P., Butorina, O. T., & Solovenchuk, L. L. (1994). [Mitochondrial DNA RFLP in inhabitants of Magadan]. *Genetika*, 30(1), 112-114.
- Malyarchuk, B. A., Rogozin, I. B., Berikov, V. B., & Derenko, M. V. (2002). Analysis of phylogenetically reconstructed mutational spectra in human mitochondrial DNA control region. *Hum Genet*, 111(1), 46-53.
- Manica, A., Prugnolle, F., & Balloux, F. (2005). Geography is a better determinant of human genetic differentiation than ethnicity. *Hum Genet*, 118(3-4), 366-371.
- Marais, G. A., Campos, P. R., & Gordo, I. (2010). Can intra-Y gene conversion oppose the degeneration of the human Y chromosome? A simulation study. *Genome Biol Evol*, 2, 347-357.
- Margulis, L. (1970). *Origin of eukaryotic cells; evidence and research implications for a theory of the origin and evolution of microbial, plant, and animal cells on the Precambrian earth*. New Haven,: Yale University Press.
- Martynova, G. S. (1988). The beginning of the Hunnic epoch in South Siberia. *Arctic Anthropology*, 25(2), 61-83.
- Mayr, E. (1963). *Animal species and evolution*. Cambridge, Mass.: Belknap Press of Harvard Univ. Press.
- McElreavey, K., & Quintana-Murci, L. (2003). Male reproductive function and the human Y chromosome: is selection acting on the Y? *Reprod Biomed Online*, 7(1), 17-23.
- Meiklejohn, C. D., Montooth, K. L., & Rand, D. M. (2007). Positive and negative selection on the mitochondrial genome. *Trends Genet*, 23(6), 259-263.
- Menges, K. H. (1968). *The Turkic languages and peoples: an introduction to Turkic studies*. Wiesbaden: Otto Harrassowitz.
- Merriwether, D. A., Hall, W. W., Vahlne, A., & Ferrell, R. E. (1996). mtDNA variation indicates Mongolia may have been the source for the founding population for the New World. *Am J Hum Genet*, 59(1), 204-212.
- Metspalu, M., Kivisild, T., Metspalu, E., Parik, J., Hudjashov, G., Kaldma, K., et al. (2004). Most of the extant mtDNA boundaries in south and southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. *BMC Genet*, 5, 26.
- Meyer, S., Weiss, G., & von Haeseler, A. (1999). Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA. *Genetics*, 152(3), 1103-1110.

- Mirabal, S., Regueiro, M., Cadenas, A. M., Cavalli-Sforza, L. L., Underhill, P. A., Verbenko, D. A., et al. (2009). Y-Chromosome distribution within the geo-linguistic landscape of northwestern Russia. *Eur J Hum Genet*, 17(10), 1260-1273.
- 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.
- MITOMAP. (2009). MITOMAP: A Human Mitochondrial Genome Database. from <http://www.mitomap.org>
- Moilanen, J. S. (2003). *Non-neutral sequence variation in human mitochondrial DNA: selection against deleterious mutations and haplogroup-related polymorphisms*. University of Oulu.
- Moilanen, J. S., Finnila, S., & Majamaa, K. (2003). Lineage-specific selection in human mtDNA: lack of polymorphisms in a segment of MTND5 gene in haplogroup J. *Mol Biol Evol*, 20(12), 2132-2142.
- Moilanen, J. S., & Majamaa, K. (2003). Phylogenetic network and physicochemical properties of nonsynonymous mutations in the protein-coding genes of human mitochondrial DNA. *Mol Biol Evol*, 20(8), 1195-1210.
- Moiseyev, V. (2006). Nonmetric traits in Early Iron Age cranial series from Western and Southern Siberia. *Archaeology, Ethnology and Anthropology of Eurasia*, 25(1), 145-152.
- Mooder, K. P., Schurr, T. G., Bamforth, F. J., Bazaliiski, V. I., & Savel'ev, N. A. (2006). Population affinities of Neolithic Siberians: a snapshot from prehistoric Lake Baikal. *Am J Phys Anthropol*, 129(3), 349-361.
- Moran, P. A. (1975). Wandering distributions and the electrophoretic profile. *Theor Popul Biol*, 8(3), 318-330.
- Muller, H. J. (1918). Genetic Variability, Twin Hybrids and Constant Hybrids, in a Case of Balanced Lethal Factors. *Genetics*, 3(5), 422-499.
- Muller, H. J. (1964). The Relation of Recombination to Mutational Advance. *Mutat Res*, 106, 2-9.
- Mulligan, C. J., Kitchen, A., & Miyamoto, M. M. (2006). Comment on "Population size does not influence mitochondrial genetic diversity in animals". *Science*, 314(5804), 1390.

- Murphy, E., Gokhman, I., Chistov, Y., & Barkova, L. (2002). Prehistoric Old World scalping: new cases from the cemetery of Aymyrlyg, South Siberia. *American Journal of Archaeology*, *106*(1), 1-10.
- Myres, N. M., Rootsi, S., Lin, A. A., Jarve, M., King, R. J., Kutuev, I., et al. (2010). A major Y-chromosome haplogroup R1b Holocene era founder effect in Central and Western Europe. *Eur J Hum Genet*, *19*(1), 95-101.
- Nachman, M. W., Brown, W. M., Stoneking, M., & Aquadro, C. F. (1996). Nonneutral mitochondrial DNA variation in humans and chimpanzees. *Genetics*, *142*(3), 953-963.
- Naumov, I. V., & Collins, D. N. (2006). *The history of Siberia*. London: Routledge.
- Naumova, O., Khaiat, S., & Rychkov, S. (2009). [Mitochondrial DNA diversity in Kazym Khanty]. *Genetika*, *45*(6), 857-861.
- Naumova, O., Rychkov, S., Morozova, I., Khaiat, S., Semikov, A. V., & Zhukova, O. V. (2008). [Mitochondrial DNA diversity in Siberian Tatars of the Tobol-Irtysh basin]. *Genetika*, *44*(2), 257-268.
- Nazarenko, S. A., & Puzyrev, V. P. (1985). Genetic drift of marker Y chromosome del(Y)(q12) in Khanty from the lower Ob river. *Hum Genet*, *71*(2), 100-102.
- Nei, M. (1978). Estimation of Average Heterozygosity and Genetic Distance from a Small Number of Individuals. *Genetics*, *89*(3), 583-590.
- Nei, M. (1987). *Molecular evolutionary genetics*. New York: Columbia University Press.
- Nei, M., & Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci U S A*, *76*(10), 5269-5273.
- Nei, M., & Tajima, F. (1981). DNA polymorphism detectable by restriction endonucleases. *Genetics*, *97*(1), 145-163.
- Nei, M., & Tajima, F. (1985). Evolutionary change of restriction cleavage sites and phylogenetic inference for man and apes. *Mol Biol Evol*, *2*(3), 189-205.
- Nichols, R. (2001). Gene trees and species trees are not the same. *Trends Ecol Evol*, *16*(7), 358-364.
- Novembre, J., & Stephens, M. (2008). Interpreting principal component analyses of spatial population genetic variation. *Nat Genet*, *40*(5), 646-649.
- O'Rourke, D. H., Hayes, M. G., & Carlyle, S. W. (2000). Spatial and temporal stability of mtDNA haplogroup frequencies in native North America. *Hum Biol*, *72*(1), 15-34.

- Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature*, 246(5428), 96-98.
- Ohta, T., & Kimura, M. (1973). A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genet Res*, 22, 201-204.
- Okladnikov, A. P. (1964). Ancient population of Siberia and its culture. In M. G. Levin & L. P. Potapov (Eds.), *The peoples of Siberia* (pp. 13-98). Chicago: The University of Chicago Press.
- Okladnikov, A. P. (1990). Inner Asia at the dawn of history. In D. Sinor (Ed.), *The Cambridge history of early Inner Asia* (pp. 41-96). Cambridge, England: Cambridge University Press.
- Osipova, L. P., & Sukernik, R. I. (1978). [Polymorphism of immunoglobulin Gm- and Km-allotypes in northern Altaians (western Siberia)]. *Genetika*, 14(7), 1272-1275.
- Pakendorf, B., Novgorodov, I. N., Osakovskij, V. L., Danilova, A. P., Protod'jakonov, A. P., & Stoneking, M. (2006). Investigating the effects of prehistoric migrations in Siberia: genetic variation and the origins of Yakuts. *Hum Genet*, 120(3), 334-353.
- Pakendorf, B., Novgorodov, I. N., Osakovskij, V. L., & Stoneking, M. (2007). Mating patterns amongst Siberian reindeer herders: inferences from mtDNA and Y-chromosomal analyses. *Am J Phys Anthropol*, 133(3), 1013-1027.
- Pakendorf, B., & Stoneking, M. (2005). Mitochondrial DNA and human evolution. *Annu Rev Genomics Hum Genet*, 6, 165-183.
- Pakendorf, B., Wiebe, V., Tarskaia, L. A., Spitsyn, V. A., Soodyall, H., Rodewald, A., et al. (2003). Mitochondrial DNA evidence for admixed origins of central Siberian populations. *Am J Phys Anthropol*, 120(3), 211-224.
- Pala, M., Achilli, A., Olivieri, A., Kashani, B. H., Perego, U. A., Sanna, D., et al. (2009). Mitochondrial haplogroup U5b3: a distant echo of the Epipaleolithic in Italy and the legacy of the early Sardinians. *Am J Hum Genet*, 84(6), 814-821.
- Palanichamy, M. G., Sun, C., Agrawal, S., Bandelt, H. J., Kong, Q. P., Khan, F., et al. (2004). Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing: implications for the peopling of South Asia. *Am J Hum Genet*, 75(6), 966-978.
- Parsch, J., Zhang, Z., & Baines, J. F. (2009). The influence of demography and weak selection on the McDonald-Kreitman Test: an empirical study in *Drosophila*. *Mol Biol Evol*, 26(3), 691-698.

- Parsons, T. J., & Irwin, J. A. (2000). Questioning evidence for recombination in human mitochondrial DNA. *Science*, 288(5473), 1931.
- Parsons, T. J., Muniec, D. S., Sullivan, K., Woodyatt, N., Alliston-Greiner, R., Wilson, M. R., et al. (1997). A high observed substitution rate in the human mitochondrial DNA control region. *Nat Genet*, 15(4), 363-368.
- Passarino, G., Semino, O., Magri, C., Al-Zahery, N., Benuzzi, G., Quintana-Murci, L., et al. (2001). The 49a,f haplotype 11 is a new marker of the EU19 lineage that traces migrations from northern regions of the Black Sea. *Hum Immunol*, 62(9), 922-932.
- Pastorini, J., Martin, R. D., Ehresmann, P., Zimmermann, E., & Forstner, M. R. (2001). Molecular phylogeny of the lemur family cheirogaleidae (primates) based on mitochondrial DNA sequences. *Mol Phylogenet Evol*, 19(1), 45-56.
- Pena, S. D., Santos, F. R., Bianchi, N. O., Bravi, C. M., Carnese, F. R., Rothhammer, F., et al. (1995). A major founder Y-chromosome haplotype in Amerindians. *Nat Genet*, 11(1), 15-16.
- 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.
- Pereira, L., Freitas, F., Fernandes, V., Pereira, J. B., Costa, M. D., Costa, S., et al. (2009). The diversity present in 5140 human mitochondrial genomes. *Am J Hum Genet*, 84(5), 628-640.
- Pereira, L., Macaulay, V., Torroni, A., Scozzari, R., Prata, M. J., & Amorim, A. (2001). Prehistoric and historic traces in the mtDNA of Mozambique: insights into the Bantu expansions and the slave trade. *Ann Hum Genet*, 65(5), 439-458.
- Perez-Lezaun, A., Calafell, F., Seielstad, M., Mateu, E., Comas, D., Bosch, E., et al. (1997). Population genetics of Y-chromosome short tandem repeats in humans. *J Mol Evol*, 45(3), 265-270.
- Phillips-Krawczak, C., Devor, E., Zlojutro, M., Moffat-Wilson, K., & Crawford, M. H. (2006). MtDNA variation in the Altai-Kizhi population of southern Siberia: a synthesis of genetic variation. *Hum Biol*, 78(4), 477-494.
- Pimenoff, V. N., Comas, D., Palo, J. U., Vershubsky, G., Kozlov, A., & Sajantila, A. (2008). Northwest Siberian Khanty and Mansi in the junction of West and East Eurasian gene pools as revealed by uniparental markers. *Eur J Hum Genet*, 16(10), 1254-1264.



