A novel approach for elucidating the complex maternal prehistories of Siberian ethnolinguistic groups using complete mitochondrial genomes

Fakultät für Biowissenschaften, Pharmazie und Psychologie

der Universität Leipzig

genehmigte

DISSERTATION

zur Erlangung des akademischen Grades

Doktor rerum naturalium

Dr. rer. nat.

vorgelegt

von

Christopher Mark Whitten, MSc

geboren am 21.09.1974 in Memphis, TN, USA.

Dekan: Prof. Dr. Tilo Pompe

Gutachter: Prof. Dr. Svante Pääbo

Asst Prof. Dr. Graciela S. Cabana

Tag der Verteidigung: 18.11.2016

Publications:

Duggan AT, **Whitten M**, Wiebe V, Crawford M, Butthof A, Spitsyn V, Makarov S, Novgorodov I, Osakovsky V, Pakendorf B. (2013) Investigating the Prehistory of Tungusic Peoples of Siberia and the Amur-Ussuri Region with Complete mtDNA Genome Sequences and Y-chromosomal Markers. *PLoS ONE* 8(12).

Bower, M.A., **Whitten, M.**, Nisbet, R.E.R., Spencer, M., Dominy, K.M., Murphy, A.M., Cassidy, R., Barrett, E., Hill, E.W. and Binns, M. (2013), Thoroughbred racehorse mitochondrial DNA demonstrates closer than expected links between maternal genetic history and pedigree records. *J Animal Breeding and Genetics*, 130.

Bower, M. A., Campana, M. G., Nisbet, R. E. R., Weller, R., **Whitten, M.**, Edwards, C. J., Stock, F., Barrett, E., O'Connell, T. C., Hill, E. W., Wilson, A. M., Howe, C. J., Barker, G. and Binns, M. (2012), Truth in the Bones: Resolving the identity of the founding elite thoroughbred racehorses. *Archaeometry*, 54.

Barbieri C, **Whitten M**, Beyer K, Schreiber H, Li M, and Pakendorf B. (2011) Contrasting maternal and paternal histories in the linguistic context of Burkina Faso. *Mol. Biol.and Evol*. 29.

Campana, MG, DL Lister, **CM Whitten**, CJ Edwards, F Stock, G Barker, MA Bower (2011) Complex relationships between mitochondrial and nuclear DNA preservation in historic DNA extracts. *Archaeometry,* 53.

Maricic T, **Whitten M**, Pääbo S (2010) Multiplexed DNA Sequence Capture of Mitochondrial Genomes Using PCR Products. *PLoS ONE* 5(11).

Cesare de Filippo, Chiara Barbieri, **Mark Whitten**, Sununguko Wata Mpoloka, Ellen Drofn Gunnarsdóttir, Koen Bostoen, Terry Nyambe, Klaus Beyer, Henning Schreiber, Peter de Knijff, Donata Luiselli, Mark Stoneking, and Brigitte Pakendorf (2010) Y-chromosomal variation in Sub-Saharan Africa: insights into the history of Niger-Congo groups, *Mol. Biol. and Evol*. 28.

Bower, MA, MG Campana, **M Whitten**, CJ Edwards, H Jones, E Barrett, R Cassidy, RER Nisbet, EW Hill, CJ Howe & M Binns (2010) The Cosmopolitan maternal heritage of the Thoroughbred racehorse breed shows a significant contribution from British and Irish native mares. *Biology Letters*.

Campana MG, **Whitten CM**, Edwards CJ, Stock F, Murphy AM, et al. (2010) Accurate Determination of Phenotypic Information from Historic Thoroughbred Horses by Single Base Extension. *PLoS ONE* 5(12).

Table of Contents

Bibliographic Data of the Dissertation

Christopher Mark Whitten

A novel approach for elucidating the complex maternal prehistories of Siberian ethnolinguistic groups using complete mitochondrial genomes

Fakultät für Biowissenschaften, Pharmazie und Psychologie
der Universität Leipzig
Dissertation
126 pages, 7 tables, 27 figures

Siberia is an ideal region for exploring population histories from a molecular anthropological perspective given the diverse human populations, in terms of linguistic affiliation and lifestyle, currently inhabiting this geographically large region. As such, this thesis explores new methodologies for the investigation of the genetic histories of Siberian populations. While previous genetic work in this area of the world was able to provide detailed insights into paternal histories based on Y chromosomal data, it was not as successful on the maternal side. There existed difficulties in exploring the complex maternal demographic histories due to high levels of sequence identity between individuals in different populations when using only a very small region of the mitochondrial DNA (mtDNA), known as the hypervariable region I (HV1). This realization led to the initial focus of this dissertation which was to identify and test improved methods of sequencing entire mtDNA genomes. This was necessary because the mtDNA genomes that were published for human Siberian populations and across the globe prior to the work described here were chosen based on specific sub-sample selection criteria that introduced an ascertainment bias rendering them unusable for population-wide analyses. After testing multiple next generation DNA sequencing methods, I helped develop a sequencing library preparation method based on multiplexing and hybridization enrichment of mtDNAs for sequencing by synthesis that has since become widely used in labs across the globe. Comparing the same samples sequenced by both the traditional and new methods for five ethnolinguistic populations showed that these new methods were robust and could lead to different inferences about population histories while avoiding a sampling bias. Based on the results of this thesis it is now recommended for researchers to sequence complete mtDNA genomes for all relevant samples within a collection. By applying these methods to additional Siberian populations it was possible to better describe maternal population contact and identify demographic changes over time. This additional information allowed for the identification of putative drops in the maternal effective population sizes in the Siberian populations examined here. When examining the potential migrations and population contact between Turkic-speaking Yakuts and the Tungusic-speaking Even and Evenks, there exists a differential sharing of haplotypes suggesting that the Tungusic speaking populations herein were already in the northern region and split prior to the expansion of the Yakuts into their territory. The putative origin of the Yakuts as being around Lake Baikal was given additional support from the analyses included in this study and the origins of the Dolgans were shown to predominately include the admixture of Yakuts and Evenks.

Christopher Mark Whitten

A novel approach for elucidating the complex maternal prehistories of Siberian ethnolinguistic groups using complete mitochondrial genomes

Summary

Fakultät für Biowissenschaften, Pharmazie und Psychologie
der Universität Leipzig
2016

Introduction

Research in the field of molecular anthropology focuses on an exploration of human populations across the globe. As humans exhibit a variety of mating and dispersal patterns that often do not correlate to observed ethnic groupings, population genetic analyses can be used to uncover their histories, such as migrations and population contacts. Historically, uniparental markers have been used as humans often exhibit a sex-bias in dispersal patterns. Mitochondrial DNA (mtDNA) is predominately passed from mother to child and analyses of these sequences can trace maternal histories while analyses of Y chromosomal markers, such as single nucleotide polymorphisms (SNPs) and short tandem repeats (STRs), are used to examine paternal histories. Prior to the advent of modern 'next-generation sequencing' (NGS) technologies, most work on mtDNA focused on a small area of the mtDNA genome known as the hypervariable region I (HV1). This region mutates at a rapid rate relative to the rest of the mtDNA genome and is the most commonly used marker for maternal analyses as its small size is easy to sequence using traditional Sanger methods. In the past decade, new technologies like pyrosequencing and sequencing-by-synthesis have emerged and have slowly been incorporated into the field of molecular anthropology. Methods to enrich samples for mtDNA and tag them so that they can be combined in the sequencing process have allowed for a reduction of costs and time relative to traditional methods. The higher throughput of NGS therefore means that sequencing complete mtDNA genomes is now possible for full sample collections.

Prior to the research presented in this thesis, it was not known whether using complete mtDNA sequences would lead to better population inferences; though this has been a commonly held notion. Since the majority of published complete mtDNA sequences were sub-selected based on specific haplogroups of interest or sequence identity in the HV1, most publically available mtDNA genomes are not usable for exploring maternal population histories

Siberia provides an ideal landscape for exploring the historical movement and contact of human populations. It is a geographically large region with highly diverse human populations in terms of lifestyle and linguistic affiliations. Previous research into Siberian population histories have shown structure on the paternal side, but, using only HV1, were unable to uncover maternal histories due to the sharing of HV1 sequence types across populations. It was therefore hypothesized that full mtDNA sequences would be able to elucidate these populations' maternal histories if indeed they had the power to do so.

Goals of this thesis

The aims of this thesis are threefold. First, it was necessary to identify and test new methods of sequencing complete mtDNA genomes using NGS. Techniques for creating libraries to enrich for mtDNA and tag samples for multiplexing greatly helped to utilize the full benefits of NGS systems. The second aim was to improve our understanding of the potential benefits of moving to full mtDNA genome sequencing by exploring the putative bias in sub-selecting samples for complete mtDNA sequencing that was done in many studies. A comparison of population inferences based on HV1 and complete mtDNA genomes was performed to identify differences. The final goal of this thesis was to apply the new methods to specific populations in Siberia to explore their population contact and histories.

Chapter 1 provides an introduction to the topics presented in the rest of the thesis and Chapter 2 gives an overview of the methods used both for laboratory techniques and analyses. This includes specifics on sequencing library preparation, creating and verifying consensus sequences as well as detailing the specifics of the analytical methods used. In Chapter 3, there is an exploration of ascertainment bias and HV1 sequences are compared against complete mtDNA sequences. The following Chapters 4 and 5 explore Siberian maternal population histories on a large scale and small scale, respectively. Finally, a discussion of the utility of the methods employed is given along with their application to improving our understanding of maternal population histories and contact in Siberia with some concluding remarks.

Results

The approach used for generating the complete mtDNA sequences used in this thesis involved a sequencing library preparation method of multiplexing and hybridization-based enrichment of mtDNAs. An initial subset of 379 samples underwent a comparison between HV1 sequences generated using traditional Sanger sequencing and the HV1 region from the Illumina-generated complete mtDNA sequences. An investigation into the two types of sequencing showed that the majority of the differences were due to an N being called in one of the methods or a C being called in a Sanger sequence that wasn’t seen in the corresponding Illumina sequence. Most of these (82%) were in a region of the HV1 known as the poly-C which is known to be problematic in sequencing and was, thus, excised from all subsequent analyses. Because the Illumina sequences had much higher coverage (50-100X) vs only 2x in Sanger sequences, the Illumina base was chosen when there was a discrepancy. Next, there was an exploration of whether HV1 sequence identity is a predictor of complete mtDNA genome sequence identity. Only 17% of pairs of individuals with the same HV1 are also identical for the entire mtDNA genome, so this method of sub-selecting samples for complete mtDNA sequencing based on HV1 identity introduces a bias to the data. Similarly, sub-selecting samples for complete mtDNA sequencing from specific haplogroups of interest also introduces a bias when performing population based analyses based on analyses shown in Chapter 3.

An exploration of maternal population histories in central and eastern Siberia using 715 complete mtDNA genomes described in Chapter 4 was able to shed new light on population contact and histories. Here the focus was on migrations of Turkic and Tungusic speaking populations into Siberia from further south, putatively around Lake Baikal. Additional populations were included for comparison including Mongolic speakers. Taking advantage of the complete mtDNA genomes allowed for a more full understanding of the histories of these populations through the use of haplotype sharing analyses and Bayesian skyline plots to examine changes in maternal effective population sizes (Nef). There was a drop in Nef around 6,000 years ago visible in all Siberian populations but absent in others (Mongolian and Kyrgyz). A subsequent increase in Nef observed in the Yakuts (the largest Turkic speaking population in Siberia) around 1,000 years ago is in agreement with results from Y chromosomal and archaeological research showing a northern migration and following expansion. This shared Nef drop combined with mtDNA homogeneity suggests a common origin for the Siberian populations under study. In regard to which populations migrated northward first, there is a differential haplotype sharing between the Yakuts and Evens (Tungusic speakers) and Yakuts and Evenks (another Tungusic speaking population) and it can be inferred from this that Tungusic speakers were already in the region and had split prior to the expansion of the Yakuts into their territories.

An exploration of the maternal histories in a smaller region of Siberia, the Taimyr Peninsula, focuses on a population called the Dolgan in Chapter 5. A further 149 complete mtDNA genomes were sequenced and analyzed from Dolgans and Nenets (Samoyedic speakers living in the region) and compared with those generated for the analyses in Chapter 4. The Dolgans are also Turkic speakers but have adopted the lifestyle of Tungusic speakers (reindeer herding and hunting rather than cattle and horse pastoralist Yakuts). Thought to be recently formed through the admixture of Yakuts, Evenks, Samoyedic speakers and possibly Russians, it has not been known to what degree these populations contributed to their formation. Analyses show that a very high percentage of haplotype sharing (compared to all other Siberian populations examined in this study) exists between the Dolgans, the Evenks on the Taimyr, and a nearby population of Yakut-speaking Evenks. This sharing is highest between the most geographically close sub-populations and less with linguistic relatives suggesting high levels of recent admixture. Sharing between Dolgans and multiple geographically farther populations suggests maternal lineages being brought to the region in waves of migrations by both the Yakuts and the Tungusic speakers. Therefore, the emerging picture is a complex one involving both shared ancestry and continuing recent admixture.

The analyses provided in this thesis underscore the need to move from HV1 sequencing to complete mtDNA genome sequencing for investigating maternal population histories. These methods have allowed for greater insights into the prehistories of Siberian populations included here as well as in multiple other populations studied around the globe.

Christopher Mark Whitten

Ein neuer Ansatz zur Aufklärung der komplexen mütterlichen Vorgeschichte von sibirischen, ethnolinguistischen Gruppen anhand vollständiger mitochondrialer Genome
Zusammenfassung

Fakultät für Biowissenschaften, Pharmazie und Psychologie
der Universität Leipzig
2016

Einführung

Die Forschung auf dem Gebiet der molekularen Anthropologie konzentriert sich auf die Untersuchung menschlicher Populationen auf der ganzen Welt. Menschen zeigen eine Vielzahl von Paarungs- und Verbreitungsmuster, die häufig nicht mit beobachteten, ethnischen Gruppierungen korrelieren. Populationsgenetische Analysen können verwendet werden um solche Migrationsbewegungen und Kontakte zwischen Populationen aufzudecken. Historisch wurden uniparentale Marker verwendet, da Menschen bei ihren Verbreitungsmustern eine geschlechtsspezifische Verzerrung (sex-bias) aufweisen. Mitochondriale DNA (mtDNA) wird hauptsächlich von der Mutter auf das Kind übertragen, so dass die Analyse solcher Sequenzen die mütterliche Vergangenheit nachvollziehen kann. Währenddessen lässt die Analyse von Y-Chromosom spezifischen Markern, wie einzelnen informativen nuklearen Polymorphismen (SNPs) und kurzen Tandemrepeats (STRs), Rückschlüsse auf die väterliche Geschichte zu. Vor dem Aufkommen der modernen "Next-Generation-Sequencing" (NGS) -Technologien, konzentrierten sich die meisten Arbeiten zur mtDNA auf einen kleinen Bereich des mtDNA-Genoms, der als die hypervariable Region I (HV1) bekannt ist. Diese Region mutiert im Vergleich zum restlichen mtDNA-Genoms mit einer relativ schnellen Rate und stellt den meistverwendeten Marker für maternale Analysen dar, da sie aufgrund ihrer kleinen Größe leicht mit der traditionellen Sanger-Methode zu sequenzieren ist. In den letzten zehn Jahren entstanden neue Technologien wie Pyrosequenzierung und Sequenzierung-durch-Synthese, welche allmählich in das Gebiet der molekularen Anthropologie aufgenommen wurden. Methoden um mtDNA in verschiedenen Proben anzureichern und zu markieren, so dass sie in einem Sequenzierungsprozess kombiniert werden konnten, reduzierten die Kosten und Zeit gegenüber traditionellen Methoden. Der höhere Durchsatz von NGS bedeutet, dass die Sequenzierung vollständiger mtDNA Genome für eine Vielzahl von Probensammlungen heutzutage möglich ist.

Vor der Forschung und Auswertung, welche in dieser Arbeit präsentiert wird, war es unklar ob die Sequenzierung des vollständigen mtDNA-Genoms bessere Rückschlüsse über Bevölkerungsentwicklungen zulassen würde; obwohl dies allgemein angenommen wurde. Da die Mehrheit der veröffentlichten, vollständigen mtDNA-Sequenzen aufgrund einer Vorauswahl entstanden, die auf bestimmten Haplogruppen oder Sequenzidentitäten in der HV1 basierten, sind die meisten öffentlich zugänglichen mtDNA-Genome nicht für die Erforschung der maternalen Populationsgeschichte verwendbar.

Sibirien stellt eine ideale Möglichkeit dar um die historischen Bewegungen und Berührungen menschlicher Bevölkerungen zu erkunden. Es ist ein große geographisch Region, deren unterschiedlichen Bevölkerungsgruppen in Bezug auf Lebensstil und ihre sprachliche Zugehörigkeit eine hohe Diversität aufweisen. Bisherige Untersuchungen der Geschichte sibirischer Bevölkerungen zeigten eine Struktur auf der väterlichen Seite. Die limitierte Verwendung der HV1 war nicht in der Lage maternale Strukturen zu entschlüsseln, da viele Bevölkerungen HV1 Sequenztypen teilen. Es wurde daher vermutet, dass vollständige mtDNA-Sequenzen in der Lage wären, die maternale Geschichte dieser Populationen zu erhellen, wenn deren Analyse aussagekräftig genug wäre.

Ziele dieser Arbeit

Diese Arbeit verfolgt dreierlei Ziele. Zuerst war es notwendig neue Methoden der Sequenzierung vollständiger mtDNA Genome mittels NGS zu identifizieren und zu testen. Techniken zur Erstellung von Bibliotheken mit angereicherter mtDNA, sowie das markieren verschiedener Proben (multiplexing) halfen enorm um die Vorzüge des NGS-Systems umfänglich auszuschöpfen. Das zweite Ziel war es, unser Verständnis für die potenziellen Vorteile vollständiger mtDNA-Genom-Sequenzierung zu verbessern. Hierzu sollte die vermeintliche Verzerrung, aufgrund von voreingenommener Probenauswahl vieler früherer Studien untersucht werden. Ein Vergleich von Populationsschlussfolgerungen basierend auf HV1 und vollständigen mtDNA-Genomen wurde durchgeführt, um mögliche Unterschiede zu identifizieren. Das endgültige Ziel dieser Arbeit war es, die neuen Methoden auf bestimmte Bevölkerungsgruppen in Sibirien anzuwenden um ihre Kontakte und Geschichte zu erforschen.

Kapitel 1 bietet eine Einführung in die Themen, die im Rest der Arbeit präsentiert werden und Kapitel 2 beschreibt die angewendeten Labor- und Analysemethoden. Dazu gehören Einzelheiten zur Vorbereitung von Sequenzierungsbibliotheken, der Erstellung und Überprüfung von Konsensus-Sequenzen sowie die detaillierte Beschreibung der analytischen Methoden. In Kapitel 3 wird die Erforschung der Verzerrung aufgrund von Probenauswahl (ascertainment bias) behandelt und HV1 Sequenzen mit vollständigen mtDNA Sequenzen verglichen. Die folgenden Kapitel 4 und 5 erforschen die sibirischen maternalen Populationsgeschichten in großem wie im kleinen Maßstab. Abschließend wird eine Diskussion über die Nützlichkeit der angewandten Methoden geführt. Dabei wird darauf eingegangen inwieweit diese unser Verständnis der mütterlichen Populationsgeschichte und Kontakte in Sibirien verbessern können.