- Poehlman, E. T., & Toth, M. J. (1995). Mathematical ratios lead to spurious conclusions regarding age- and sex-related differences in resting metabolic rate. *Am J Clin Nutr*, 61(3), 482-485.
- Popov, A. A., & Dolgikh, B. O. (1964). The Kets. In M. G. Levin & L. P. Potapov (Eds.), *The peoples of Siberia* (pp. 607-619). Chicago: The University of Chicago Press.
- Potapov, L. P. (1962). The origins of the Altayans. In H. N. Michael (Ed.), *Studies in Siberian ethnogenesis* (Vol. Arctic Institute of North America Anthropology of the North: translations from Russian sources, pp. 169-196). Toronto: University of Toronto Press.
- Potapov, L. P. (1964a). The Altays. In M. G. Levin & L. P. Potapov (Eds.), *The Peoples of Siberia* (pp. 305-341). Chicago: The University of Chicago Press.
- Potapov, L. P. (1964b). Historical-ethnographic survey of the Russian population of Siberia in the prerevolutionary period. In M. G. Levin & L. P. Potapov (Eds.), *The peoples of Siberia* (pp. 105-199). Chicago: University of Chicago Press.
- Potapov, L. P. (1964c). The Khakasy. In M. G. Levin & L. P. Potapov (Eds.), *The peoples of Siberia* (pp. 342-379). Chicago: University of Chicago Press.
- Potapov, L. P. (1964d). The Shors. In M. G. Levin & L. P. Potapov (Eds.), *The peoples of Siberia* (pp. 440-473). Chicago: University of Chicago Press.
- Potapov, L. P. (1964e). The Tuvans. In M. G. Levin & L. P. Potapov (Eds.), *The peoples of Siberia* (pp. 380-422). Chicago: University of Chicago Press.
- Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., & Feldman, M. W. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol Biol Evol*, 16(12), 1791-1798.
- Provine, W. B. (1971). *The origins of theoretical population genetics*. Chicago: University of Chicago Press.
- Puzyrev, V. P., Stepanov, V. A., Golubenko, M. V., Puzyrev, K. V., Maksimova, N. R., Khar'kov, V. N., et al. (2003). [MtDNA and Y-chromosome lineages in the Yakut population]. *Genetika*, 39(7), 975-981.
- Quintana-Murci, L., Chaix, R., Wells, R. S., Behar, D. M., Sayar, H., Scozzari, R., et al. (2004). Where west meets east: the complex mtDNA landscape of the southwest and Central Asian corridor. *Am J Hum Genet*, 74(5), 827-845.
- Quintana-Murci, L., Krausz, C., & McElreavey, K. (2001). The human Y chromosome: function, evolution and disease. *Forensic Science International*, 118(2-3), 169-181.

- Quintana-Murci, L., Krausz, C., Zerjal, T., Sayar, S. H., Hammer, M. F., Mehdi, S. Q., et al. (2001). Y-chromosome lineages trace diffusion of people and languages in southwestern Asia. *Am J Hum Genet*, 68(2), 537-542.
- Ramakrishnan, U., & Mountain, J. L. (2004). Precision and accuracy of divergence time estimates from STR and SNPSTR variation. *Mol Biol Evol*, 21(10), 1960-1971.
- Regueiro, M., Cadenas, A. M., Gayden, T., Underhill, P. A., & Herrera, R. J. (2006). Iran: tricontinental nexus for Y-chromosome driven migration. *Hum Hered*, 61(3), 132-143.
- 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.
- Repping, S., Skaletsky, H., Brown, L., van Daalen, S. K., Korver, C. M., Pyntikova, T., et al. (2003). Polymorphism for a 1.6-Mb deletion of the human Y chromosome persists through balance between recurrent mutation and haploid selection. *Nat Genet*, 35(3), 247-251.
- Repping, S., van Daalen, S. K., Korver, C. M., Brown, L. G., Marszalek, J. D., Gianotten, J., et al. (2004). A family of human Y chromosomes has dispersed throughout northern Eurasia despite a 1.8-Mb deletion in the azoospermia factor c region. *Genomics*, 83(6), 1046-1052.
- Rice, W. R. (1987). Genetic hitchhiking and the evolution of reduced genetic activity of the Y sex chromosome. *Genetics*, 116(1), 161-167.
- Richards, M., Corte-Real, H., Forster, P., Macaulay, V., Wilkinson-Herbots, H., Demaine, A., et al. (1996). Paleolithic and neolithic lineages in the European mitochondrial gene pool. *Am J Hum Genet*, 59(1), 185-203.
- Richards, M., Macaulay, V., Hickey, E., Vega, E., Sykes, B., Guida, V., et al. (2000). Tracing European founder lineages in the Near Eastern mtDNA pool. *Am J Hum Genet*, 67(5), 1251-1276.
- Richards, M., Macaulay, V. A., Bandelt, H. J., & Sykes, B. C. (1998). Phylogeography of mitochondrial DNA in western Europe. *Ann Hum Genet*, 62(3), 241-260.
- Rocha, E. P., Smith, J. M., Hurst, L. D., Holden, M. T., Cooper, J. E., Smith, N. H., et al. (2006). Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J Theor Biol*, 239(2), 226-235.
- Roewer, L., Kayser, M., Dieltjes, P., Nagy, M., Bakker, E., Krawczak, M., et al. (1996). Analysis of molecular variance (AMOVA) of Y-chromosome-specific microsatellites in two closely related human populations. *Hum Mol Genet*, 5(7), 1029-1033.