Ergebnisse

Der in dieser Arbeit verwendete Ansatz zur Erzeugung vollständiger mtDNA-Sequenzen, umfasste die Herstellung von Sequenzierungs-Bibliotheken anhand von Hybridisierungs-Anreicherungsmethoden und Markierungen (multiplexing). Eine erste Teilmenge von 379 Proben, diente dem Vergleich zwischen HV1 Sequenzen, die mit traditionellen Sanger-Sequenzierung erzeugt wurden und vollständigen mtDNA-Sequenzen die mit der Illumina-Technologie entstanden. Eine Untersuchung der beiden Sequenzierungsmethoden zeigte, dass die Mehrheit der Unterschiede nichtbestimmter Basen (N) in einem der beiden Verfahren betraf oder ein C in der Sanger-Sequenzierungsmethode ermittelt wurde, das bei der Illumina-Methode nicht gesehen wurde. Die meisten von ihnen (82%) entfielen auf einen Bereich der HV1 der als poly-C bekannt, und sich für die Sequenzierung als problematisch darstellt und deshalb in allen nachfolgenden Analysen entfernt wurde. Da die Illumina-Sequenzen eine vielfach höhere Abdeckung einzelner Basen hatte (50-100X) im Vergleich zu 2X der Sanger-Sequenzen, wurde bei Diskrepanzen die Illumina-Base verwendet. Als nächstes wurde untersucht, ob die HV1 Sequenzidentität ein Indikator für die Identität des vollständigen mtDNA-Genoms ist. Nur 17% der Paare von Personen mit derselben HV1 waren ebenfalls für bei der vollständigen mtDNA identisch, so dass durch die Vorauswahl von mtDNA Sequenzierungen aufgrund ihrer HV1-Identitäten eine Verzerrung in den Datensatz eingeführt wird. In ähnlicher Weise führt eine Vorauswahl basierend auf Haplogruppen ebenso zu einer Verzerrung, die sich in der Analyse von Populationen widerspiegelt und in Kapitel 3 beschrieben wird.

Eine Untersuchung der maternalen Populationsgeschichte in Zentral- und Ostsibirien, unter Verwendung von 715 vollständige mtDNA-Genomen in Kapitel 4, war in der Lage ein neues Licht auf Bevölkerungskontakte und ~geschichten zu werfen. Hier lag der Fokus auf Migrationen von türkisch- und tungusisch-sprechenden Populationen nach

Sibieien, aus weiter südlichen Regionen, mutmaßlich um den Baikalsee. Weitere Populationen, wie z.B. mongolisch sprechende Völker, wurden zum Vergleich mit einbezogen. Die Verwendung der gesamten mtDNA hatte den Vorteil ein umfassendes Verständnis der Geschichte dieser Populationen zu erhalten. Hierbei kamen Analysen der Haplotyp-Übereinstimmungen sowie Bayesian-Skyline Plots zum Einsatz, welche die maternale effektive Populationsgröße (N_{eff}) untersuchen. Es gab einen Abfall der N_{eff} vor rund 6.000 Jahren, der in allen sibirischen Populationen sichtbar ist, aber in mongolischen und kirgisischen Populationen fehlt. Eine später folgende Erhöhung der N_{eff}, die bei den Jakuten (der größten Turk-sprechenden Bevölkerung in Sibirien) vor rund 1.000 Jahren beobachtet werden konnte, steht in Übereinstimmung mit den Ergebnissen von Y-Chromosom und archäologischen Forschung, die eine nördliche Migration und anschließende Expansion zeigt. Dieser gemeinsame N_{eff}-Abfall in Kombination mit mtDNA-Homogenität legt einen gemeinsamen Ursprung der sibirischen Populationen nahe. Zur Frage, welche Bevölkerung zuerst nach Norden gewanderte, lässt sich sagen, dass es ein unterschiedliche Übereinstimmungen der Haplotypen zwischen Jakuten und Ewenen (tungusisch-sprechend) und Jakuten und Ewenken (eine andere tungusisch sprachige Bevölkerung) gibt. Man kann ableiten, dass tungusisch-sprachige Populationen in dieser Region bereits präsent und voneinander getrennt bevor sich die Expansion der Jakuten in ihr Territorium vollzog.

Eine Untersuchung der maternalen Geschichte in einer kleineren Region Sibiriens, der Taimyrhalbinsel, konzentriert sich auf eine Population die Dolganen genannt wird findet in Kapitel 5 statt. Weitere 149 vollständige mtDNA Genome von Dolganen und Nenzen (samojedisch-sprechende Population, die in der Region lebt) wurden sequenziert und analysiert und mit den Analysen aus Kapitel 4 verglichen. Obwohl die Dolganen ebenfalls turk-sprachig sind, haben sie sich den Lebensstil der tungusisch-sprechenden Bevölkerung angeeignet (Rentierzucht und Jagd statt pastorale Rinder- und

Pferdehaltung der Jakuten). Bisher nahm man an, dass die Dolganen durch die Vermischung der Jakuten, Ewenken, samojedisch-sprechenden Gruppen und möglicherweise Russen entstanden. Unbekannt war allerdings, inwieweit diese Gruppen zu der Formierung der Dolganen beitrugen. Die Analysen zeigen, dass ein sehr hoher Prozentsatz der Haplotyp-Übereinstimmungen (im Vergleich zu allen anderen sibirischen Populationen, die in dieser Studie untersucht wurden) zwischen den Dolganen und den Ewenken der Taimyr-Region und ortsnahen jakutisch-sprachigen Ewenken existiert. Diese Übereinstimmungen sind zwischen den geografisch naheliegenden Subpopulationen am größten und bei sprachlich Verwandten Gruppen geringer, was auf ein hohes Niveau von Durchmischung in der jüngeren Vergangenheit schließen lässt. Übereinstimmungen zwischen Dolganen und mehreren geographisch weiter entfernten Populationen deutet darauf hin, dass maternale Abstammungslinien durch Migrationswellen der jakutisch- und tungusisch-sprachigen Populationen in diese Region eingebracht wurden. Daher ist das sich darstellende Bild komplex und beinhaltet sowohl eine gemeinsamen Abstammung als auch eine fortdauernde jüngere Durchmischung.

Die Analysen in dieser Arbeit unterstreichen die Notwendigkeit, von HV1 Sequenzierungen zu vollständigen mtDNA Genomen zu wechseln, wenn man die maternalen Populationsvergangenheit untersuchen will. Diese Methoden ermöglichten in dieser Arbeit tiefere Einblicke in die Vorgeschichte der sibirischen Populationen, sowie in anderen Studien die Untersuchungen vieler Populationen auf der ganzen Welt.

Chapter 1

General Introduction

Introduction

The field of molecular anthropology aims to elucidate the prehistories of human populations across the globe using biological markers. Over the past three decades, continually improving technologies have allowed for larger amounts of data to be gathered for increasing numbers of populations. Important questions in the field of anthropology revolve around whether and how genes and languages migrate together. It is possible for either genes or languages to be transmitted from one population to another alone. It is also possible for both to be passed along together. There are a number of different combinations by which genes can be passed between populations that come into contact with one another. Sometimes only men transfer from one group to another, other times only women emigrate between the populations, and sometimes populations can fuse together. Migration between populations can happen in one direction or bi-directionally. In order to differentiate between types of admixture, molecular anthropologists often focus on a specific region of the world and examine the genetic diversity in linguistically divergent populations. Following in this tradition, this study focuses on Northeast Asia due to its linguistic diversity (Turkic, Tungusic, Samoyedic, and Mongolic speakers as well as a linguistic isolate) and divergent subsistence patterns ranging from hunting/gathering to cattle and horse pastoralism and reindeer herding.

Siberia is a geographical region which stretches from the Ural Mountains in the west to the Kamchatka Peninsula in the east and from the Arctic Ocean in the north down to the northern borders of Kazakhstan, Mongolia and China. It constitutes roughly ¾ of Russia, though it has one of the lowest population densities due to its harsh climate and relative lack of infrastructure. This study focuses primarily on the Central portion of Siberia, while also including some relevant populations located in the Northeast and more southern populations from Mongolia and Kazakhstan.

In this monograph, the term ethnolinguistic group will be used to mean a group of people that share a combination of a common language, culture and subsistence pattern. Groups are always self-identified and some are subsequently divided by geographic location for the purposes of analyses. The primary focus here will be on populations that speak Turkic and Tungusic languages and, for comparative purposes, also includes Mongolic speakers, Samoyedic speakers, and speakers of the linguistic isolate, Yukaghir.

Benefitting from the fact that some DNA is inherited paternally and some maternally, it is possible to use parental-specific markers to uncover information about our species' past from multiple perspectives. Given the diverse residence marriage and residence patterns existing in human populations, these uniparental markers can help to disentangle complex male and female population contact situations. Mitochondrial DNA (mtDNA) is passed on from mother to offspring and has been used since the 1980s to study human population history and was used to trace back a common mtDNA ancestor (Cann et al., 1987). Most work up until the past few years has focused primarily on a very short segment of the mtDNA genome, the hypervariable region 1 (HV1), in order to elucidate maternal population histories. As the name implies, this small region mutates at a higher rate than the rest of the mtDNA genome. There has long existed a need to investigate the putative benefits of performing population-wide analyses based on complete mtDNA genomes rather than solely focusing on one rapidly mutating region. However, this has largely remained too expensive and time consuming for most research groups. Second- and third-generation sequencing technologies have however made this possible and the data analyzed in this thesis were generated by a novel, tailored method of preparing sequencing libraries by enriching for mtDNA that was developed in collaboration with colleagues using some of the Siberian samples examined here (Maricic et al., 2010). Although “2nd generation sequencing” methods have been around for over a decade, they are still relatively new to the field of molecular anthropology and remain under-utilized. Here they are used to examine maternal prehistories amongst human populations in Siberia.

Studies of maternal population history in Siberia have previously been limited because they largely focused on mtDNA HV1. A few studies have included complete mtDNA genomes, but only limited population inferences could be made from these data because the samples selected for complete mtDNA genome sequencing were chosen based on haplogroups of interest or sub-selected due to having unique HV1 sequences and were, thus, non-random (Achilli et al., 2005; Derenko et al., 2014; Derenko et al., 2007; Fedorova et al., 2013; Starikovskaya et al., 2005; Volodko et al., 2008) With the advent of multiplexing and hybridization methods for use in 2nd generation sequencing technologies, it is now feasible to completely sequence the mtDNA genomes of all the individuals in a collection, rather than just a small selection of samples, which increases our ability to study population demographic parameters (Barbieri et al., 2013a; Barbieri et al., 2012; Delfin et al., 2014; Duggan et al., 2014; Duggan et al., 2013; Gunnarsdóttir et al., 2011a; Gunnarsdóttir et al., 2011b; Maricic et al., 2010; Meyer and Kircher 2010; Vyas et al., 2016). We recently published the first paper on Siberian maternal population history utilizing these methods on a comparatively unbiased collection of Siberian populations with a focus on Tungusic speaking populations (Duggan et al., 2013) using samples sequenced during the course of the work for this dissertation. In addition to more closely examining population contact, the use of complete mtDNA genomes allows for additional analyses that can elucidate effective population sizes over time to provide a look deeper into the past to identify demographic changes.

The Y chromosome is passed from father to son and can therefore help us to gain a better understanding of paternal population histories. This chromosome has traditionally been analyzed using two different types of markers, single nucleotide polymorphisms (SNPs) and short tandem repeats (STRs). Because of the high rate of mutations of STRs, it is possible to distinguish more recent population events, while SNPs involved in the analyses of Y chromosomal data are in relatively stable (mutationally speaking) positions, so we can use them to look in the paternal past (Jobling and Tyler-Smith 1995). While the focus of this thesis is on using complete mtDNA genomes to disentangle maternal population histories, some limited Y chromosomal data will be discussed for the purpose of comparison.

The specific research goals of this study can be divided into three parts. First, it was necessary to find and test new methods of generating complete mtDNA sequences from human samples and to verify these methods against the standard, accepted Sanger sequencing technology. The second goal was to explore the putative benefits of these novel methods by determining whether the common practice of sub-selecting samples for complete mtDNA sequencing from a collection introduces a bias (which could be avoided with the improved methods) and if inferences of population histories based on analyses of full mtDNAs are different from those made using only HV1 data. The third goal of this research was to use these new tested and verified methods to gain a better understanding of the histories of the Siberian populations included herein. To accomplish these goals, the work performed is laid out in the chapters as described below.

Chapter 2 provides an overview of all the laboratory methods used in this dissertation from DNA extraction through library preparation and sequencing. Additionally, the analysis of DNA sequences is described, including sequence assembly methods and analytical tools used in studying the populations included herein. Information on individual populations and sample numbers used in different parts of this study are provided in the “Materials” sections in the relevant chapters (3-5).

In Chapter 3, we will look at the benefits of using complete mtDNA genomes in a side-by-side comparison with HV1 sequences and determine whether different interpretations of population histories could be inferred. This chapter will also examine the potential bias that could result from using existing published data that has undergone a form of sub-selection of specific samples used in complete mtDNA sequencing; a practice which has been common in many past studies and remains in use today.

Next, Chapter 4 examines the largest ethnolinguistic populations in Siberia as well as some of their linguistic and geographic neighbors using complete mtDNA genomes. The goal of this chapter is to obtain a deeper understanding of maternal histories to help uncover putative past migrations and changes in effective population sizes as well as to look at potential contact and admixture on a large scale.

Chapter 5 will explore a small region in the far north of Siberia, the Taimyr Peninsula, in order to shed more light on the origin of the Dolgans, the northernmost population of Turkic language speakers whose subsistence strategies more closely match those of Tungusic-speakers. This ethnolinguistic group is thought to have been formed relatively recently and their origins have been unclear. Therefore, complete mtDNA genome analyses on the relevant populations in this chapter and their comparison to those from Chapter 4 and presented along with a discussion of some Y preliminary chromosomal analyses.

The final chapter (6) summarizes the findings of the previous chapters and examines the key conclusions. A review of the goals of this study will be provided, including the creation and substantiation of improved sequencing methods and the usefulness of these methods in examining population histories in the Siberian populations analyzed in this study both in a large scale exploration of maternal population demographies and a more fine-scaled examination of a recently formed ethnolinguistic group. Overall conclusions are provided along with a discussion of the effects of the work leading to this dissertation on the field of molecular anthropology in general.

Chapter 2

Sequencing and Analytical Methods

Sequencing and Analytical Methods

Initial explorations into generating complete mtDNA genomes

Prior to deciding on the final methods used to sequence the samples in this study, considerable testing was done using a method of parallel tagged sequencing on the Roche 454 GS and GS FLX analyzers based on work described in Meyer et al., (2008). For complete mtDNA sequencing, this method of library preparation was based on the amplification of two, overlapping, long range PCR products covering the entire mtDNA genome that were then sheared down to small fragments. Sequences generated using these systems were not of adequate coverage across the mtDNA genomes, often showing some regions completely uncovered. Attempts were made to improve this protocol for our purposes by varying the long range PCR products used, switching the method of shearing, and using whole genome amplification using a Repli-G kit (Qiagen). Given that further testing still showed significant drawbacks to multiplexed sequencing on the Roche 454 platform, it was decided not to continue with 454 sequencing and those full methods are given in Gunnarsdóttir et al., (2011a). Rather, I switched to testing Illumina Genome Analyzers as they offered higher coverage without some of the problems inherent in 454 sequencing such as the difficulties with homopolymer regions. Optimization of library preparation methods led to the protocol described below.

Illumina library preparation and sequencing

The methods given in Meyer and Kircher (2010) were used with the following modifications: Genomic DNAs were sheared using the Bioruptor UCD-200 (Diagenode) to approximately 200bps to 800bps. The adapter fill-in step used

Dynabeads® MyOne™ Streptavidin C1 (Invitrogen, cat no. 650.01). Beads were prepared by aliquoting 25µl bead suspension for each sample, washing twice with 2X-BWT buffer (Maricic et al., 2010) and eluting in 25µl 2X-BWT buffer. Adapted sample libraries were added to the bead suspension and incubated at room temperature for 15 minutes. The supernatant was discarded and the beads were washed twice with 100µl 1X-BWT buffer. The master mix described in Meyer and Kircher (2010) was used for the fill-in after removing the buffer, and no SPRI purification was necessary.

Specific indexes were attached to individual samples by performing a PCR amplification using the Phusion Mastermix (NEB, cat. no. F-531S). Samples were pooled in equimolar ratios after indexing. The resulting pooled libraries were hybridized to sheared, long-range mtDNA PCR products according to Maricic et al., (2010). Quantitative PCRs were performed on the sample pools after hybridization with the DyNAmo qPCR kit (NEB, cat. no. F-410). Based on the qPCR amplification plots, the sample pools were amplified using the Phusion Mastermix so they could be accurately quantified on an Agilent DNA-1000 chip (Agilent Technologies, cat. no 5067-1504).

Each of the library pools was sequenced with 36+7 or 76+7 cycles in one lane of an Illumina flow cell (Cluster Generation kit V2 and V4, sequencing chemistry V3 and V3+). The sequencing primer used can be found in Meyer and Kircher (2010). The manufacturer's instructions for Single Read Multiplex sequencing on the Genome Analyzer IIx platform were followed.

Assembly of Complete mtDNA Genomes

The runs were processed using SCS versions 2.4, 2.5, and 2.6/RTA 1.4, 1.5, and 1.6 (Illumina Inc.). Base calling was performed using an alternative base-caller, Ibis (Kircher et al., 2009). The raw sequences from individual samples were separated using the index reads (Meyer and Kircher 2010). Reads were scanned for adapter sequence and trimmed correspondingly. Identical reads were collapsed and assembled with the software MIA (Briggs et al., 2009) by mapping the reads to the revised Cambridge reference sequence (rCRS) (Andrews et al., 1999). Additional information on assembly and quality filtering parameters can be found in Li et al., (2010). Once assembled, the reads were checked using clview, an interactive, graphical file viewer (http://compbio.dfci.harvard.edu/tgi/software/).

Consensus sequence clean-up

Because of the method used to align the reads to the rCRS, the intergenic COII/tRNALYS 9-bp repeat was called incorrectly in some samples. The rCRS has two copies of this repeat and, thus, many samples with only one copy were automatically constructed as having two in the consensus. To verify the number of repeats in this region, the reads were checked using an in-house Perl script[1] to identify the number of copies of the 9-bp repeat by using only those reads that span the region. Additionally, samples with the deletion, with potential heteroplasmy of the deletion, and individuals assigned to haplogroup B that appeared to have two copies of the repeat were amplified using the primers from Wrischnik et al., (1987) and the

[1] Thanks to Mingkun Li for the script to check repeats

protocol specified in Tarskaia et al., (2006) and visualized on a 4% agarose gel[2]. There were discrepancies in the number of copies of the repeat between the initially called consensus sequences and the sequences identified using the Perl script as well as the length of the fragment after amplification and visualization for six individuals. The consensus sequences of those six individuals were changed to the number of repeats resulting from the gel verification and reads spanning the entire region.

Prior to performing a multiple alignment, all positions with 1x coverage were replaced with an N. Average coverage of all samples ranged from around 50X to 100X. Additionally, a cut-off was set such that the majority base was kept if it was present in over 70% of the reads. Up to 0.5% of missing data was allowed in the consensus sequences. Samples with more missing data were excluded. A multiple alignment of the consensus sequences was done using MAFFT v6.708b (Katoh et al., 2002).