- Rogers, A. R., & Harpending, H. (1992). Population growth makes waves in the distribution of pairwise genetic differences. *Mol Biol Evol*, 9(3), 552-569.
- Rootsi, S., Zhivotovsky, L. A., Baldovic, M., Kayser, M., Kutuev, I. A., Khusainova, R., et al. (2007). A counter-clockwise northern route of the Y-chromosome haplogroup N from Southeast Asia towards Europe. *Eur J Hum Genet*, 15(2), 204-211.
- Rosser, Z. H., Zerjal, T., Hurles, M. E., Adojaan, M., Alavantic, D., Amorim, A., et al. (2000). Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *Am J Hum Genet*, 67(6), 1526-1543.
- Rozen, S., Marszalek, J. D., Alagappan, R. K., Skaletsky, H., & Page, D. C. (2009). Remarkably little variation in proteins encoded by the Y chromosome's single-copy genes, implying effective purifying selection. *Am J Hum Genet*, 85(6), 923-928.
- Rozen, S., Skaletsky, H., Marszalek, J. D., Minx, P. J., Cordum, H. S., Waterston, R. H., et al. (2003). Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature*, 423(6942), 873-876.
- Rubicz, R., Schurr, T. G., Babb, P. L., & Crawford, M. H. (2003). Mitochondrial DNA variation and the origins of the Aleuts. *Hum Biol*, 75(6), 809-835.
- Rubinstein, S., Dulik, M. C., Gokcumen, O., Zhadanov, S., Osipova, L., Cocca, M., et al. (2008). Russian Old Believers: Genetic Consequences of Their Persecution and Exile, as Shown by Mitochondrial DNA Evidence. *Human Biology*, 80(3), 203-237.
- Rudenko, S. I. (1970). *Frozen tombs of Siberia, the Pazyryk burials of Iron Age horsemen*. Berkeley: University of California Press.
- Ruiz-Pesini, E., Mishmar, D., Brandon, M., Procaccio, V., & Wallace, D. C. (2004). Effects of purifying and adaptive selection on regional variation in human mtDNA. *Science*, 303(5655), 223-226.
- Ruiz Linares, A., Nayar, K., Goldstein, D. B., Hebert, J. M., Seielstad, M. T., Underhill, P. A., et al. (1996). Geographic clustering of human Y-chromosome haplotypes. *Ann Hum Genet*, 60(Pt 5), 401-408.
- Ruvolo, M., Disotell, T. R., Allard, M. W., Brown, W. M., & Honeycutt, R. L. (1991). Resolution of the African hominoid trichotomy by use of a mitochondrial gene sequence. *Proc Natl Acad Sci USA*, 88(4), 1570-1574.
- Ruvolo, M., Pan, D., Zehr, S., Goldberg, T., Disotell, T. R., & von Dornum, M. (1994). Gene trees and hominoid phylogeny. *Proc Natl Acad Sci U S A*, 91(19), 8900-8904.

- Rychkov Iu, G., & Udina, I. G. (1985). [Population genetics of taiga hunters-deer breeders. Characteristics of the prevalence of HLA system markers among the native population of Central Siberia]. *Genetika*, 21(5), 861-867.
- Rychkov, S., Naumova, O., Falunin, D. A., Zhukova, O. V., & Rychkov Iu, G. (1995). [Gene pool of residents of northeastern Eurasia in light of data on polymorphism of mitochondrial DNA. I. New data on polymorphism of restriction sites of the D-loop of mtDNA in aboriginal populations of the Caucasus and Siberia]. *Genetika*, 31(1), 118-127.
- Sagan, L. (1967). On the origin of mitosing cells. *J Theor Biol*, 14(3), 255-274.
- Saillard, J., Forster, P., Lynnerup, N., Bandelt, H. J., & Norby, S. (2000). mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. *Am J Hum Genet*, 67(3), 718-726.
- Salas, A., Carracedo, A., Richards, M., & Macaulay, V. (2005). Charting the ancestry of African Americans. *Am J Hum Genet*, 77(4), 676-680.
- Salas, A., Richards, M., De la Fe, T., Lareu, M. V., Sobrino, B., Sanchez-Diz, P., et al. (2002). The making of the African mtDNA landscape. *Am J Hum Genet*, 71(5), 1082-1111.
- Salas, A., Richards, M., Lareu, M. V., Scozzari, R., Coppa, A., Torroni, A., et al. (2004). The African diaspora: mitochondrial DNA and the Atlantic slave trade. *Am J Hum Genet*, 74(3), 454-465.
- Salas, A., Richards, M., Lareu, M. V., Sobrino, B., Silva, S., Matamoros, M., et al. (2005). Shipwrecks and founder effects: divergent demographic histories reflected in Caribbean mtDNA. *Am J Phys Anthropol*, 128(4), 855-860.
- Santos, F. R., Pandya, A., Tyler-Smith, C., Pena, S. D., Schanfield, M., Leonard, W. R., et al. (1999). The central Siberian origin for native American Y chromosomes. *Am J Hum Genet*, 64(2), 619-628.
- Santos, F. R., Pena, S. D., & Tyler-Smith, C. (1995). PCR haplotypes for the human Y chromosome based on alphoid satellite DNA variants and heteroduplex analysis. *Gene*, 165(2), 191-198.
- Scheinfeldt, L., Friedlaender, F., Friedlaender, J., Latham, K., Koki, G., Karafet, T., et al. (2006). Unexpected NRY chromosome variation in Northern Island Melanesia. *Mol Biol Evol*, 23(8), 1628-1641.
- Schmidt, H. A., Strimmer, K., Vingron, M., & von Haeseler, A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, 18(3), 502-504.

- Schneider, S., & Excoffier, L. (1999). Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA. *Genetics*, 152(3), 1079-1089.
- Schofield, W. N. (1985). Predicting basal metabolic rate, new standards and review of previous work. *Hum Nutr Clin Nutr*, 39 Suppl 1, 5-41.
- Schurr, T. G., Ballinger, S. W., Gan, Y. Y., Hodge, J. A., Merriwether, D. A., Lawrence, D. N., et al. (1990). Amerindian mitochondrial DNAs have rare Asian mutations at high frequencies, suggesting they derived from four primary maternal lineages. *Am J Hum Genet*, 46(3), 613-623.
- Schurr, T. G., & Sherry, S. T. (2004). Mitochondrial DNA and Y chromosome diversity and the peopling of the Americas: evolutionary and demographic evidence. *Am J Hum Biol*, 16(4), 420-439.
- Schurr, T. G., Sukernik, R. I., Starikovskaya, Y. B., & Wallace, D. C. (1999). Mitochondrial DNA variation in Koryaks and Itel'men: population replacement in the Okhotsk Sea-Bering Sea region during the Neolithic. *Am J Phys Anthropol*, 108(1), 1-39.
- Schurr, T. G., & Wallace, D. C. (2003). Genetic prehistory of Paleoasiatic-speaking populations of northeastern Siberia and their relationships to Native Americans. In L. Kendall & I. Krupnik (Eds.), *Constructing cultures then and now: celebrating Franz Boas and the Jesup North Pacific Expedition* (Vol. Contributions to circumpolar anthropology, pp. 239-258). Washington, D.C.: Arctic Studies Center, National Museum of Natural History, Smithsonian Institution.
- Schwartz, M., & Vissing, J. (2002). Paternal inheritance of mitochondrial DNA. *N Engl J Med*, 347(8), 576-580.
- Schwartz, M., & Vissing, J. (2003). New patterns of inheritance in mitochondrial disease. *Biochem Biophys Res Commun*, 310(2), 247-251.
- Semino, O., Passarino, G., Oefner, P. J., Lin, A. A., Arbuzova, S., Beckman, L. E., et al. (2000). The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. *Science*, 290(5494), 1155-1159.
- Sengupta, S., Zhivotovsky, L. A., King, R., Mehdi, S. Q., Edmonds, C. A., Chow, C. E., et al. (2006). Polarity and temporality of high-resolution Y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. *Am J Hum Genet*, 78(2), 202-221.
- Sergeyev, M. A. (1964). The Tofalars. In M. G. Levin & L. P. Potapov (Eds.), *The peoples of Siberia* (pp. 474-484). Chicago: University of Chicago Press.

- Serre, D., & Paabo, S. (2004). Evidence for gradients of human genetic diversity within and among continents. *Genome Res*, *14*(9), 1679-1685.
- Sharma, S., Rai, E., Bhat, A. K., Bhanwer, A. S., & Bamezai, R. N. (2007). A novel subgroup Q5 of human Y-chromosomal haplogroup Q in India. *BMC Evol Biol*, *7*, 232.
- Shen, P., Lavi, T., Kivisild, T., Chou, V., Sengun, D., Gefel, D., et al. (2004). Reconstruction of patrilineages and matrilineages of Samaritans and other Israeli populations from Y-chromosome and mitochondrial DNA sequence variation. *Hum Mutat*, *24*(3), 248-260.
- Shen, P., Wang, F., Underhill, P. A., Franco, C., Yang, W. H., Roxas, A., et al. (2000). Population genetic implications from sequence variation in four Y chromosome genes. *Proc Natl Acad Sci USA*, *97*(13), 7354-7359.
- Sheremet'eva, V. A., & Gorshkov, V. A. (1977). [Koryaks of Kamchatka. The genetic characteristics and formation of a gene pool]. *Genetika*, *13*(6), 1119-1125.
- Sheremet'eva, V. A., & Gorshkov, V. A. (1981). [Koryak of Kamchatka. The genetic differentiation of the population]. *Genetika*, *17*(7), 1309-1312.
- Shi, W., Ayub, Q., Vermeulen, M., Shao, R. G., Zuniga, S., van der Gaag, K., et al. (2010). A worldwide survey of human male demographic history based on Y-SNP and Y-STR data from the HGDP-CEPH populations. *Mol Biol Evol*, *27*(2), 385-393.
- Shields, G. F., Schmiechen, A. M., Frazier, B. L., Redd, A., Voevoda, M. I., Reed, J. K., et al. (1993). mtDNA sequences suggest a recent evolutionary divergence for Beringian and northern North American populations. *Am J Hum Genet*, *53*(3), 549-562.
- Sigurdardottir, 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.
- Sinor, D. (1990). The establishment and dissolution of the Türk empire. In D. Sinor (Ed.), *The Cambridge History of Early Inner Asia* (pp. 285-316). Cambridge: Cambridge University Press.
- Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P. J., Cordum, H. S., Hillier, L., Brown, L. G., et al. (2003). The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature*, *423*(6942), 825-837.
- Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, *139*, 457 - 462.

- 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.
- Sokolova, L. (2007). Okunev cultural tradition in the stratigraphic aspect. *Archaeology, Ethnology and Anthropology of Eurasia*, 30(1), 41-51.
- Solovenchuk, L. L., & Avanesova, G. P. (1977). [Populational-age dynamics of types of haptoglobins among the inhabitants of the northeast USSR]. *Genetika*, 13(9), 1648-1652.
- Solovenchuk, L. L., Deviatkina, S. D., & Avanesova, G. P. (1976). [Serum alkaline phosphatase polymorphism among Chukchis]. *Genetika*, 12(9), 144-149.
- SPSS Inc. (2001). SPSS for Windows Release 11.0.0. Chicago IL: SPSS Inc.
- 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(1), 67-89.
- Starikovskaya, E. B., Sukernik, R. I., Schurr, T. G., Kogelnik, A. M., & Wallace, D. C. (1998). mtDNA diversity in Chukchi and Siberian Eskimos: implications for the genetic history of Ancient Beringia and the peopling of the New World. *Am J Hum Genet*, 63(5), 1473-1491.
- Stepanov, V. A., & Puzyrev, V. P. (2000a). [Analysis of allele frequency of seven microsatellite loci of Y chromosome in three Tuva populations]. *Genetika*, 36(2), 241-248.
- Stepanov, V. A., & Puzyrev, V. P. (2000b). [Haplotypes of two diallelic Y chromosome loci in the indigenous and migrant populations of Siberia]. *Genetika*, 36(1), 87-92.
- Stepanov, V. A., & Puzyrev, V. P. (2000c). [Microsatellite haplotypes of the Y-chromosome demonstrate the absence of subdivisions and presence of several components in the Tuvian male gene pool]. *Genetika*, 36(3), 377-384.
- Stewart, J. B., Freyer, C., Elson, J. L., Wredenberg, A., Cansu, Z., Trifunovic, A., et al. (2008). Strong purifying selection in transmission of mammalian mitochondrial DNA. *PLoS Biol*, 6(1), e10.
- Stoneking, M. (2000). Hypervariable sites in the mtDNA control region are mutational hotspots. *Am J Hum Genet*, 67(4), 1029-1032.
- Stoneking, M., Sherry, S. T., Redd, A. J., & Vigilant, L. (1992). New approaches to dating suggest a recent age for the human mtDNA ancestor. *Philos Trans R Soc Lond B Biol Sci*, 337(1280), 167-175.