Verification of sequences

For 379 of the samples that underwent complete mtDNA sequencing for this thesis, the hypervariable region 1 (HV1) had been previously sequenced to at least 2-fold coverage using Sanger sequencing methods (Pakendorf et al., 2006; Pakendorf et al., 2007). This allowed for a direct comparison to be performed in this region between the consensus sequences generated from each technology. When differences were identified (as was the case in 72 individuals, with one to six

[2] I would like to thank Serena Tucci for performing the laboratory work to check for the 9-bp deletions/repeats.

differences per sample), the trace files and the Illumina Genome Analyzer IIx reads were investigated to determine possible reasons for the mismatches. The vast majority of the differences were in the poly-C regions which were removed from analyses. The specifics of these differences are discussed in Chapter 3.

Haplogroup assignment

After performing a multiple alignment, the resulting consensus sequences were manually aligned to the rCRS by removing insertions such that all samples had the same numbering. They were then processed using a Perl script created in-house that used the full polymorphism data from Phylotree.org Build 12 (Oven and Kayser 2009) as a reference. This script provides information on the closest haplogroup that is available in Phylotree.org by comparing all variable positions in the sequences against those provided in the tree.

mtDNA Data analyses

Summary statistics, pairwise Φst between populations, and analysis of molecular variance (AMOVA) were calculated using Arlequin ver. 3.11 (Excoffier et al., 2005). A Bonferroni correction for multiple tests was applied to the matrix of Φst values. MDS plots were created using Statistica ver. 8 (StatSoft, Inc.). Correspondence analysis plots were generated using an in-house R script[3]. Haplotype sharing plots were also created using an in-house R script (de Filippo et al., 2011). Network v.4.516 was used to generate phylogenetic networks from complete mtDNA genomes after removing indels and all positions containing Ns. Bayesian Skyline

[3] Thanks to Chiara Barbieri for her Correspondence Analysis plot script

Plots (BSPs) of populations using the coding region (positions 577-16023) were generated in BEAST v. 1.6.1 to estimate the female effective population size (N_{ef}) over time (Drummond and Rambaut 2007). A generation time of 28 years was used to convert to N_{ef} from τ (N_{ef}×generation time) (Fenner 2005). MCMC samples were based on runs between 30,000,000 and 50,000,000 generations, depending on population size, such that ESS values were above 200. Samples were taken every 3,000 or 4,000 generations, with the first 3,000,000 or 4,000,000 generations discarded as burn-in. The rate of molecular evolution was set at 1.691×10^{-8} substitutions/site/year (Atkinson et al., 2008). The plots were produced by Tracer v. 1.5 (Rambaut and Drummond 2003).

Conclusions

The methods of DNA sequencing library preparation described above were modified and created for this project and were used on the Siberian samples for this study described in the next chapters. The protocol that came out of this work enhanced the ability for research labs to multiplex samples and enrich for mtDNA thus reducing cost while maintaining high coverage leading to more accurate sequences. The combination of these laboratory methods with the subsequent assembly and consensus calling pipeline set up in the Genetics Department of the Max Planck Institute of Evolutionary Anthropology allowed for me to train multiple researchers from various institutions and led to an improvement in DNA sequencing not only in many geographic regions in the field of molecular anthropology, but also across other disciplines.

Chapter 3

Investigating ascertainment bias in sample selection and comparison of complete mtDNA sequencing against HV1 sequencing

Investigating ascertainment bias in sample selection and comparison of complete mtDNA sequencing against HV1 sequencing

Introduction

The number of complete mtDNA sequences deposited in Genbank has grown rapidly in the past decade. However, the majority of these sequences were chosen based on haplogroups of interest or by only using samples with distinctive HV1 sequences (Achilli et al., 2005; Derenko et al., 2014; Derenko et al., 2007; Fedorova et al., 2013; Starikovskaya et al., 2005; Volodko et al., 2008). In part, this is due to limited access of molecular anthropology research groups to facilities able to perform 2nd generation sequencing methods which could make complete mtDNA genomes more easily obtainable. Also, prior to the methods described in Chapter 2, Meyer and Kircher (2010), and Maricic et al. (2010), there were no cost effective and acceptable ways of even utilizing the advances in sequencing technologies for the purposes of population wide mtDNA genome sequencing. Given the higher costs and time associated with sequencing entire mtDNA genomes using traditional Sanger methods, sub-sampling has often been employed when deciding which samples to perform complete sequencing on. This is largely due to the size difference between the HV1 (~360bps) and the complete mtDNA genome (~16,600bps) and the inability to sequence samples in a highly parallel manner.

Sub-selecting samples for complete mtDNA genome sequencing therefore raises two questions: *(i)* is an ascertainment bias (resulting in a non-random sample) introduced when selecting samples based on unique HV1 haplotypes, and *(ii)* do the results of population genetic analyses differ when using HV1 sequences vs.

complete mtDNA genomes? To investigate these questions I used complete mtDNA sequences from Siberian populations for which HV1 sequences were previously generated (Pakendorf et al., 2006, Pakendorf et al., 2007) and performed side-by-side analyses with both sets of sequences.

Materials

A subset of 379 samples from five Siberian populations was used for the analyses in this chapter (Figure 3.1). The Yakuts, Yakut-speaking Evenks (YSE), and Tuvans speak Turkic languages, and the Evens and Evenks speak Northern Tungusic languages. The Yakut sequences were generated for this project and were first published in Duggan et al., (2013). The Tuvan and YSE sequences have not been published elsewhere. More comprehensive descriptions of these populations are given in Chapter 4.

To verify the accuracy of the sequences, the HV1 regions of the newly generated sequences were compared with previously published HV1 sequences (Pakendorf et al., 2006, Pakendorf et al., 2007) for all samples described above as explained in Chapter 2. As ~20% of tested samples showed some difference between the Sanger and Illumina HV1 sequences, a closer investigation was conducted. Most of the differences between the two technologies were either due to an 'N' being called in one of the sequences (24%) which could be due to heteroplasmy, NUMTs, or sequencing error, or a 'C' was called in a Sanger sequence that was not present in the Illumina sequence (62%). The majority of mismatches occurred in, or adjacent to, the 'poly-C' region in HV1 (82%). The C-stretches in HV1 and HV2 (16184-16193 and 303-315) were therefore excluded from all analyses. Given the high coverage of the Illumina sequences (50-100X) against the 2X coverage of the Sanger sequences, for those few discrepancies still present after removing the poly-C regions, the Illumina base was preferred if it met the cutoff described in Chapter 2 (70% of reads having the majority base), otherwise an N was given.

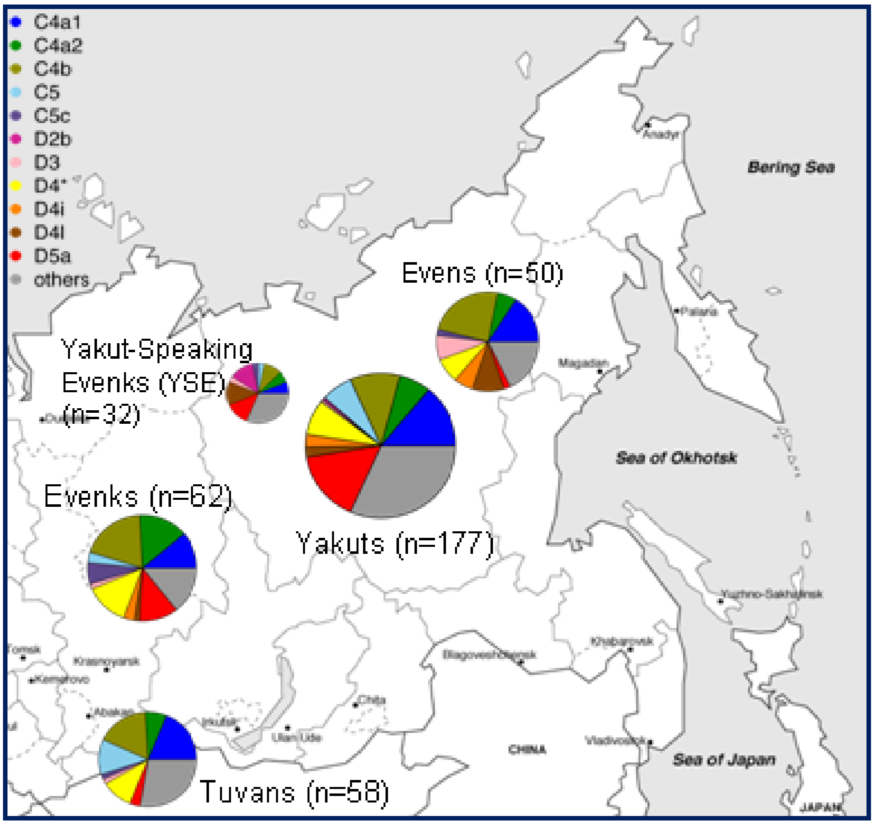

Fig. 3.1. Map of population locations, sizes and haplogroup percentages of samples based on complete mtDNA genomes used in this chapter. Colors indicate haplogroup attribution

Results

Sequence identity

When selecting samples based on unique HV1 haplotypes, there is an underlying assumption: if HV1 sequences are identical, then the complete mtDNA should also be identical or at least very similar. Therefore, HV1 has been historically used as a proxy for the whole mtDNA genome. If the samples used are biasing the currently existing set of sequences, this could have an impact on what population information is inferred. To determine whether an ascertainment bias is introduced when selecting samples for complete mtDNA sequencing based on HV1 identity, pairwise comparisons were performed using an in-house script[4], between the published HV1 sequences and the newly sequenced complete mtDNA genomes. Figure 3.2 compares the HV1 pairwise differences against those outside of HV. As can be seen, the number of base pair differences within HV1 is not a good predictor of the number of differences in the rest of the mtDNA genome. This pattern is not unique to the Siberian populations included in this study, as similar results were subsequently found in Filipino populations (Gunnarsdóttir et al., 2011a).

[4] Thanks to Mingkun Li for writing the script to compare sequences.

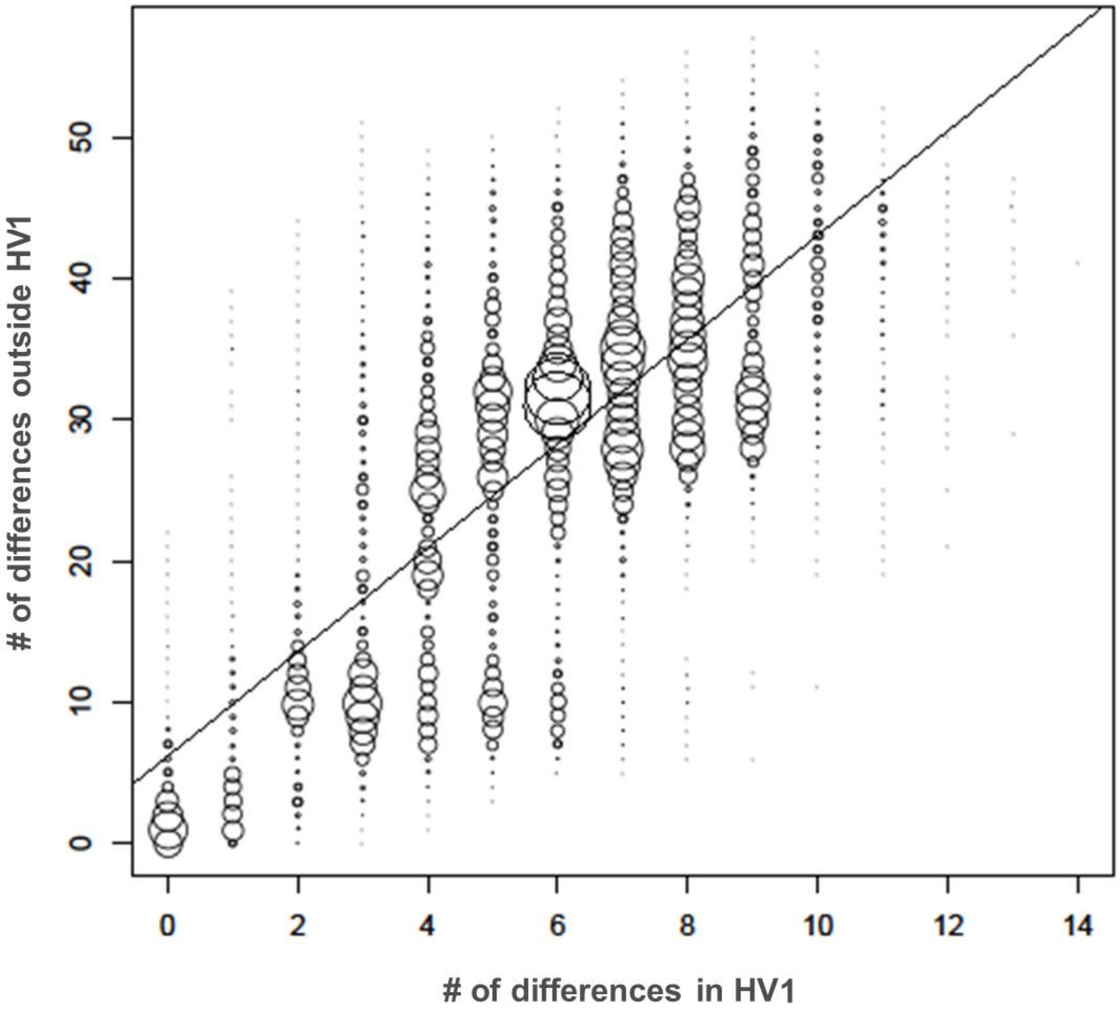

Fig. 3.2. Numbers of base pair differences inside versus outside HV1 for pairs of individuals. Circle size represents number of pairs of individuals represented in comparison and a best fit line is shown.

When looking at all pairs of individuals having identical HV1 sequences (as seen in the far left of Figure 3.2), only 17% of the pairs are also identical for the complete mtDNA genome (Figure 3.3). The majority of differences between complete mtDNA genomes for HV1-identical individuals lies between one and eight, but can go as high as 22.

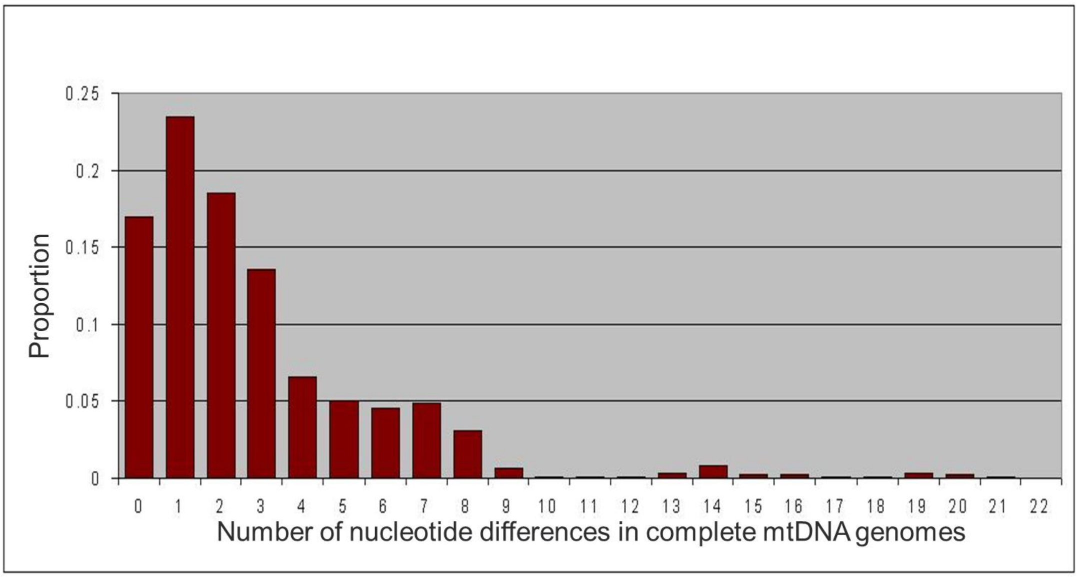

Fig. 3.3. Distribution of nucleotide differences in complete mtDNA genomes for HV1-identical individuals

To explore the potential effects of sub-selection based on HV1 identity further, 100 simulations were carried out using an in-house R script[5] on complete mtDNA genomes for each of the three most frequently found haplogroups (C, D excluding D5a, and D5a) by randomly sampling one sequence from each set of individuals whose HV1 sequences were identical and calculating four common measures of diversity and then comparing the results to the same summary statistics generated using all individuals from these haplogroups (Figure 3.4). These statistics include the number of segregating sites (S), Tajima's D, mean number of pairwise differences, and θs (which is a variation of the Watterson estimator (θ) for exploring population mutation rates using segregating sites and sample size). Statistically significant differences (P-value < 0.05) were seen in all but two out of the 12 comparisons.

[5] The script for these simulations was written by Cesare de Filippo

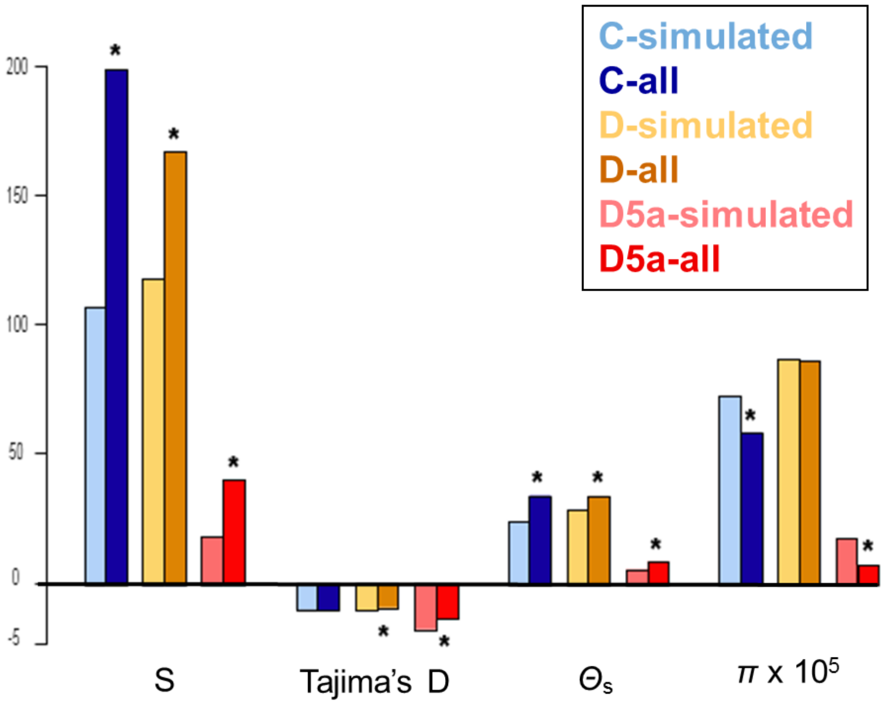

Fig. 3.4. Simulations of sampling based on HV1 identity showing results from four commonly used diversity measures: S = number of segregating sites; π = mean number of pairwise differences. The asterisks denote statistically significant differences.

HV1 versus complete mtDNA

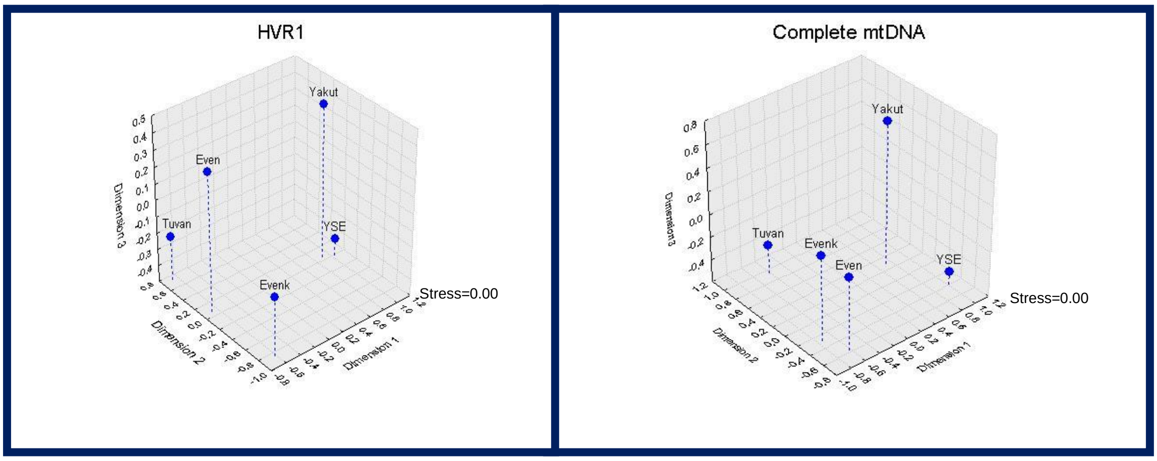

Fig. 3.5. MDS plots based on pairwise Φst values for HV1 and complete mtDNA

In the MDS plot based on HV1 sequences shown in Figure 3.5, which were cut out of the complete mtDNA sequences, the two Northern Tungusic populations (Even and Evenk) are separated by dimensions 2 and 3, while using complete mtDNA genomes shows them to be closely related. The Yakut-speaking Evenks (YSE) claim Evenk ethnicity and have the same mode of subsistence, though they have spoken Yakut for at least three generations. Analyses based on HV1 sequences group them with Yakuts in dimensions 1 and 2 and show them to be separated from the Evenks in all dimensions. However, when complete mtDNA genomes are used for the same analysis, the YSE are more differentiated from the Yakuts in dimensions 2 and 3 and closer to the Evenks in dimension 2.

Discussion

The underlying assumption that identity in the HV1 region is correlated with identity across the genome is not true, at least in these populations. Selecting samples for complete mtDNA sequencing based on HV1 identity does appear to introduce a bias into this data set. Most of the differences in summary statistics between using all sequences in a data set and biased subsamples based on HV1 identity are significant, so sequencing all available samples in a collection is recommended now that rapid, cost-effective methods as described in Chapter 2 are available. Additionally, there is an obvious bias introduced when samples are selected based on haplogroups of interest, so any population inferences based on such statistics from biased sample selection would also be questionable. This was supported by analyses in Gunnarsdóttir et al. (2011a) where a Bayesian skyline plot was constructed based on sub-selecting samples from different haplogroups or sub-haplogroups and compared against an unbiased sample and showed strikingly different population demographic histories. Taken together, these analyses show that biased sub-sets of mtDNA genomes should not be used in these types of analyses of human populations. It is absolutely essential that care is taken to avoid any such sequences when conducting population-wide analyses for if were one to simply download all available mtDNA genomes for a given population, there is a danger of unintentionally biasing the data.

As for the question of using complete mtDNA genomes or HV1 sequences, it is clear based on the populations examined herein that two different pictures of maternal population histories emerge based on whether only the HV1 or the complete mtDNA genome are used in these MDS analyses. Similar work was undertaken to look at the differences in analyzing the HV1 versus the entire control region but excluding the larger coding region (Johnson 2013). This study showed that while some analyses were not significantly different (such as AMOVAs and neutrality tests), others, including MDS plots and neighbor joining trees, showed obvious differences.

While the conclusion of that study was reported as being ambiguous, and it was put forward that issues such as budget must be taken into account when deciding whether to sequence only the HV1 or complete control region, the increased information was an improvement and, combined with the analyses shown here, should ideally be taken to the logical next step of generating the entire mtDNA genomes for full sample collections. Therefore, given the improvements in cost and speed previously discussed, it is no longer recommended to use HV1 as a proxy for the complete mtDNA genome when exploring maternal population histories.

Chapter 4

Maternal Population Histories of Central / Eastern Siberia

Maternal population histories of central / eastern Siberia

Introduction

Archaeological evidence indicates that the region around Lake Baikal in Siberia has been inhabited since the Upper Palaeolithic (Naumov 2010; Weber et al., 2010). However, as shown by studies of burial sites to the west of Lake Baikal (Cis-Baikal), this region was initially inhabited by groups differing both genetically and culturally from the populations who currently reside in this area (Mooder et al., 2006). The existence of divergent cultures between the Early and Late Neolithic separated by a hiatus in burial sites during most of the 7th millennium BP has been uncovered by archaeological evidence (Weber et al., 2010). In an attempt to elucidate the origins of modern Siberians, Mooder et al., (2006) examined the hypervariable region 1 (HV1) of mitochondrial DNA (mtDNA) of both ancient and modern populations and showed affinities of modern Siberians to the Late Neolithic, but not the Early Neolithic, Cis-Baikal populations.

Many of the indigenous peoples currently inhabiting central and north-eastern Siberia are thought to be relatively recent newcomers to these regions. Three widespread groups, the Tungusic-speaking Evens and Evenks and the Turkic-speaking Yakuts, moved into the territories they currently inhabit only within the past millennium. It has been proposed that the ancestors of these peoples come from the area around Lake Baikal and that they may have migrated north because of pressure from other groups migrating or expanding into their territories (Alekseev 1996; Janhunen 1996; Vasilevich 1969). Another Turkic-speaking population, the Kyrgyz, also migrated to their current location in Central Asia from the Upper Yenissei region west of Lake Baikal as recently as the 15th to 17th centuries (Bregel 2003; Lebeynsky

2007). The modern inhabitants of the Baikal-Altai region, who might represent the descendants of the Late Neolithic populations, are the Mongolic-speaking Buryats and the Turkic-speaking Tofa, Tuvan, and Altai peoples.

The Evens and Evenks were the first of the aforementioned populations to migrate north into what is currently Siberia. A common origin of the Evens and Evenks has been suggested from mtDNA and Y-chromosomal analyses (Pakendorf et al., 2007). There is, however, debate over the geographic location of their origins. Vasilevich (1969) proposes an ancestral population south of Lake Baikal that moved into the Lake Baikal region before being split into the ancestors of modern-day Evens and Evenks by the arrival of Turkic populations in the mid-first millennium AD, which furthermore led to the northward migration of the Evenks and Evens. An alternative hypothesis, provided by Janhunen (1996), is that these Northern Tungusic populations come from ancestral populations who lived in Manchuria, followed by a migration to the Baikal region and subsequent move northward in the 12th century. In their northern territory they were split due to pressures of the Yakut immigration and subsequent expansion. Both of these hypotheses allow for contact between the northern Tungusic-speaking populations and the Turkic-speaking populations inhabiting the Baikal region who are thought to be the ancestors of the Yakuts. In both of these scenarios, the Evens and Evenks moved north prior to the Yakut ancestors. The primary difference between these two hypotheses is whether the Tungusic peoples were split due to arrival of Turkic populations around Lake Baikal and then migrated north, or rather split due to arrival and expansion of Yakut ancestors after they had already migrated north. With the latter, the Tungusic-speaking populations should be more similar to one another than in the former hypothesis.

Recent data based on autosomal genome-wide data has shown low levels of differentiation and no structure in the Even and Evenks and suggests that rather than originating in the region around Lake Baikal, they come from the area around the Amur River (Pugach et al., 2015). This refutation of a Baikal origin is based primarily on the lack of a type of European signal in the Tungusic-speaking populations that is otherwise seen in the populations around Lake Baikal which is lacking in the Tungusic-speaking populations. It was also suggested, based on Y chromosomal data in Duggan et al. (2013), that, depending on which mutation rate is used, it is possible to support either the early or late split described above. However, there exists a preference for the later split due to the pedigree based mutation rate being shown in a study on Yakut history (Pakendorf et al., 2006) to align better with associated archaeological and linguistic evidence (Duggan et al., 2013). Taken with the more recent split date given based on Y chromosomal haplogroup C data presented in (Malyarchuk et al., 2010) it was therefore suggested that the Janhunen (1996) hypothesis described has more support based on paternal data (Duggan et al., 2013). The mtDNA data was unable to lend support to either hypothesis, but no study has looked at complete mtDNA genomes across these Tungusic populations and the Yakut while also including multiple South Siberian populations. These combined data could better address this question and provide a last line of evidence in determining which of these hypotheses is more likely.

Based on archaeological finds around Lake Baikal, and associated runic inscriptions, the ancestors of the Yakuts are thought to have been a Turkic-speaking population called the Kurykans (Alekseev 1996; Gogolev 1993; Konstantinov 1975; Okladnikov 1955). Their lifestyle, including cattle- and horse-pastoralism, is a distinctly southern feature. A recent study on Yakut horses (Librado et al., 2015) shows that they are not descended from the now-extinct, native horse populations which populated this region prior to the arrival of the Yakuts and, thus, were likely brought with the Yakuts when they migrated north. They suggest a rapid evolutionary adaptation to the cold similar to what has been found in wooly mammoths and some humans. This can explain their more stout body shape and heavy winter coats compared to other domesticated horses. Linguistic evidence points to a southern origin for the Yakuts in that a significant percentage of the Yakut lexicon can be attributed to borrowing from Mongolic (Kaluzynski 1962; Pakendorf and Novgorodov 2009). Additionally, linguistic evidence shows structural changes which the Yakut language has undergone due to Evenki influence (Pakendorf 2007). Studies on mtDNA and the Y-chromosome from both modern and ancient DNA show that the Yakuts underwent a founder event roughly 800 to 1000 years ago (Crubezy et al., 2010; Pakendorf et al., 2006; Zlojutro et al., 2008; Zlojutro et al., 2009). These studies have also shown an affinity between the Yakuts, South Siberian and Mongolian populations.

The ancestors of the Kyrgyz underwent conflicts with various Turkic populations and eventually defeated the Uygurs in the 9th century CE. By the 10th century, after being forced to abandon their lands in northwestern Mongolia, they were located in the upper Yenisei where they remained until being defeated by and integrated into the Mongol empire in the 13th century. Eventually they migrated to what is now modern Kyrgyzstan in the 15th to 17th centuries (Lebeynsky 2007).

Despite linguistic and cultural differences between populations, previous research on mtDNA hypervariable region 1 (HV1) sequences has revealed a large amount of sequence type sharing among Siberian populations (Bregel 2003; Fedorova et al., 2003; Pakendorf et al., 2006; Pakendorf et al., 2007; Zlojutro et al., 2008). There are multiple possibilities for the pattern of mtDNA HV1 sequence type sharing seen between these populations that are separated by vast distances. First, this could be a result of relatively recent shared ancestry, given that the Turkic and Tungusic-speaking populations are thought to have inhabited a common territory. Second, this pattern could be caused by admixture, either between ancestral populations in southern Siberia around Lake Baikal, or more recent, post-migration admixture between populations after their northward expansion. Alternatively, it could be a combination of both shared ancestry and recent admixture.

In this chapter, I report on the sequencing and analyses of complete mtDNA genomes from 715 individuals from 18 populations belonging to 11 ethnolinguistic groups using the Illumina Genome Analyzer IIx. These populations include (1) the modern inhabitants of the region between Lake Baikal and the Altai Mountains, (2) those, like the Yakuts, Kyrgyz, Evenks, and Evens, whose ancestors are thought to have lived there in the past and (3) additional comparative populations including geographic and linguistic neighbors. I examine the genetic structure of these populations to uncover their potential shared ancestry and admixture, and the putative migrations of populations away from the Lake Baikal region.

Materials

DNA samples

715 samples that were collected with informed consent and ethical approval by the relevant institutional review boards were used in research leading up to this study. Table 4.1 includes the sample sizes, approximate locations and linguistic affiliations

of the populations. Of these samples, the complete mtDNA genomes for most of the Yakuts (excluding the CEPH Yakuts), Even, Evenk, and Yukaghir were published in Duggan et al. (2013). There are minor differences in numbers of individuals in the following populations: Yukaghir includes 2 new sequences here; Central Yakut includes 1 new sequence here, Northeast Yakut includes 1 new sample here, Vilyuy Yakut includes 1 new sample here and the CEPH panel Yakuts are new here along with all collected samples from the Southern populations described below. Because the CEPH Yakuts are of Central Yakutian origin and clustered with the Central Yakuts in preliminary analyses, they were grouped with the Central Yakuts for analyses presented here. The Mongolian samples were collected in various locations in Mongolia, the Eastern Buryats were collected in the village of Elesun in the Buryat Republic, the Western Buryat samples were collected in the village of Gakhany in the Ust Orda Okrug of the Irkutsk Oblast, and the Tofa samples were collected in Alygdjer village of the Nizhne Udinsk district of the Irkutsk Oblast. Details on the Even, Evenk, Altai, Kyrgyz, Tuvan, Yakut, YSE and Yukaghir samples can be found in previous publications (Kaessmann et al., 2002; Martinez-Cruz et al., 2010; McComb et al., 1996; Pakendorf et al., 2006; Pakendorf et al., 2007; Phillips-Krawczak et al., 2006; Ségurel et al., 2008). Even and Evenk samples are grouped differently than in Duggan et al. (2013) due to the focus being more on the Central and Southern populations and those labels are given in Table 4.1. Additionally some of the more eastern populations were excluded (some Evens, Koryak, Nivkh). The full breakdown of the populations described here is given in Table 4.1. Most of the population subgroups are based on geographical designations with the exception of the Kyrgyz which are given the acronyms Kyrgyz B and Kyrgyz L denoted in Martinez-Cruz et al. (2010) and come from Eastern Kyrgyzstan. Because of the reasons given in Chapter 3 on ascertainment bias in existing published complete mtDNA genomes, additional samples from published data were not incorporated in these analyses with the exception of those highlighted herein from our previous study on Tungusic peoples (Duggan et al., 2013) that we generated in-house without any sub-selection of samples.

Table 4.1. Affiliations and sample sizes of the populations included in this study.

Population	Ethnolinguistic Affiliation	Acronym	n	Linguistic Affiliation
Mongolian	Mongolian	Mongol	56	Mongolic
Eastern Buryat	Buryat	E_Bur	27	Mongolic
Western Buryat	Buryat	W_Bur	48	Mongolic
Kyrgyz B	Kyrgyz	Kyr_B	30	Turkic
Kyrgyz L	Kyrgyz	Kyr_L	27	Turkic
Altai	Altai	Altai	33	Turkic
Tofalar	Tofalar	Tofa	23	Turkic
Tuvan	Tuvan	Tuvan	59	Turkic
Central Yakut	Yakut	C_Yak	89	Turkic
CEPH Yakut	Yakut	C_Yak	24	Turkic
Vilyuy Yakut	Yakut	V_Yak	55	Turkic
Northeastern Yakut	Yakut	NE_Yak	33	Turkic
Yakut-speaking Evenk	Yakut-speaking Evenk	YSE	32	Turkic
Stony Tunguska Evenk	Evenk	ST_Evk	39	Tungusic
Nyukzha Evenk	Evenk	Ny_Evk	45	Tungusic
Iengra Evenk	Evenk	Ie_Evk	23	Tungusic
Central Even	Even	C_Evn	26	Tungusic
Western Even	Even	W_Evn	24	Tungusic
Yukaghir	Yukaghir	Yukag	22	Isolate

Results

Haplogroup composition

Many of the Siberian populations included in this study are similar with respect to their haplogroup composition (Figure 4.1, Table 4.2), despite the vast geographic distances that separate them and regardless of their linguistic differences. Haplogroups C and D constitute the most frequent haplogroups in most of the populations included in this study, with frequencies ranging from 7.4% in the Kyrgyz L to 76.9% in the Stony Tungusic Evenks for haplogroup C, and frequencies of 0% in the Tofa to 56.5% in the Iengra Evenks for haplogroup D (Table 4.2). Haplogroup C is particularly common in populations of northeastern Siberia (Evens, Evenks, Yukaghirs, and NE Yakuts), while it is less common in southern Siberia; a notable exception are the South Siberian Turkic-speaking Tuvans and Tofa, who have 55.9% and 60.9% haplogroup C, respectively. Haplogroup D is found at lower frequencies overall, relative to haplogroup C. Sub-haplogroup D5 (excl. D5b) has a more limited distribution: it is present at frequencies of more than 10% in some Yakut populations, YSE and Iengra Evenks, but is otherwise rare or absent.

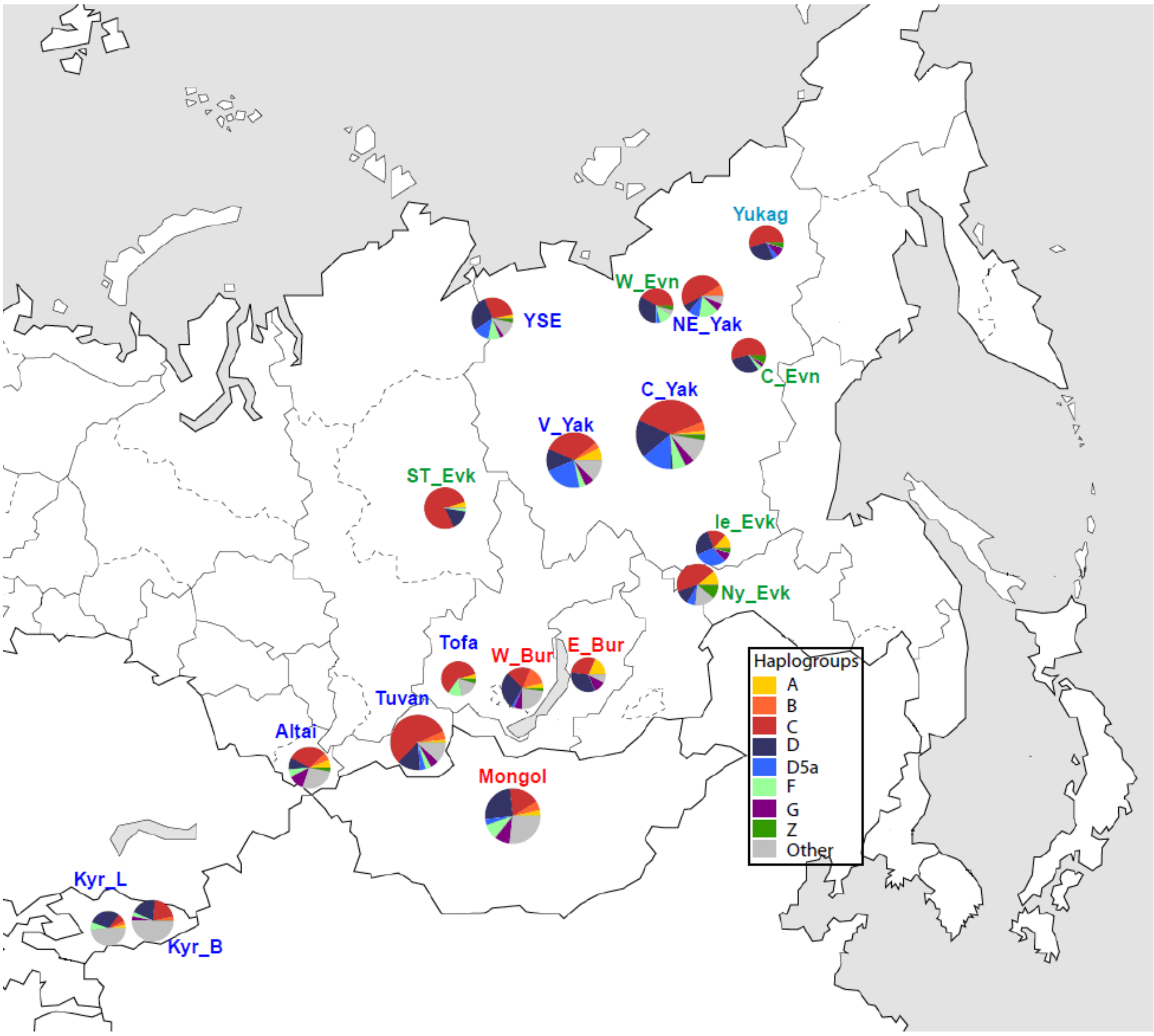

Fig. 4.1. Map of Siberia showing the approximate position of sampling locations and frequencies of major mtDNA haplogroups. The size of each pie chart represents the relative sample sizes of the populations included in this study. The colors of the population labels vary according to linguistic affiliation (Turkic, blue; Tungusic, green; Yukaghir, aqua; Mongolic, red). Haplogroup D shown in dark blue excludes D5a.