- Sukernik, R. I., Abanina, T. A., Karafet, T. M., Osipova, L. P., & Galaktionov, O. K. (1979). [Population structure of forest Nenets. I. Blood group distribution in six subpopulations]. *Genetika*, 15(2), 327-332.
- Sukernik, R. I., Gol'tsova, T. V., Karafet, T. M., Osipova, L. P., & Galaktionov, O. K. (1977). [Genetic structure of an isolated group of the indigenous population of northern Siberia--Nnganasans (Tavgiitsi) of Taymyr. I. History, erythrocyte and serum blood systems, isoenzymes]. *Genetika*, 13(9), 1653-1661.
- Sukernik, R. I., Karafet, T. M., Abanina, T. A., Korostyshevskii, M. A., & Bashlai, A. G. (1977). [Genetic structure of 2 isolated populations of native inhabitants of Siberia (Northern Altaics) according to the results of a study of blood groups and isoenzymes]. *Genetika*, 13(5), 911-918.
- Sukernik, R. I., Karafet, T. M., & Osipova, L. P. (1977). [Genetic structure of an isolated native population group of northern Siberia, the Nnganasani (Tavgi) of the Taimyr. II. An analysis of intrapopulation variability]. *Genetika*, 13(10), 1855-1864.
- Sukernik, R. I., Karafet, T. M., Osipova, L. P., & Posukh, O. L. (1985). [Population structure of the forest Nentsi. V. F-statistics, genetic distances and the average heterozygosity (compared to the Nnganasani)]. *Genetika*, 21(4), 646-657.
- Sukernik, R. I., Lemza, S. V., Karaphet, T. M., & Osipova, L. P. (1981). Reindeer Chukchi and Siberian Eskimos: studies on blood groups, serum proteins, and red cell enzymes with regard to genetic heterogeneity. *Am J Phys Anthropol*, 55(1), 121-128.
- Sukernik, R. I., & Osipova, L. P. (1976). [Distribution of hereditary variants of haptoglobin and transferrin in several human populations in Siberia]. *Genetika*, 12(9), 139-143.
- Sukernik, R. I., & Osipova, L. P. (1982). Gm and Km immunoglobulin allotypes in Reindeer Chukchi and Siberian Eskimos. *Hum Genet*, 61(2), 148-153.
- Sukernik, R. I., Osipova, L. P., Karafet, T. M., Vibe, V. P., & Kirpichnikov, G. A. (1986). [Genetic and ecological study of aboriginal populations of northeastern Siberia. I. Gm-haplotypes and their frequency in 10 chukchi populations. Genetic structure of reindeer chukchi]. *Genetika*, 22(9), 2361-2368.
- Sukernik, R. I., Schurr, T. G., Starikovskaia, E. B., & Uolles, D. K. (1996). [Mitochondrial DNA variation in native inhabitants of Siberia with reconstructions of the evolutionary history of the American Indians. Restriction polymorphism]. *Genetika*, 32(3), 432-439.
- Sukernik, R. I., Vibe, V. P., Karafet, T. M., Osipova, L. P., & Posukh, O. L. (1986). [Genetic and ecological study of aboriginal populations of northeastern Siberia. II. Polymorphic blood systems, immunoglobulin allotypes and other genetic markers



- in asian eskimos. Genetic structure of Bering sea eskimos]. *Genetika*, 22(9), 2369-2380.
- 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.
- Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105(2), 437-460.
- Tajima, F. (1989a). The effect of change in population size on DNA polymorphism. *Genetics*, 123(3), 597-601.
- Tajima, F. (1989b). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3), 585-595.
- Tajima, F. (1996). The amount of DNA polymorphism maintained in a finite population when the neutral mutation rate varies among sites. *Genetics*, 143(3), 1457-1465.
- Tambets, K., Rootsi, S., Kivisild, T., Help, H., Serk, P., Loogvali, E. L., et al. (2004). The western and eastern roots of the Saami--the story of genetic "outliers" told by mitochondrial DNA and Y chromosomes. *Am J Hum Genet*, 74(4), 661-682.
- 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.
- Tamura, K., Dudley, J., Nei, M., & Kumar, S. (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol*, 24(8), 1596-1599.
- Tamura, K., & Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol*, 10(3), 512-526.
- 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.
- Templeton, A. R. (2006). *Population genetics and microevolutionary theory*. Hoboken, N.J.: Wiley-Liss.
- Thomas, D. H. (2000). *Skull wars: Kennewick man, archaeology, and the battle for Native American identity*. New York: Basic Books.
- Thomson, R., Pritchard, J. K., Shen, P., Oefner, P. J., & Feldman, M. W. (2000). Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. *Proc Natl Acad Sci USA*, 97(13), 7360-7365.

- Tkacheva, N. A., & Tkachev, A. A. (2008). The role of migration in the evolution of the Andronov community. *Archaeology, Ethnology and Anthropology of Eurasia*, 35(3), 88-96.
- Tokarev, S. A., & Gurvich, I. S. (1964). The Yakuts. In M. G. Levin & L. P. Potapov (Eds.), *The peoples of Siberia* (pp. 243-304). Chicago: University of Chicago Press.
- Torroni, A., Bandelt, H. J., D'Urbano, L., Lahermo, P., Moral, P., Sellitto, D., et al. (1998). mtDNA analysis reveals a major late Paleolithic population expansion from southwestern to northeastern Europe. *Am J Hum Genet*, 62(5), 1137-1152.
- Torroni, A., Bandelt, H. J., Macaulay, V., Richards, M., Cruciani, F., Rengo, C., et al. (2001). A signal, from human mtDNA, of postglacial recolonization in Europe. *Am J Hum Genet*, 69(4), 844-852.
- Torroni, A., Huoponen, K., Francalacci, P., Petrozzi, M., Morelli, L., Scozzari, R., et al. (1996). Classification of European mtDNAs from an analysis of three European populations. *Genetics*, 144(4), 1835-1850.
- Torroni, A., Lott, M. T., Cabell, M. F., Chen, Y. S., Lavergne, L., & Wallace, D. C. (1994). mtDNA and the origin of Caucasians: identification of ancient Caucasian-specific haplogroups, one of which is prone to a recurrent somatic duplication in the D-loop region. *Am J Hum Genet*, 55(4), 760-776.
- Torroni, A., Neel, J. V., Barrantes, R., Schurr, T. G., & Wallace, D. C. (1994). Mitochondrial DNA "clock" for the Amerinds and its implications for timing their entry into North America. *Proc Natl Acad Sci U S A*, 91(3), 1158-1162.
- Torroni, A., Rengo, C., Guida, V., Cruciani, F., Sellitto, D., Coppa, A., et al. (2001). Do the four clades of the mtDNA haplogroup L2 evolve at different rates? *Am J Hum Genet*, 69(6), 1348-1356.
- Torroni, A., Richards, M., Macaulay, V., Forster, P., Villems, R., Norby, S., et al. (2000). mtDNA haplogroups and frequency patterns in Europe. *Am J Hum Genet*, 66(3), 1173-1177.
- 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.
- Torroni, A., Schurr, T. G., Yang, C. C., Szathmary, E. J., Williams, R. C., Schanfield, M. S., et al. (1992). Native American mitochondrial DNA analysis indicates that the Amerind and the Nadene populations were founded by two independent migrations. *Genetics*, 130(1), 153-162.