The non-Siberian populations (Mongolians and Kyrgyz) stand out by exhibiting higher frequencies of haplogroups that are either absent or at very low frequencies in the Siberian populations. For example, haplogroup U, typically found in Western Eurasia is present at a frequency of 10.5% in the Kyrgyz and 7.1% in the Mongolians. Other haplogroups of interest that are found Siberia and in parts of East Asia and are present in multiple populations in this study include G, B and A.

Haplogroups B and A are also found in indigenous peoples of the Americas. Haplogroup G is absent in the Evens and Evenks but is present in frequencies ranging from 5.0% and 5.1% in the Yakuts and Tuvans to 12.1% in the Altai. This haplogroup is also found at frequencies close to 10% in the Mongolians, Buryats and Yukaghirs. Haplogroup B is absent in the Evens, Evenks, and Yukaghirs, but present at between 5% and 10% in the Altai, Buryats, Mongolians, Tuvans and Yakut. Haplogroup A is of note because it exhibits a frequency of 9.3% in both the Buryats and Evenks, but is missing in the Evens and Yukaghirs. Of these, the vast majority (83%) are in A4.

Table 4.2. Major haplogroup frequencies

Haplogroup:		A	B	C	D (excl. D5)	D5 (excl. D5b)	D5b	F	G	H	HV	J	K	M	N	R	T	U	W	Y	Z
Population:	n=																				
Mongol	56	3,6	5,4	17,9	25,0	3,6		8,9	8,9	5,4	3,6			8,9				7,1	1,8		
W_Bur	48	4,2	14,6	18,8	29,2	2,1			6,3	6,3			4,2	4,2	2,1	2,1	4,2				2,1
E_Bur	27	18,5		29,6	33,3				11,1	3,7								3,7			
Buryat Total	*75*	*9,3*	*9,3*	*22,7*	*30,7*	*1,3*			*8,0*	*5,3*			*2,7*	*2,7*	*1,3*	*1,3*	*2,7*	*1,3*			*1,3*
Kyr_B	30			20,0	20,0			6,7	3,3	3,3	6,7	10,0	3,3	6,7				16,7		3,3	
Kyr_L	27	3,7	3,7	7,4	18,5		11,1	7,4		25,9		3,7	3,7		7,4		3,7	3,7			
Kyrgyz Total	*57*	*1,8*	*1,8*	*14,0*	*19,3*		*5,3*	*7,0*	*1,8*	*14,0*	*3,5*	*7,0*	*3,5*	*3,5*	*3,5*		*1,8*	*10,5*		*1,8*	
Altai	33	6,1	6,1	30,3	9,1			6,1	12,1	3,0			6,1	12,1				6,1			3,0
Tofa	23	4,3		60,9				13,0		8,7		8,7									4,3
Tuvan	59	1,7	5,1	55,9	13,6	3,4		3,4	5,1					3,4			1,7	1,7		5,1	
C_Yak	89	2,2	4,5	39,3	16,9	14,6		6,7	3,4	1,1		1,1		1,1		1,1	1,1	2,2	1,1	2,2	1,1
CEPH_Yak	24		4,2	33,3	16,7	12,5	4,2	4,2	8,3			4,2					4,2				8,3
V_Yak	55	7,3	3,6	32,7	12,7	21,8		3,6	5,5	3,6	3,6			1,8			1,8		1,8		
NE_Yak	33		9,1	48,5	6,1	9,1		15,2	6,1	3,0	3,0										
Yakut Total	*201*	*3,0*	*5,0*	*38,3*	*13,9*	*15,4*	*0,5*	*7,0*	*5,0*	*2,0*	*1,5*	*1,0*		*1,0*		*0,5*	*1,5*	*1,0*	*1,0*	*1,0*	*1,5*
YSE	32	3,1		28,1	28,1	12,5		9,4	3,1	9,4				3,1							3,1
ST_Evk	39	5,1		76,9	15,4			2,6													
Ny_Evk	45	11,1		44,4	11,1	6,7				8,9		4,4		2,2							11,1
Ie_Evk	23	13,0		17,4	26,1	30,4			8,7												4,3
Evenk Total	*107*	*9,3*		*50,5*	*15,9*	*9,3*		*0,9*	*1,9*	*3,7*		*1,9*		*0,9*							*5,6*
C_Evn	26			53,8	30,8			3,8	3,8												7,7
W_Evn	24			41,7	33,3	4,2		12,5						4,2							4,2
Even Total	*50*			*48,0*	*32,0*	*2,0*		*8,0*	*2,0*					*2,0*							*6,0*
Yukag	22			54,5	27,3	4,5			9,1												4,5

Patterns of mtDNA diversity

Table 4.3 shows the mtDNA diversity values in the populations analyzed in this study. The highest diversity values are present in the Mongolians and the Kyrgyz populations in the south, with gene diversity values (GD, the probability of randomly choosing two different haplotypes in the sample) from 0.989 to 0.999 and mean number of pairwise differences (MPD, across the whole mtDNA genome excluding the poly-C region) over 30. Conversely, the Tofa have the lowest GD value of 0.834, which could be explained by genetic drift, while exhibiting an intermediate value of 25.91 for MPD. The Evens, Evenks and Yukaghirs, while exhibiting relatively high GD values (0.939-0.972), mostly show lower MPD values (18.40-26.03), with the Central Evens, Iengra Evenks, Stony Tungusic Evenks, and Yukaghirs exhibiting particularly low MPD values (18.40-21.76) .

Table 4.3. mtDNA diversity values

Population	**N**	**n**	**GD**	**SE**	**S**	**MPD**	**SE**
Mongol	56	54	0.999	0.0037	404	32.6	14.43
E_Bur	27	20	0.969	0.0206	135	26.76	12.1
W_Bur	48	33	0.986	0.0065	254	29.37	13.06
Kyr_B	30	29	0.998	0.0094	269	31.76	14.26
Kyr_L	27	23	0.989	0.0131	223	30.08	13.56
Altai	33	19	0.964	0.0146	196	29.67	13.31
Tofa	23	10	0.834	0.0675	101	25.91	11.8
Tuvan	59	40	0.981	0.0079	273	25.71	11.44
C_Yak	113	79	0.986	0.0047	391	27.64	12.19
V_Yak	55	40	0.979	0.0107	240	27.67	12.3
NE_Yak	33	24	0.979	0.0124	169	28.89	12.96
YSE	32	19	0.958	0.0176	160	27.46	12.35
ST_Evk	39	18	0.941	0.0176	112	18.4	8.34
Ny_Evk	45	24	0.967	0.011	154	26.03	11.63
Ie_Evk	23	13	0.945	0.0259	87	20.17	9.25
C_Evn	26	19	0.972	0.0183	120	21.76	9.91
W_Evn	24	17	0.964	0.0239	125	25.33	11.52
Yukag	22	14	0.939	0.0325	92	20.37	9.36

N, number of samples; n, number of haplotypes; GD, gene diversity; S, number of segregating sites; MPD, mean number of pairwise differences; SE, standard error

An initial correspondence analysis (CA), a statistical technique used to summarize categorical data into a 2D graph, based on mtDNA haplogroups including all populations, showed the Kyrgyz populations to be quite distant from each other

(Figure 4.2). These two populations, along with the E. Buryats were considered to be outliers and removed from further CA analyses.

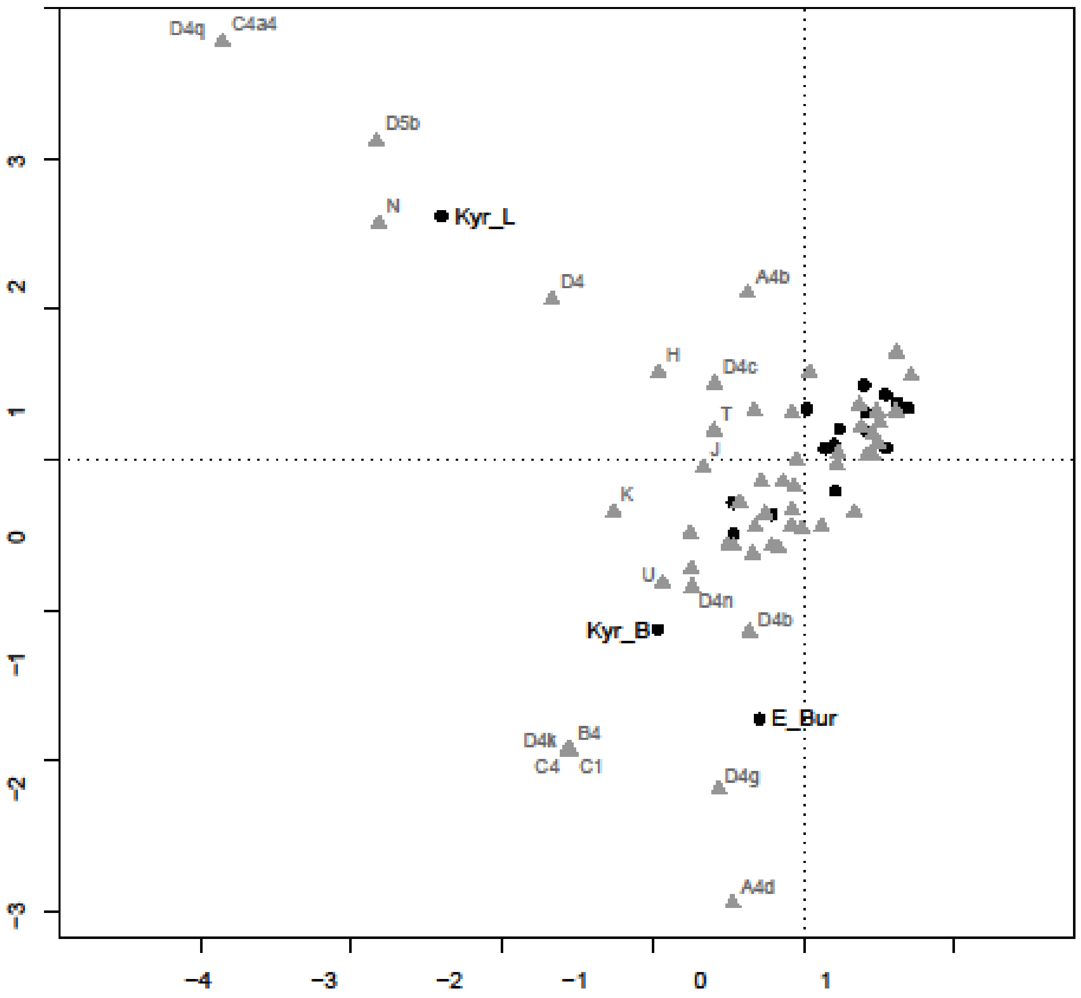

Fig. 4.2. Correspondence analysis (CA) based on haplogroups for all populations included in this study, highlighting the outliers.

The CA based on haplogroup frequencies shows little genetic structure in Siberia, though the W. Buryats cluster with the Mongolians (Figure 4.3). The separation of Evenki populations is apparent and shows that even populations speaking dialects of the same language do not always cluster together.

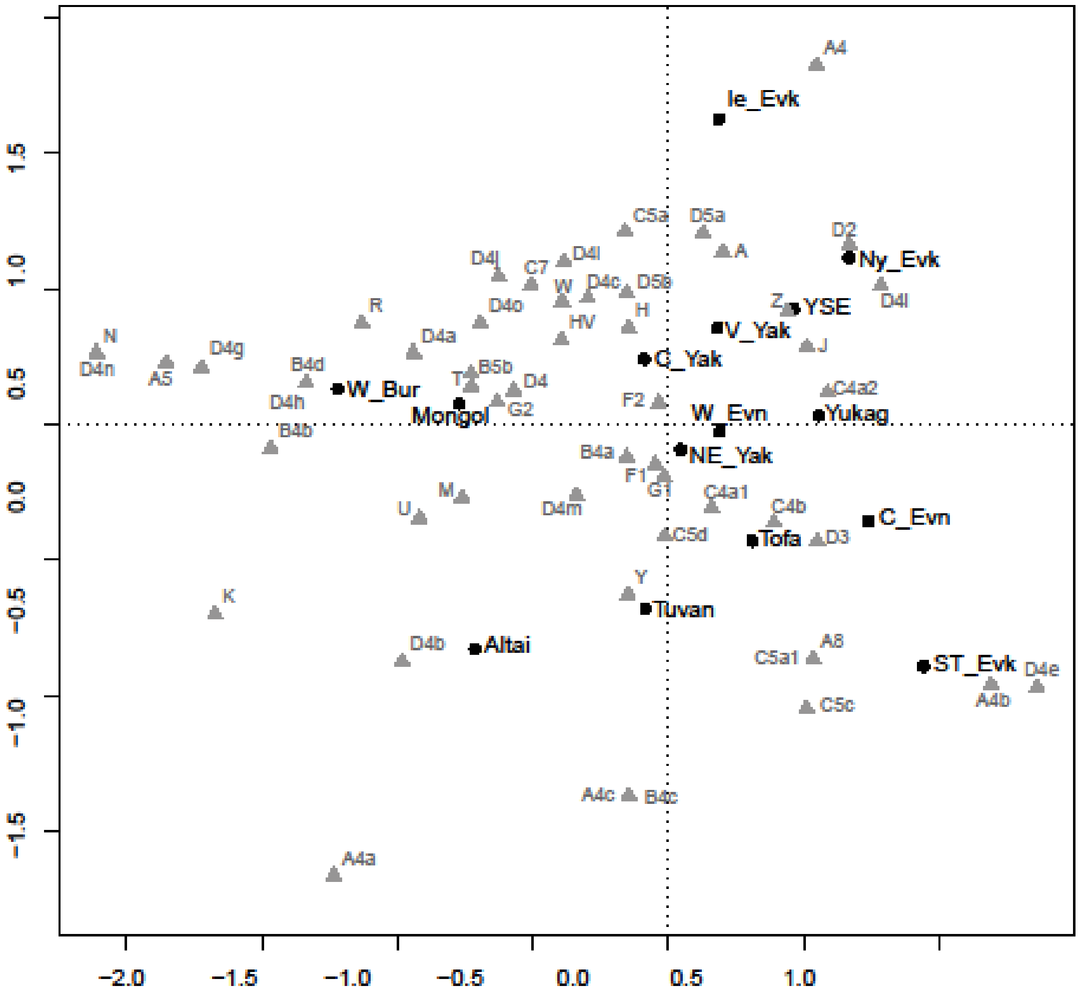

Fig. 4.3. Correspondence analysis (CA) based on sub-haplogroups. The Kyrgyz populations and E_Bur were removed for better visualization.

The main populations that are separated along the x-axis in Figure 4.3 are the Western Buryats, Mongolians and the Altai, all populations presently residing in the south. However, the Tuvans and Tofa, two Turkic-speaking populations also settled in southern Siberia, show affinities with northeastern Siberian populations rather than with the other southern groups, though in the y-axis the Tuvans are closer to the

Altai. The Stony Tunguska Evenks and Iengra Evenks are also pulled away from the main group showing distinct separation between the Evenk populations.

This limited genetic structure among the populations of Siberia is further confirmed by a Multidimensional Scaling plot (MDS) based on pairwise Φst values (Figure 4.4). This lack of structure is strikingly exemplified by the Tuvans. Geographically, the Tuvans are located near their close linguistic relatives, the Altai and Tofa, as well as in the vicinity of the Mongolic-speaking populations. However, in the MDS plot they are located closer to the Tungusic-speaking populations (and are especially close to the Central Evens and Yukaghirs in the third dimension, cf. Figure 4.5). The Tofa appear to be an outlier in Figure 4.4, which could be explained by genetic drift in accordance with their low diversity values. This differentiation is largely from dimension 2, as the Tofa align with the Tuvans in dimension 1 and Northeastern Yakuts and Kyrgyz L in dimension 3.

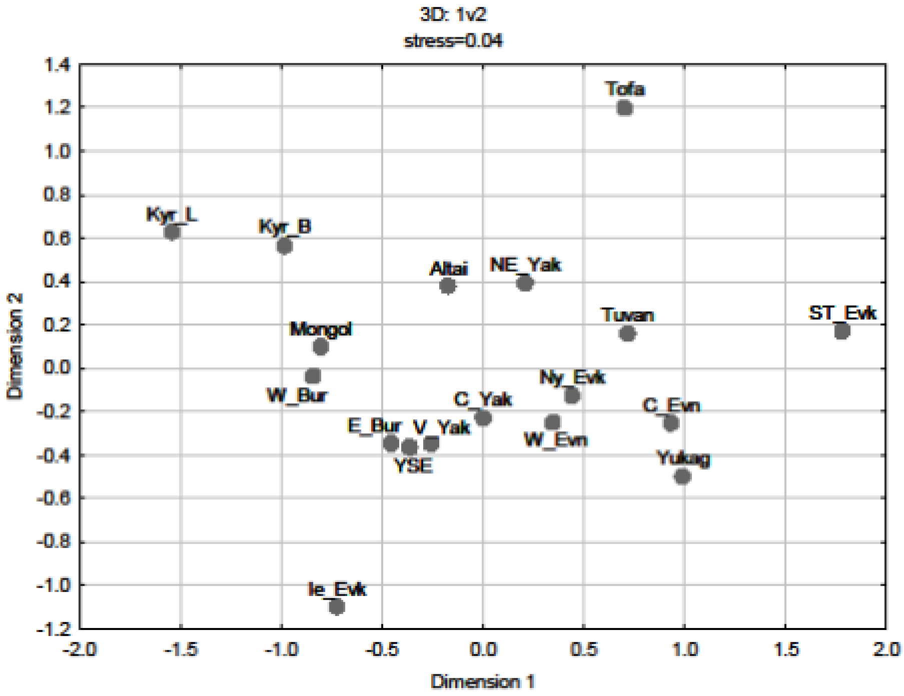

Fig. 4.4. Multidimensional Scaling plot (MDS) based on pairwise Φst values for dimensions 1 versus 2.