- Torrioni, A., Sukernik, R. I., Schurr, T. G., Starikorskaya, Y. B., Cabell, M. F., Crawford, M. H., et al. (1993). mtDNA variation of aboriginal Siberians reveals distinct genetic affinities with Native Americans. *Am J Hum Genet*, 53(3), 591-608.
- Tuppen, H. A., Blakely, E. L., Turnbull, D. M., & Taylor, R. W. (2010). Mitochondrial DNA mutations and human disease. *Biochim Biophys Acta*, 1797(2), 113-128.
- Turner, C. (1990). Paleolithic teeth of the central Siberian Altai Mountains. In A. P. Derevianko (Ed.), *Chronostratigraphy of the Paleolithic in North, Central, East Asia and America* (pp. 239-243). Novosibirsk: USSR Academy of Sciences.
- Tyler-Smith, C. (2008). An evolutionary perspective on Y-chromosomal variation and male infertility. *Int J Androl*, 31(4), 376-382.
- Tyler-Smith, C., & McVean, G. (2003). The comings and goings of a Y polymorphism. *Nat Genet*, 35(3), 201-202.
- Underhill, P. A., Jin, L., Lin, A. A., Mehdi, S. Q., Jenkins, T., Vollrath, D., et al. (1997). Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. *Genome Res*, 7(10), 996-1005.
- Underhill, P. A., Jin, L., Zemans, R., Oefner, P. J., & Cavalli-Sforza, L. L. (1996). A pre-Columbian Y chromosome-specific transition and its implications for human evolutionary history. *Proc Natl Acad Sci USA*, 93(1), 196-200.
- Underhill, P. A., & Kivisild, T. (2007). Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annu Rev Genet*, 41, 539-564.
- Underhill, P. A., Myres, N. M., Rootsi, S., Metspalu, M., Zhivotovsky, L. A., King, R. J., et al. (2009). Separating the post-Glacial coancestry of European and Asian Y chromosomes within haplogroup R1a. *Eur J Hum Genet*, 18(4), 479-484.
- Underhill, P. A., Passarino, G., Lin, A. A., Shen, P., Mirazon Lahr, M., Foley, R. A., et al. (2001). The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann Hum Genet*, 65(1), 43-62.
- Underhill, P. A., Shen, P., Lin, A. A., Jin, L., Passarino, G., Yang, W. H., et al. (2000). Y chromosome sequence variation and the history of human populations. *Nat Genet*, 26(3), 358-361.
- Vajda, E. J. (2001). *Yeniseian peoples and languages : a history of Yeniseian studies : with an annotated bibliography and a source guide*. Richmond, Surrey: Curzon Press.
- van Oven, M., & Kayser, M. (2009). Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum Mutat*, 30(2), E386-394.

- Vasil'ev, S. A. (1993). The Upper Paleolithic of Northern Asia. *Current Anthropology*, 34(1), 82-92.
- Vasilev, S. A., Kuzmin, Y. V., Orlova, L. A., & Dementiev, V. N. (2002). Radiocarbon-based chronology of the Paleolithic in Siberia and its relevance to the peopling of the New World. *Radiocarbon*, 44(2), 503-550.
- 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.
- Voevoda, M. I., Sitnikova, V. V., Chikisheva, T. A., Romashchenko, A. G., Polos'mak, N. V., Molodin, V. I., et al. (1998). [Molecular genetic analysis of mitochondrial DNA of representatives of from the Pazyryk culture of Altai (IV-II centuries B.C.)]. *Dokl Akad Nauk*, 358(4), 564-566.
- Vogt, P. H. (2005). AZF deletions and Y chromosomal haplogroups: history and update based on sequence. *Hum Reprod Update*, 11(4), 319-336.
- Volod'ko, N. V., Eltsov, N. P., Starikovskaya, Y. B., & Sukernik, R. I. (2009). [Analysis of the mitochondrial DNA diversity in Yukaghirs in the evolutionary context]. *Genetika*, 45(7), 992-996.
- 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 Pleistocene peopling of the Americas. *Am J Hum Genet*, 82(5), 1084-1100.
- Wakeley, J. (1993). Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. *J Mol Evol*, 37(6), 613-623.
- Wakeley, J. (2009). *Coalescent theory: an introduction*. Greenwood Village, Colo.: Roberts & Co. Publishers.
- Wallace, D. C. (1999). Mitochondrial diseases in man and mouse. *Science*, 283(5407), 1482-1488.
- Wallace, D. C. (2007). Why do we still have a maternally inherited mitochondrial DNA? Insights from evolutionary medicine. *Annu Rev Biochem*, 76, 781-821.
- Wallace, D. C., Garrison, K., & Knowler, W. C. (1985). Dramatic founder effects in Amerindian mitochondrial DNAs. *Am J Phys Anthropol*, 68(2), 149-155.
- Ward, R. H., Frazier, B. L., Dew-Jager, K., & Paabo, S. (1991). Extensive mitochondrial diversity within a single Amerindian tribe. *Proc Natl Acad Sci U S A*, 88(19), 8720-8724.

- Wares, J. P., Barber, P. H., Ross-Ibarra, J., Sotka, E. E., & Toonen, R. J. (2006). Mitochondrial DNA and population size. *Science*, 314(5804), 1388-1390; author reply 1388-1390.
- Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theor Popul Biol*, 7(2), 256-276.
- Weber, A. W., Katzenberg, M. A., & Schurr, T. G. (2010). *Prehistoric hunter-gatherers of the Baikal region, Siberia: bioarchaeological studies of past life ways*. Philadelphia: University of Pennsylvania Press.
- Weber, A. W., Link, D. W., & Katzenberg, M. A. (2002). Hunter-Gatherer Culture Change and Continuity in the Middle Holocene of the Cis-Baikal, Siberia. *Journal of Anthropological Archaeology*, 21(2), 230-299.
- Weber, A. W., McKenzie, H. G., & Beukens, R. (2010). Radiocarbon dating of Middle Holocene culture history in Cis-Baikal. In A. W. Weber, M. A. Katzenberg & T. G. Schurr (Eds.), *Prehistoric Hunters-Gathers of the Baikal Region, Siberia* (pp. 27-49). Philadelphia: University of Pennsylvania Press.
- Wells, R. S., Yuldasheva, N., Ruzibakiev, R., Underhill, P. A., Evseeva, I., Blue-Smith, J., et al. (2001). The Eurasian heartland: a continental perspective on Y-chromosome diversity. *Proc Natl Acad Sci U S A*, 98(18), 10244-10249.
- Whitfield, L. S., Sulston, J. E., & Goodfellow, P. N. (1995). Sequence variation of the human Y chromosome. *Nature*, 378(6555), 379-380.
- Willuweit, S., & Roewer, L. (2007). Y chromosome haplotype reference database (YHRD): update. *Forensic Sci Int Genet*, 1(2), 83-87.
- Wilson, I., Balding, D., & Weale, M. (2003). Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *Journal of the Royal Statistical Society: Series A*, 166, 155-188.
- Wixman, R. (1984). *The peoples of the USSR: an ethnographic handbook*. Armonk, N.Y.: M.E. Sharpe.
- Wright, S. (1930). The genetical theory of natural selection: a review. *J Hered*, 21, 349-356.
- Xue, Y., Wang, Q., Long, Q., Ng, B. L., Swerdlow, H., Burton, J., et al. (2009). Human Y Chromosome Base-Substitution Mutation Rate Measured by Direct Sequencing in a Deep-Rooting Pedigree. *Current Biology*, 19(17), 1453-1457.
- Xue, Y., Zerjal, T., Bao, W., Zhu, S., Lim, S. K., Shu, Q., et al. (2005). Recent spread of a Y-chromosomal lineage in northern China and Mongolia. *Am J Hum Genet*, 77(6), 1112-1116.