The Eastern Buryats form a tight group with the Vilyuy Yakuts and YSE despite being linguistically (Mongolic vs. Turkic) and geographically distant – though the third dimension separates the Buryats from the Yakut-speaking groups (Figures 4.5 and 4.6). In contrast, the Western Buryats cluster with the Mongolians (and in dimensions 1 and 3 with the Kyrgyz B). The most widely dispersed set of populations within a single ethnolinguistic group in this analysis, just as was seen in the CAs, are the Evenks.

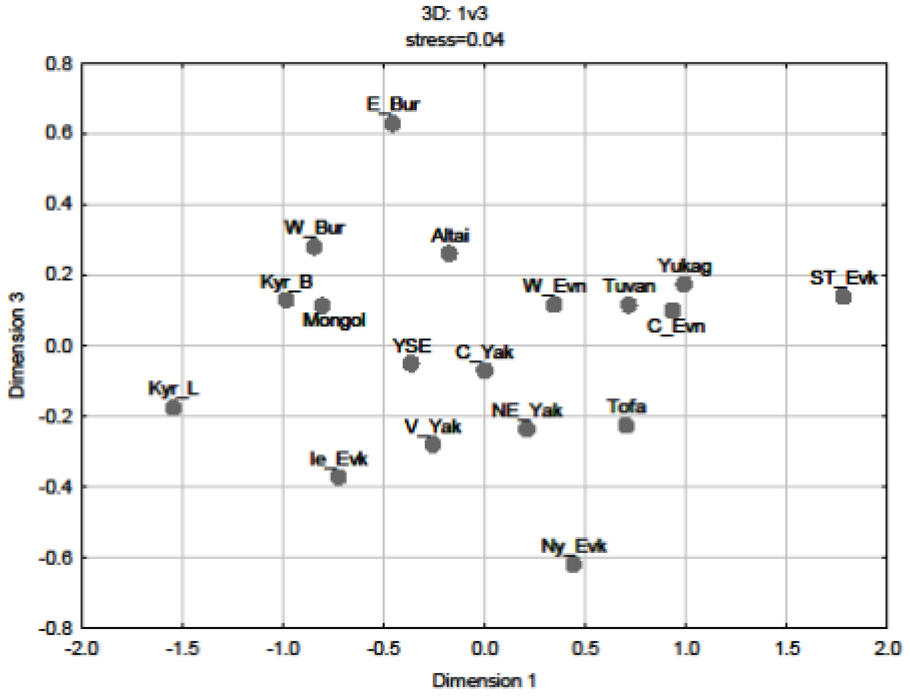

Fig. 4.5. MDS plot based on pairwise Φ_{ST} values for dimensions 1 versus 3.

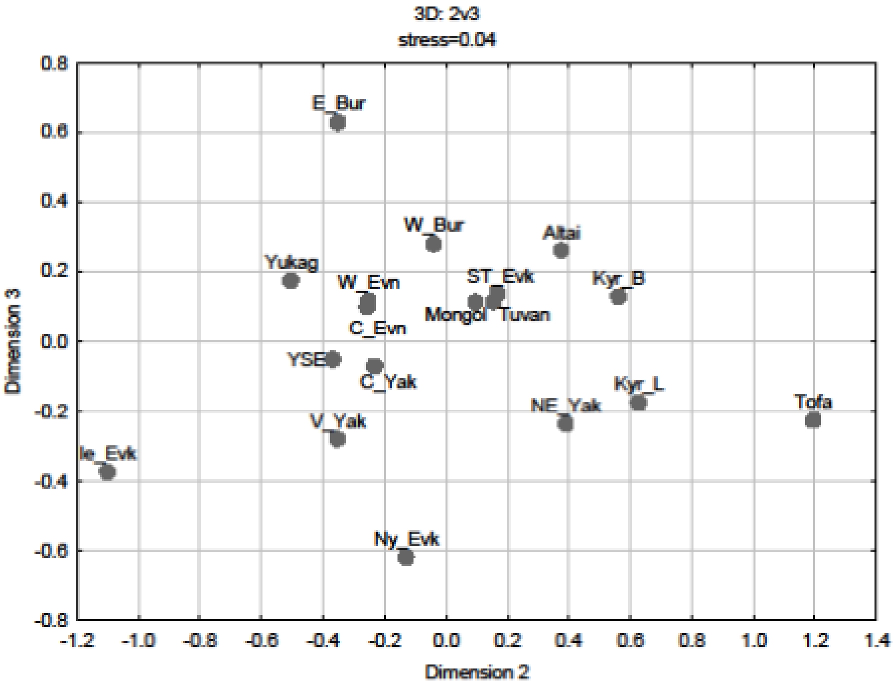

Fig. 4.6. MDS plot based on pairwise Φ_{ST} values for dimensions 2 versus 3.

Figure 4.7 shows the values of pairwise Φ_{ST} distances in a matrix, highlighting the significantly dissimilar pairs of populations after Bonferroni correction (red diamonds), and additionally those population pairs that are not significantly differentiated even without Bonferroni correction (green circles). The Kyrgyz populations stand out as being very different from the other populations. However, the Kyrgyz B population shows more affinity to the Altai, Mongolians, and Western Buryats than does Kyrgyz L, just as in the MDS plots.

As was seen in the MDS plot, the ST Evenks and Iengra Evenks are highly differentiated. The Mongolians are not significantly differentiated from the Buryats, Kyrgyz, Altai, and the YSE. The Yakut populations are all very close genetically to one another and to some of the Tungusic speaking populations, with the exception of the ST Evenks.

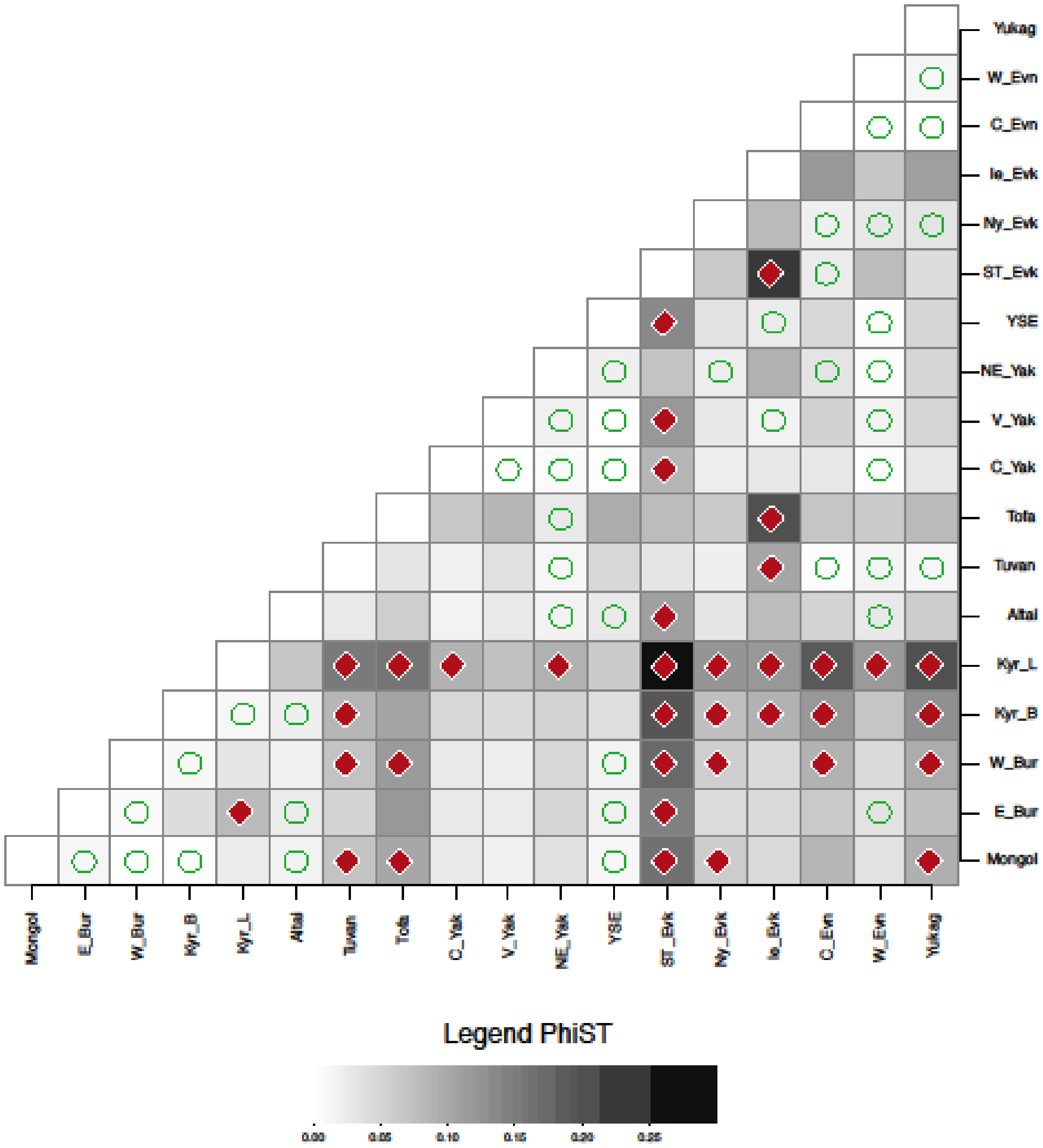

Fig. 4.7 Matrix of Φst pairwise distances. Darker shading denotes higher Φst values. Green circles mark those pairs of populations that are non-significant before Bonferroni correction (most similar). Red diamonds show the pairs of populations that are significant after Bonferroni correction (most dissimilar).

The low level of structure based on linguistic or geographic segmentation among these populations is also shown by AMOVA analyses of Φ_{ST} values (Table 4.4). The differences between the 18 populations included in this study account for 4.91% of the variance (P<0.001); grouping them according to their linguistic affiliation results in 3.73% of the variance being accounted for by differences between the populations belonging to each language family, and only 1.79% of the variation being due to differences between the linguistic groups. Similarly, the proportion of variance due to differences between geographic groups (southwest vs. northeast) is far smaller (1.4%) than that due to differences between populations within each geographic grouping (4.14%).

Table 4.4. Analysis of molecular variance (AMOVA)

Grouping	Among groups	Among pops within groups	Within pops
Linguistic	1.79*	3.73**	94.48
Geographic	1.4*	4.14**	94.46
No grouping		4.91**	95.09

*P < 0.05; **P < 0.001

The linguistic grouping followed the affiliation given in Table 4.1. The geographic grouping was done by including all populations south of, and including, the Tofa and Buryats as a southwestern group and all other populations as northeastern.

Analyses of shared haplotypes

An analysis of sequence type (identical complete mtDNA genomes excluding poly-C region) sharing shows a distinction between Southern populations including the Kyrgyz, the Mongolians, and the southern Siberian groups, who share few haplotypes with others, and populations from northeastern Siberia, where the amount of sequence type sharing is very high (Figure 4.8). This is especially apparent amongst the Yakut populations, but also between the Northeastern Yakuts and Western Evens and between the Nyukzha Evenks and Iengra Evenks. It is notable that the Tuvans and Altai do share an appreciable amount of sequence types with the northeastern Siberian populations and the W. Buryat share sequence types with the Yakuts. The Tofa and Kyrgyz stand out in that they share hardly any sequence types with the other populations included in the study (Figure 4.8), with the exception of the Tuvans, and the two Kyrgyz populations do not even share any sequence types with each other.

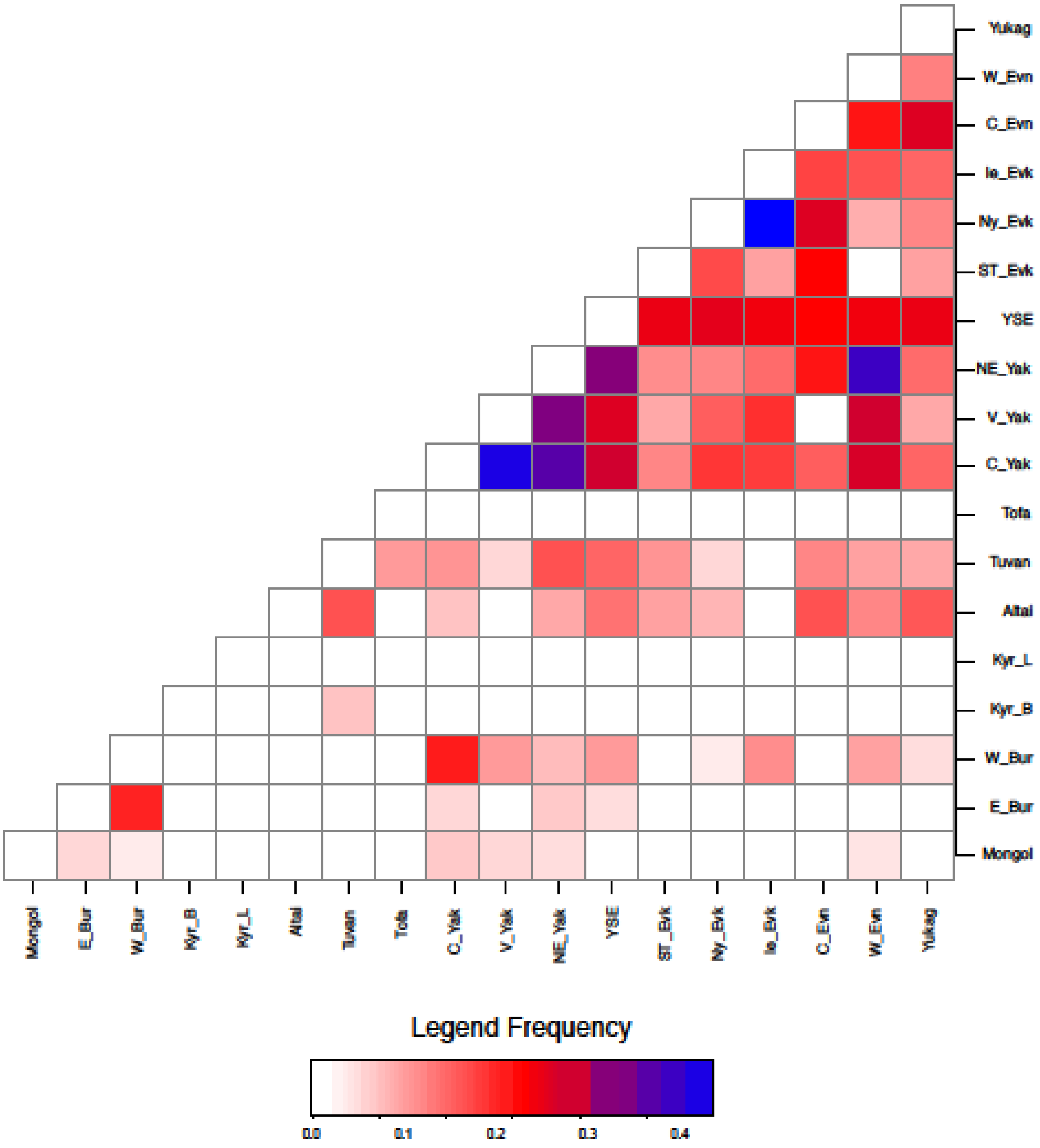

Fig. 4.8. Haplotype sharing heat plot.

Figure 4.9 shows the same analysis broken down by haplotypes and their associated haplogroup. Using this Figure and the frequencies of the haplotypes in the populations shown in Table 4.5, it is possible to infer the potential directionality of gene flow when haplotypes are present in only two populations. A two-fold difference in haplotype frequencies between pairs of populations was assumed to indicate geneflow from the population with higher frequency to the population with lower frequency.

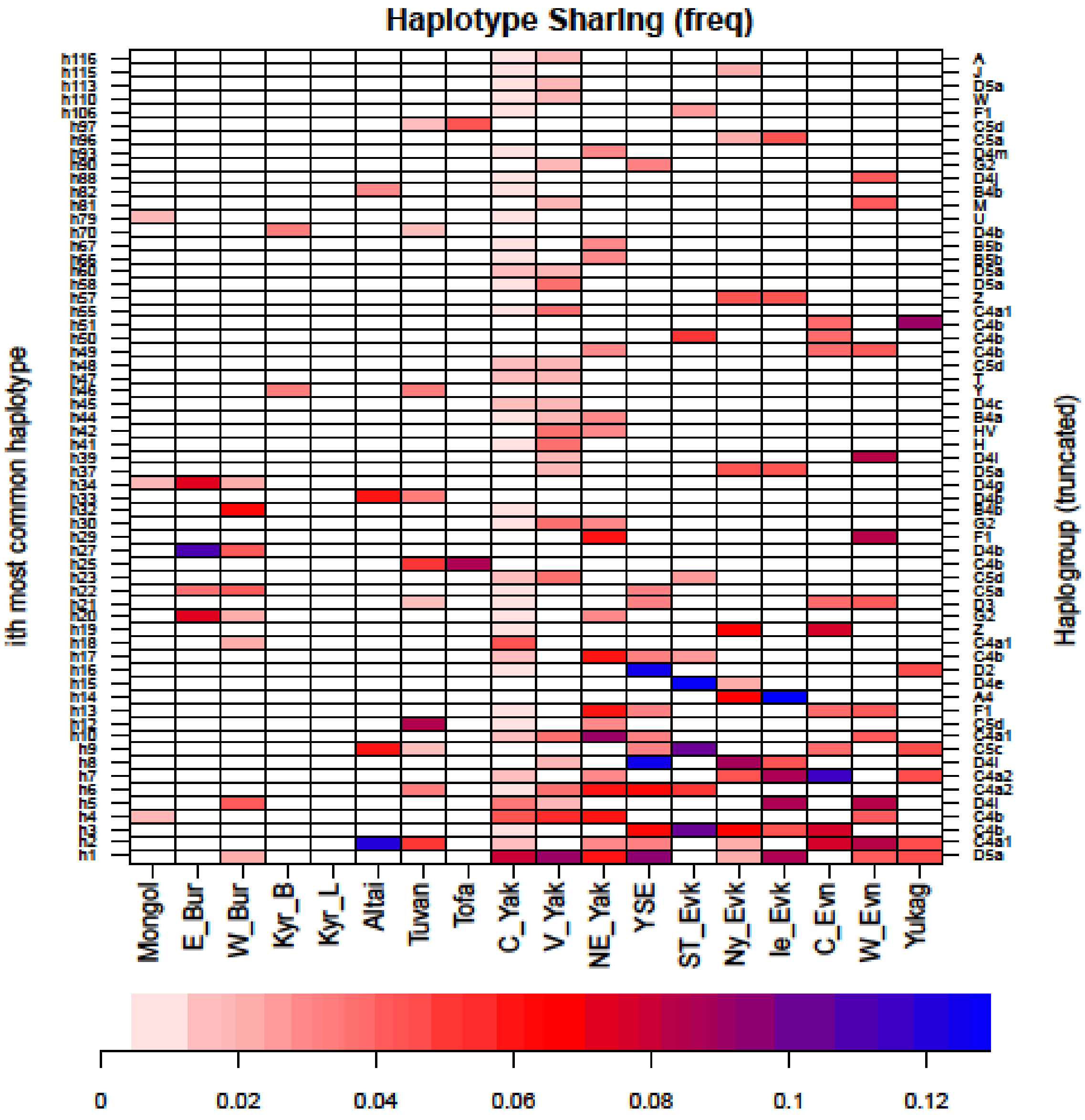

Fig. 4.9. Haplotype sharing heat plot showing all haplotypes shared by at least two populations, decreasing in sharing frequency from the bottom to the top of the plot, along with their associated haplogroups shown on the right.

Sixteen haplotypes are shared only between Yakut and Tungusic-speaking populations (Evens and Evenks), with eight of these being suggestive of geneflow from Tungusic groups to Yakuts (Table 4.5 haplotypes: 3, 8, 19, 37, 39, 49, 81, 88) and four suggestive of geneflow from Yakuts to Tungusic populations (Table 4.5 haplotypes: 4, 10, 17, 23). The other four do not meet the criteria of a twofold difference in frequencies (Table 4.5 haplotypes: 13, 29, 106, 115).

Of the haplotypes shared only between the Yakut and Tungusic populations, all but two are shared either between the Yakuts and Evens or Yakuts and Evenks which could be an indicator of relatively recent population contact rather than shared ancestral sequences. The only two haplotypes shared by all three populations have frequencies of 0.88%, 21.27%, and 7.69% (haplogroup C4b, haplotype 3), and 0.88%, 6.67%, and 7.69% (haplogroup Z, haplotype 19) for the Yakuts, Evenks, and Evens, respectively, thus suggesting a directionality of gene flow from Northern Tungusic groups to Yakuts. There are four haplotypes shared only between Yakuts and Buryats (Table 4.5 haplotypes: 18, 20, 22, 32), with three suggestive of geneflow from Buryats to Yakuts (haplogroups B4b, C5a, and G2). Additionally, there are two haplotypes shared only between Altai/Tuvan and Yakuts (haplogroups B4b and C5d, haplotypes from Table 4.5: 12, 82) suggesting directionality from Altai/Tuvan to the Yakuts. Given the putative migration of the Yakuts away from the southern region, this directionality could suggest at least a minor amount of continued contact and maternal geneflow in a northerly direction between the populations rather than a single migratory event.

Table 4.5. Percentages of haplotypes in the populations

Haplotype	Mongol	E_Bur	W_Bur	Kyr_B	Kyr_L	Altai	Tuvan	Tofa	C_Yak	V_Yak	NE_Yak	YSE	ST_Evk	Ny_Evk	Ie_Evk	C_Evn	W_Evn	Yukag	*Haplogroup*
h116									0,88	1,82									A
h115									0,88					2,22					J
h113									0,88	1,82									D5a
h110									0,88	1,82									W
h106									0,88				2,56						F1
h97							1,69	4,35											C5d
h96														2,22	4,35				C5a
h93									0,88		3,03								D4m
h90										1,82		3,13							G2
h88									0,88								4,17		D4j
h82						3,03			0,88										B4b
h81										1,82							4,17		M
h79	1,79								0,88										U
h70				3,33			1,69												D4b
h67									0,88		3,03								B5b
h66									0,88		3,03								B5b
h60									1,77	1,82									D5a
h58									0,88	3,64									D5a
h57														4,44	4,35				Z
h55									0,88	3,64									C4a1
h51																3,85		9,09	C4b
h50													5,13			3,85			C4b
h49											3,03					3,85	4,17		C4b
h48									1,77	1,82									C5d
h47									1,77	1,82									T
h46				3,33			3,39												Y
h45									1,77	1,82									D4c
h44									0,88	1,82	3,03								B4a
h42										3,64	3,03								HV
h41									0,88	3,64									H
h39										1,82							8,33		D4l
h37										1,82				4,44	4,35				D5a
h34	1,79	7,41	2,08																D4g
h33						6,06	3,39												D4b
h32			6,25						0,88										B4b
h30									0,88	3,64	3,03								G2
h29											6,06						8,33		F1
h27		11,11	4,17																D4b
h25							5,08	8,70											C4b
h23									1,77	3,64			2,56						C5d
h22		3,70	4,17						0,88			3,13							C5a
h21							1,69		0,88			3,13				3,85	4,17		D3
h20		7,41	2,08						0,88		3,03								G2
h19									0,88					6,67		7,69			Z
h18			2,08						4,42										C4a1
h17									1,77		6,06	3,13	2,56						C4b
h16									0,88			12,50						4,55	D2
h15													12,82	2,22					D4e
h14														6,67	13,04				A4
h13									0,88		6,06	3,13				3,85	4,17		F1
h12							8,47		0,88		3,03								C5d
h10									1,77	3,64	9,09	3,13					4,17		C4a1
h9						6,06	1,69					3,13	10,26			3,85		4,55	C5c
h8										1,82		12,50		8,89	4,35				D4l
h7									1,77		3,03			4,44	8,70	11,54		4,55	C4a2
h6							3,39		0,88	3,64	6,06	6,25	5,13						C4a2
h5			4,17						3,54	1,82					8,70		8,33		D4l
h4	1,79								4,42	5,45	6,06						4,17		C4b
h3									0,88			6,25	10,26	6,67	4,35	7,69			C4b
h2						12,12	5,08		1,77		3,03	3,13		2,22		7,69	8,33	4,55	C4a1
h1			2,08						7,96	9,09	6,06	9,38		2,22	8,70		4,17	4,55	D5a

Phylogenetic networks were created for haplogroups G, A/B combined, C, and D (Figures 4.10-4.13) due to the combinations of populations with these haplogroups. The network for Haplogroup G in Figure 4.10 shows sharing only in two nodes, both in sub-haplogroup G2. There is sharing between a Yakut and a YSE on the lower part of the network and between the Yakut and Buryat with some nodes of Mongolians very close in the upper left area.

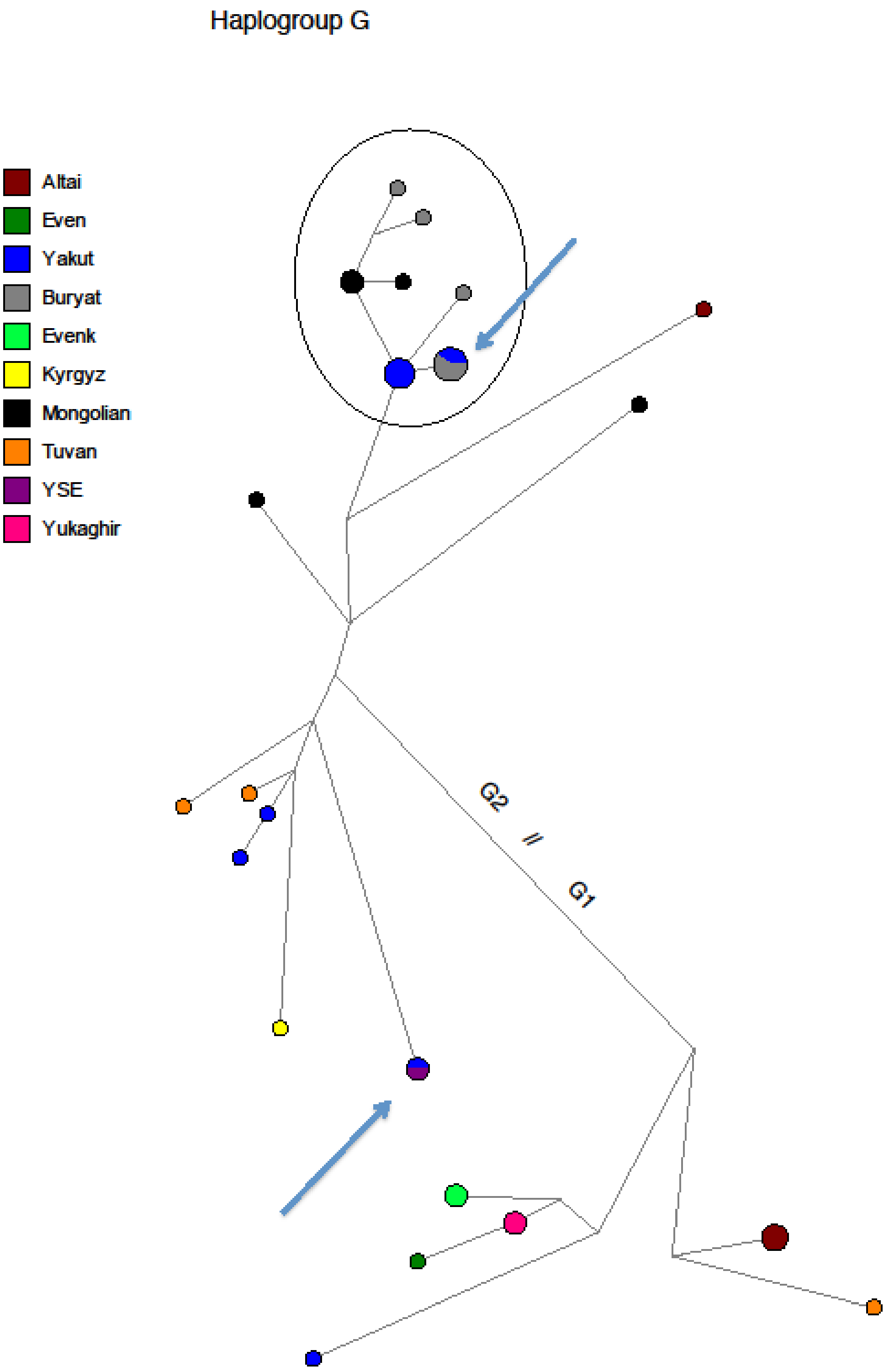

Fig. 4.10. Phylogenetic network of haplogroup G. Nodes mentioned in the text are marked with an arrow. The shortest distance between nodes corresponds to 1 mutation.

Figure 4.11 shows the haplogroup A/B network that also exhibits sharing in only two nodes. These are part of sub-haplogroup B4b and include a node with a Yakut and Altai and another node with a Yakut and three Buryats, again only one mutational step away from a Mongolian and two steps from an Altai.

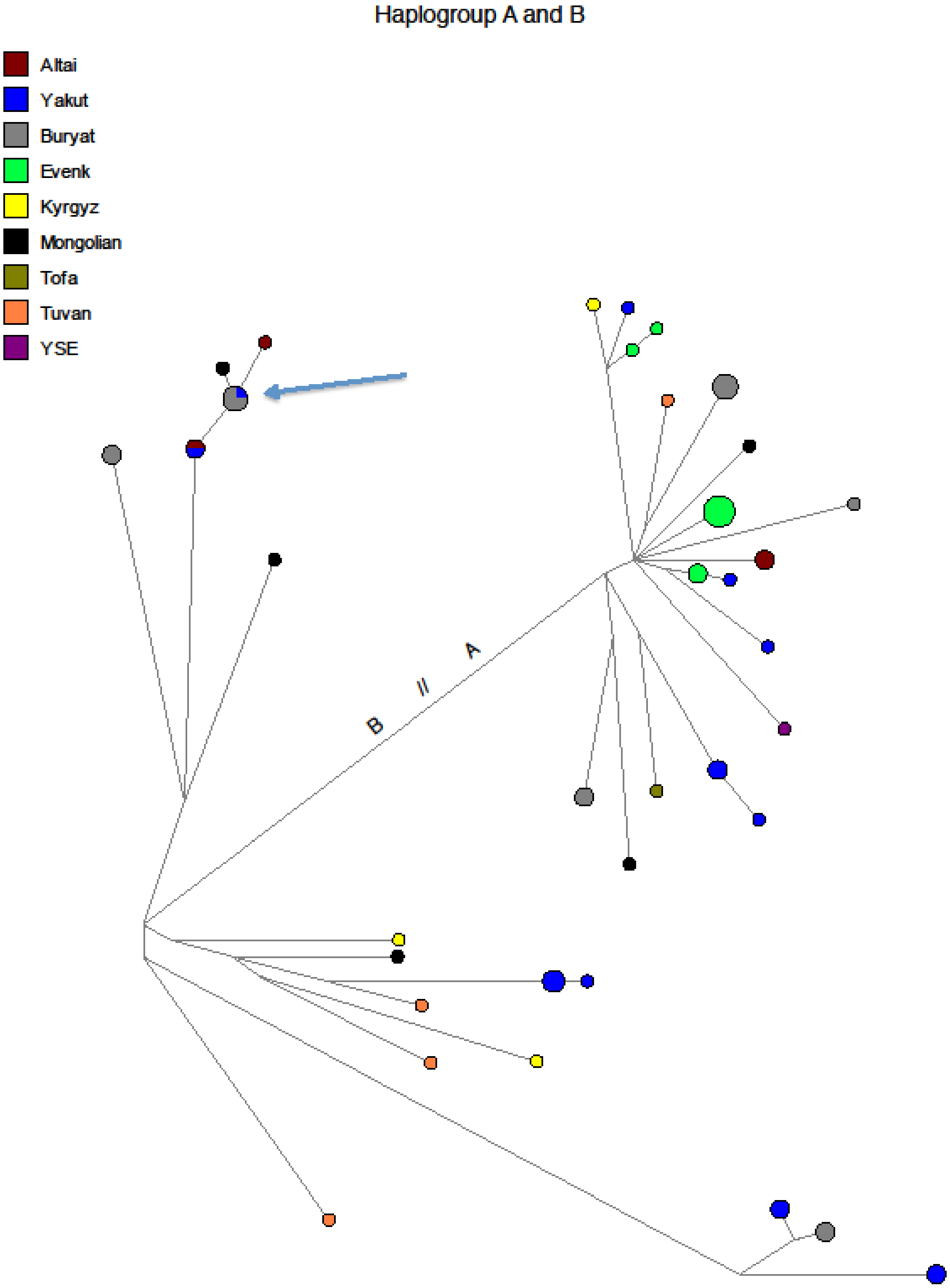

Fig. 4.11. Phylogenetic network of haplogroups A and B. Nodes mentioned in the text are marked with an arrow. The shortest distance between nodes corresponds to 1 mutation.

The network for haplogroup C in Figure 4.12 shows the largest amount of sharing throughout the entire network between all populations, though most predominately between Turkic and Tungusic speakers. While most clusters show a broad amount of sharing, there are some in the bottom right half that are predominately Yakut and the southern Siberian populations (Altai and Buryat).

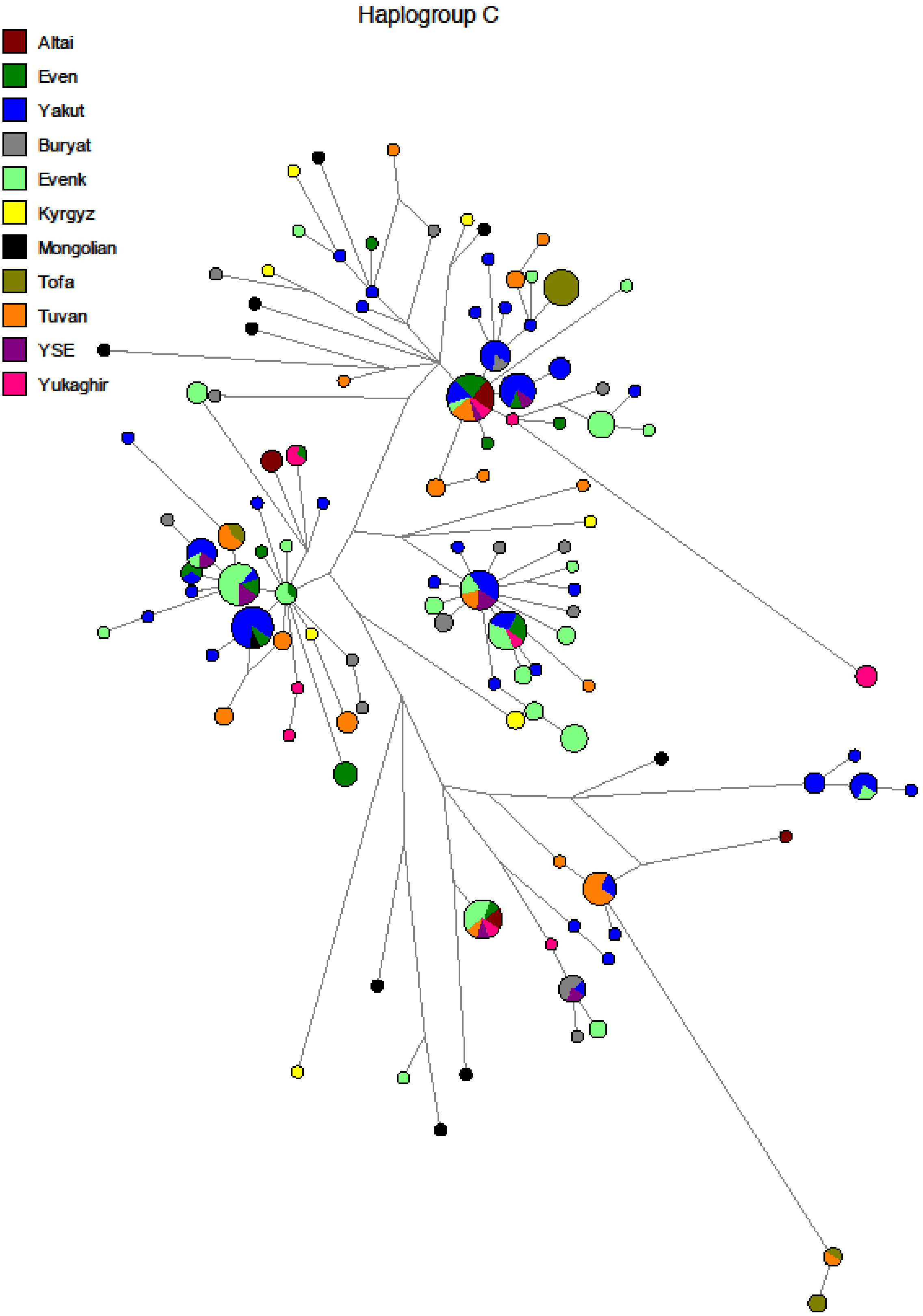

Fig. 4.12. Phylogenetic network of haplogroup C. The shortest distance between nodes corresponds to 1 mutation.

The Network for haplogroup D has many nodes shared between populations. Sub-haplogroup D5a stands out as being the largest node with a number of nodes branching off and is dominated by Yakut sequences with many other populations represented at smaller frequencies. D4l is notable for the sharing occurring only between Turkic and Tungusic speakers, with one node of Yukaghir two mutational steps away from an Even node. In the D4i sub-haplogroup, the Yakuts and Tungusic speakers only share haplotypes with the Buryats. The position of these samples close to the root of the network suggests that this could be indicative of ancestral sharing rather than recent contact as is likely to be the case in haplogroup D5a.