- Xue, Y., Zerjal, T., Bao, W., Zhu, S., Shu, Q., Xu, J., et al. (2006). Male demography in East Asia: a north-south contrast in human population expansion times. *Genetics*, *172*(4), 2431-2439.
- Xue, Y., Zerjal, T., Bao, W., Zhu, S., Shu, Q., Xu, J., et al. (2008). Modelling male prehistory in east Asia using BATWING. In S. Matsumura, P. Forster & C. Renfrew (Eds.), *Simulations, genetics and human prehistory* (pp. 79-88). Cambridge: McDonald Institute for Archaeological Research.
- Y Chromosome Consortium. (2002). A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Res*, *12*(2), 339-348.
- Yablonsky, L. T. (1995). The material culture of the Saka and historical reconstruction. In J. Davis-Kimball, V. A. Bashilov & L. T. Yablonsky (Eds.), *Nomads of the Eurasian steppes in the Early Iron Age* (pp. 201-239). Berkeley, CA: Zinat Press.
- Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. *Tree*, *11*(9), 367-372.
- Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*, *13*(5), 555-556.
- Yang, Z., & Nielsen, R. (2008). Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol Biol Evol*, *25*(3), 568-579.
- Yao, Y. G., Kong, Q. P., Bandelt, H. J., Kivisild, T., & Zhang, Y. P. (2002). Phylogeographic differentiation of mitochondrial DNA in Han Chinese. *Am J Hum Genet*, *70*(3), 635-651.
- Yao, Y. G., Kong, Q. P., Wang, C. Y., Zhu, C. L., & Zhang, Y. P. (2004). Different matrilineal contributions to genetic structure of ethnic groups in the silk road region in china. *Mol Biol Evol*, *21*(12), 2265-2280.
- Yao, Y. G., & Zhang, Y. P. (2002). Phylogeographic analysis of mtDNA variation in four ethnic populations from Yunnan Province: new data and a reappraisal. *J Hum Genet*, *47*(6), 311-318.
- Yoder, A. D., Cartmill, M., Ruvolo, M., Smith, K., & Vilgalys, R. (1996). Ancient single origin for Malagasy primates. *Proc Natl Acad Sci U S A*, *93*(10), 5122-5126.
- Yoder, A. D., & Yang, Z. (2000). Estimation of primate speciation dates using local molecular clocks. *Mol Biol Evol*, *17*(7), 1081-1090.
- Zegura, S. L., Karafet, T. M., Zhivotovsky, L. A., & Hammer, M. F. (2004). High-resolution SNPs and microsatellite haplotypes point to a single, recent entry of

- Native American Y chromosomes into the Americas. *Mol Biol Evol*, 21(1), 164-175.
- Zeng, L. W., Comeron, J. M., Chen, B., & Kreitman, M. (1998). The molecular clock revisited: the rate of synonymous vs. replacement change in *Drosophila*. *Genetica*, 102-103(1-6), 369-382.
- Zerjal, T., Dashnyam, B., Pandya, A., Kayser, M., Roewer, L., Santos, F. R., et al. (1997). Genetic relationships of Asians and Northern Europeans, revealed by Y-chromosomal DNA analysis. *Am J Hum Genet*, 60(5), 1174-1183.
- Zerjal, T., Pandya, A., Thangaraj, K., Ling, E. Y., Kearley, J., Bertoneri, S., et al. (2007). Y-chromosomal insights into the genetic impact of the caste system in India. *Hum Genet*, 121(1), 137-144.
- Zerjal, T., Wells, R. S., Yuldasheva, N., Ruzibakiev, R., & Tyler-Smith, C. (2002). A genetic landscape reshaped by recent events: Y-chromosomal insights into central Asia. *Am J Hum Genet*, 71(3), 466-482.
- Zerjal, T., Xue, Y., Bertorelle, G., Wells, R. S., Bao, W., Zhu, S., et al. (2003). The genetic legacy of the Mongols. *Am J Hum Genet*, 72(3), 717-721.
- Zhang, F., Xu, Z., Tan, J., Sun, Y., Xu, B., Li, S., et al. (2010). Prehistorical East-West admixture of maternal lineages in a 2,500-year-old population in Xinjiang. *Am J Phys Anthropol*, 142(2), 314-320.
- Zhivotovsky, L. A., Underhill, P. A., Cinnioglu, C., Kayser, M., Morar, B., Kivisild, T., et al. (2004). The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. *Am J Hum Genet*, 74(1), 50-61.
- Zhong, H., Shi, H., Qi, X.-B., Xiao, C.-J., Jin, L., Ma, R. Z., et al. (2010). Global distribution of Y-chromosome haplogroup C reveals the prehistoric migration routes of African exodus and early settlement in East Asia. *J Hum Genet*, 55(7), 428-435.
- Zlojutro, M., Tarskaia, L. A., Sorensen, M., Snodgrass, J. J., Leonard, W. R., & Crawford, M. H. (2009). Coalescent simulations of Yakut mtDNA variation suggest small founding population. *Am J Phys Anthropol*, 139(4), 474-482.
- Zouros, E. (1979). Mutation rates, population sizes and amounts of electrophoretic variation of enzyme loci in natural populations. *Genetics*, 92(2), 623-646.
- Zuckermandl, E., & Pauling, L. (1961). Molecular disease, evolution, and genetic heterogeneity. In M. Kasha & B. Pullman (Eds.), *Horizons in biochemistry* (pp. 189). New York: Academic Press.

Zuckerkandl, E., & Pauling, L. (1965). Evolutionary divergence and convergence in proteins. In V. Bryson & H. J. Vogel (Eds.), *Evolving genes and proteins*. (pp. 97-165). New York: Academic Press.