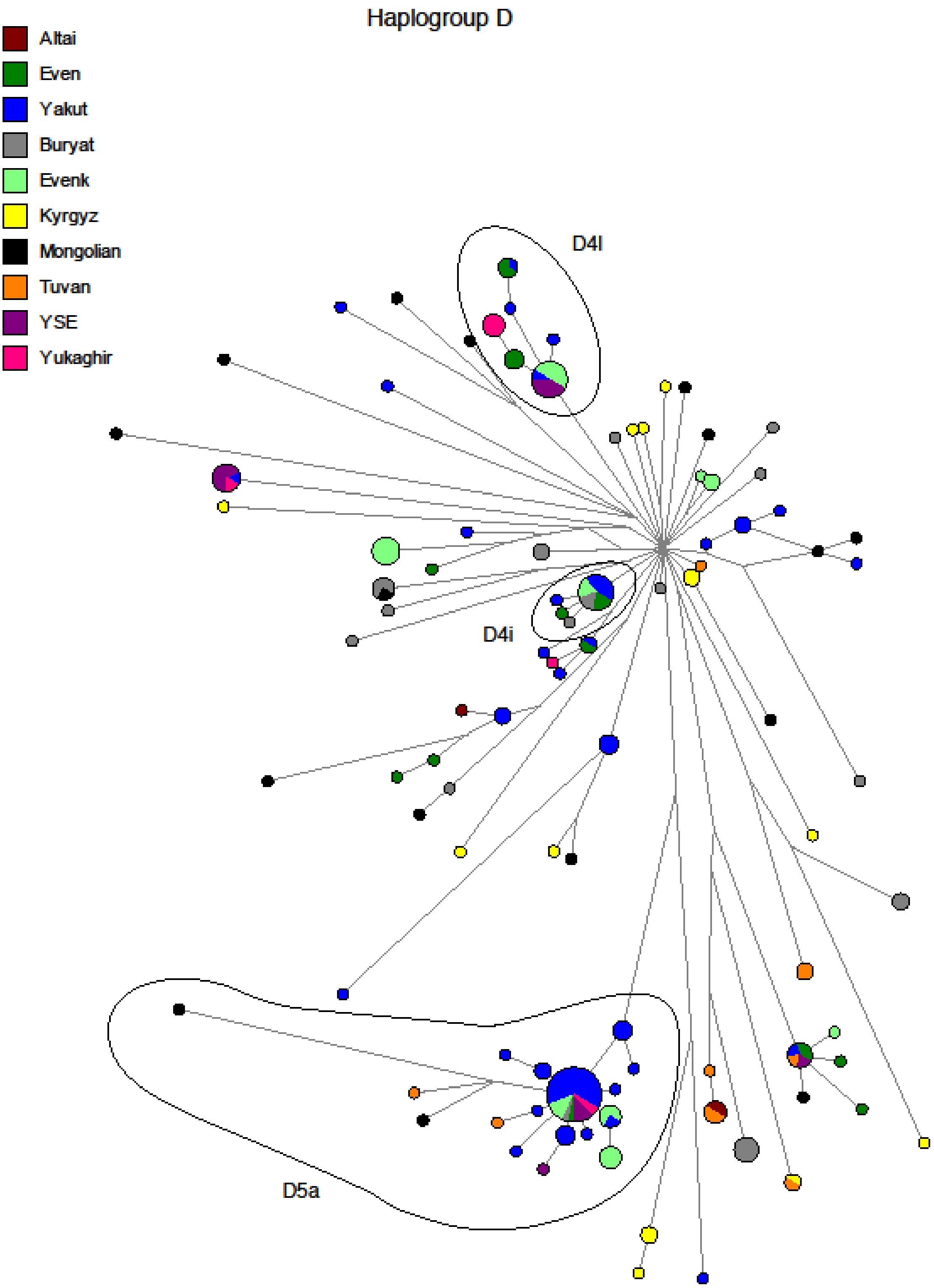

Fig. 4.13. Phylogenetic network of haplogroup D. Sub-haplogroups discussed in the text are circled. The shortest distance between nodes corresponds to 1 mutation.

Demographic changes over time

Bayesian skyline plots (BSPs) were performed for all ethnolinguistic groups and populations and based on mutation rates as described in Chapter 2. Many of the Siberian populations show an expansion in female effective population size (N_{ef}) up until around 30,000 years ago and a subsequent levelling off. This pattern is most prominent in the Turkic-speaking and Western Buryat populations shown in Figures 4.14, 4.15, and 4.16.

There are three primary trends in the past ~10,000 years that stand out in these plots. First, the Mongolian and Kyrgyz populations show an increase and levelling out in N_{ef} and there is no observable drop in N_{ef} which is observed in most other populations (Figure 4.14A and Figure 4.15). Second, the Yakuts show a sizeable drop in N_{ef} beginning at around 5,000 to 6,000 years ago, followed by a more sizeable increase beginning around 1,000 years ago (due to multiple branching events in a short time), although this increase is largely driven by the Central Yakuts (Figure 4.14B, Figure 4.16). Third, all other Siberian populations including the Northeastern and, to a lesser extent, the Vilyuy Yakuts, display a pattern similar to that seen in the Evenks in Figure 4.14C, showing a drop in N_{ef} beginning somewhere around 5,000 to 6.000 years ago, as was seen in the overall and Central Yakuts, but without the subsequent increase in size discernible in the Yakut plot (Figure 4.15). The Evens and Evenks differed in that the latter showed a more rapid drop in the last few thousand years while the Evens exhibited a more gradual decline beginning earlier.

This relatively recent difference in changes in N_{ef} between Siberian populations and the Mongolians and Kyrgyz is very pronounced. Because of the recent, sharp increase in N_{ef} in the Yakuts, an additional BSP was performed using only individuals from haplogroups C and D to see if this pattern holds or if the recent uptick in N_{ef} could be an artefact from recent admixture with immigrants (e.g. from recent Russian expansion introducing haplogroups of European descent) which could artificially inflate $N_{ef.}$ These haplogroups were chosen because they are commonly found in northeast Asia. As can be seen in Figure 4.14D, a similar pattern of a recent decline followed by a substantial increase in N_{ef} was observed in this case. Additionally, the N_{ef} of these populations is in line with what one would expect with Mongolians being at the high end followed by the Kyrgyz and Yakuts, then the Tungusic populations in the middle with Altai and Tuvans, and Yukaghirs and Tofa at the lowest end.

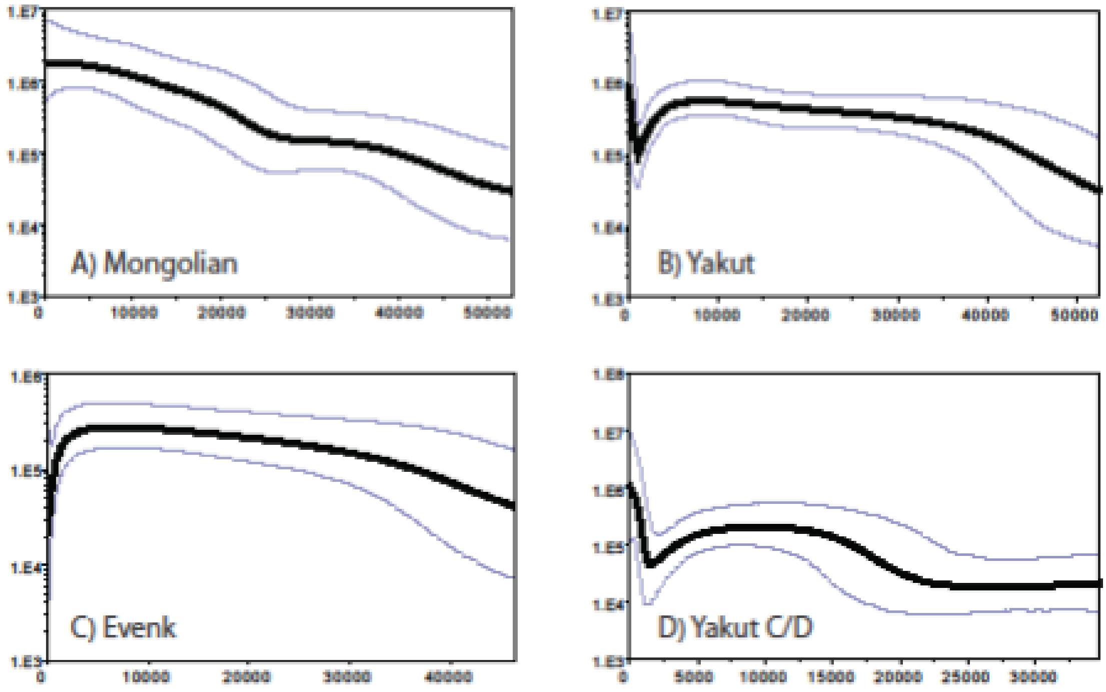

Fig. 4.14. Bayesian skyline plots showing the three main types of patterns seen amongst the populations included in this study in A, B, and C with D focusing on the primary haplogroups found in the Yakuts.

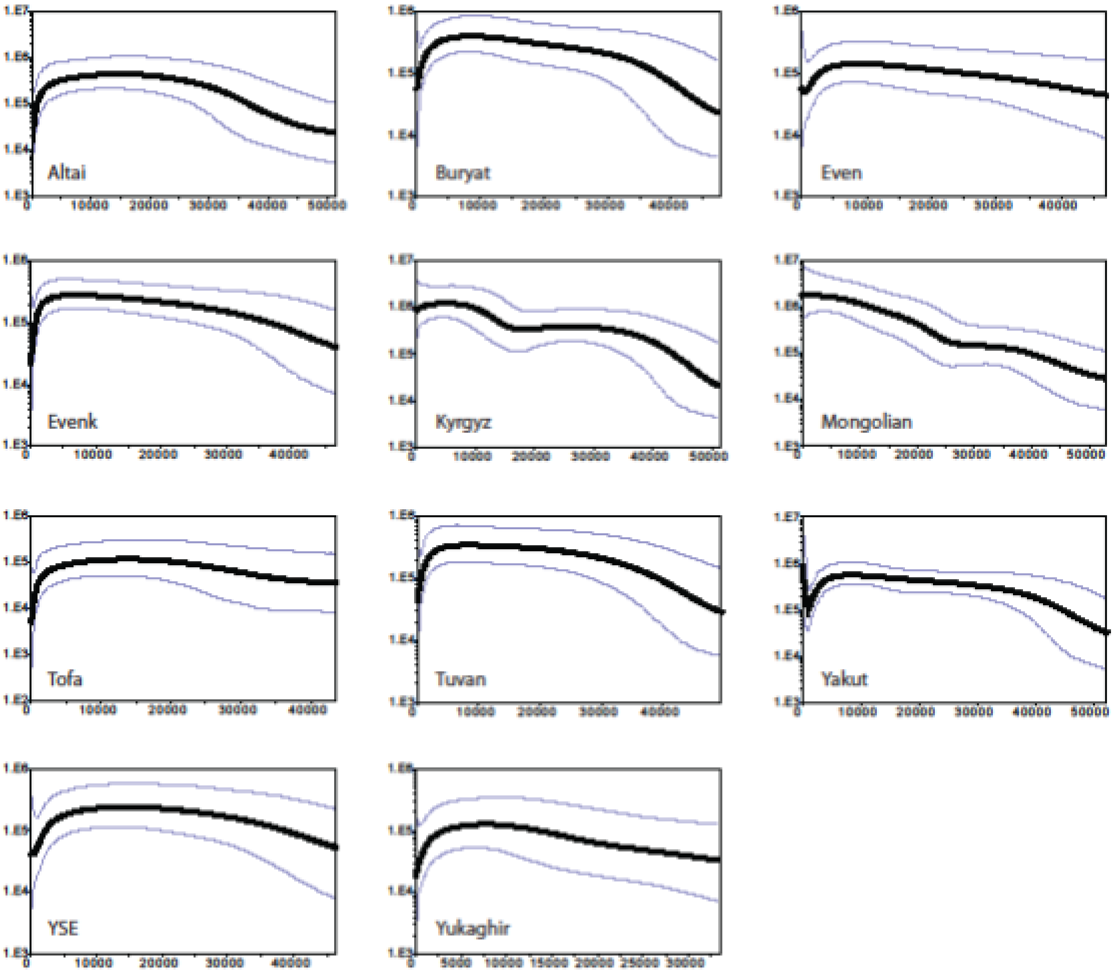

Fig. 4.15. BSPs of all ethnolinguistic groups included in this study.

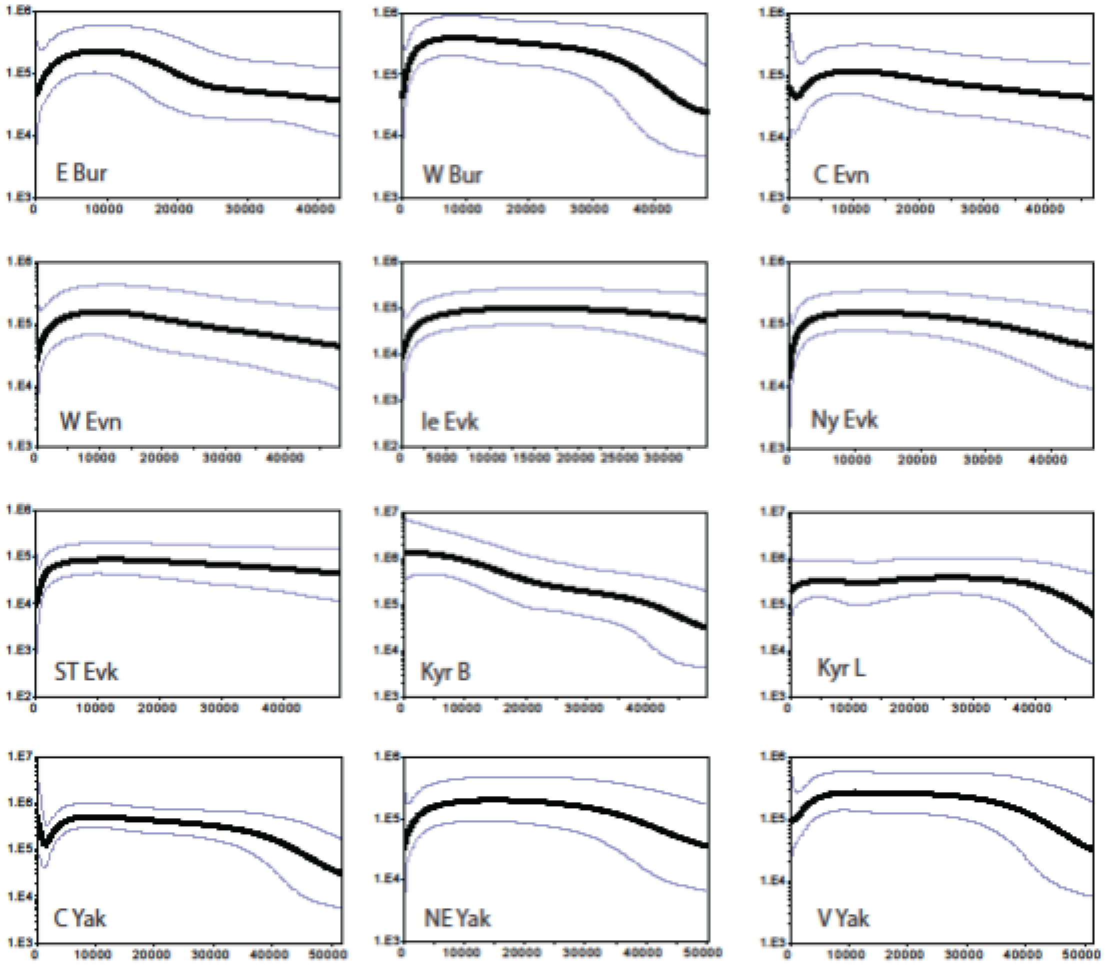

Fig. 4.16. BSPs of sub-populations.

Discussion

Genetic Structure

As has been observed in previous studies on the mtDNA of Siberian populations, the populations are relatively homogenous on the maternal side (Fedorova et al., 2003; Pakendorf et al., 2006; Pakendorf et al., 2007; Zlojutro et al., 2008). Investigations into the haplogroup structure of the Siberian populations included in this study show a high frequency of haplogroups C and D, which is most pronounced among the northeastern populations. The correspondence analysis (Figure 4.2) shows that these Siberian populations, with the exception of the Altai and Buryats, group on one side of the plot with little discernible structure. However, the benefits of switching to complete mtDNA genome sequencing will not be found in analyses based on haplogroups.

The Mongolians and Kyrgyz are shown to have the highest diversity values, which is in line with their relative population expansions and contact, while the Tofa have the lowest, pointing to a more isolated history and potential genetic drift. The Tungusic-speaking populations have lower diversity values than the Yakut. Again, this is understandable in light of the Yakut having undergone an expansion as they moved into new territories and admixed (largely on the maternal side) with the populations they came in contact with.

When one examines the MDS plots (Figures 4.3-4.5), few clusters stand out as representing a differentiation between populations based on either language or geography. Still, it is possible to see that the geographically different Yakut populations are genetically closer to one another in the MDS plots than are the Evenks or, as is seen in more resolution in Duggan et al. (2013), the Evens. What also stands out is the affinity between the Central and Vilyuy Yakuts and the YSE with the Eastern Buryats suggesting a southern origin to these populations and a closer relationship of the YSE to Yakut populations than to Tungusic-speaking populations.

The AMOVA shows that the variance between groups based on language or geography is very small, suggesting no clear structure, while the overall variance for the full set of 18 populations which, while quite low, is still significant at just under 5%. This is in stark contrast to what is seen on the paternal side. Studies of Y chromosomal structure among Siberians have repeatedly shown marked differences between populations, especially in the case of the Yakuts (Fedorova and Khusnutdinova 2010; Pakendorf et al., 2006; Pakendorf et al., 2007). This lack of mtDNA differentiation could be due to the effects of patrilocality, in that there is a higher degree of intermarriage on the maternal side. However, it is thought that patrilocality, while playing an important role at small scales, is not as important in explaining observed population structure at larger geographic scales and other factors, such as differences in N_e between males and females, could play a role (Heyer et al., 2012; Wilder et al., 2004). Nevertheless, the high mobility of some populations, in particular the Tungusic-speaking reindeer-herders, could allow for higher levels of gene flow between geographically distant populations and this is evident when exploring the sharing of haplotypes.

Population contact and shared ancestry

Approximate Bayesian Computation (ABC) analyses and Isolation with Migration (IM) analyses were attempted on the complete mtDNA genomes from these samples to attempt to distinguish between different population histories in terms of split dates and effective population sizes, but the posteriors were poor and no support could be found for any specific demographic parameters so these analyses have not been included. This is likely due to the prehistory of the region under study with the large amount of sequence type sharing obscuring any signals supporting modelled population histories. However, it is possible to use haplotype sharing and network analyses to shed some light on admixture between these populations. When looking at the networks, some sharing is apparent between the Yakuts and Tungusic speakers and the populations in the south. In haplogroups G, B and C, there are nodes with sharing purely between the Yakuts and southern populations. The D haplogroup contains a node in which the Buryats share with both the Yakuts and Tungusic speakers.

The sharing of complete mtDNA genome sequence types between linguistically or geographically separated populations could suggest recent admixture. Combining haplotype sharing data with the analyses of the Φ_{ST} values could allow for the differentiation between shared ancestry and recent contact. When Φst values are very low (non-significant before Bonferroni correction) and there is little to no haplotype sharing, this could suggest a common ancestry without recent admixture. Conversely, when Φ_{ST} values are very high (significant after Bonferonni correction) and there is haplotype sharing, this could mean recent admixture without a shared ancestry.

The recent increase in N_{ef} seen in the Yakuts dates to somewhere around 1,000 years ago. This is in line with previous work done on the Yakuts which describe a small founding population that migrated northward around 1,000 to 1,286ya and expanded according to mtDNA analyses (Zlojutro et al., 2009) and around 880ya using Y chromosomal data (Pakendorf et al., 2006). There is a very strong affinity between the Yakut populations based on the Φ_{ST} values. Combined with the high amount of haplotype sharing amongst these populations, this suggests a shared ancestry. In comparing the Yakut populations to southern populations, there is still an affinity based on Φ_{ST} values, but the haplotype sharing is lower. This could suggest a common ancestry in the south with limited admixture since their migration northward. The sharing of haplotypes between the Yakuts and the Altai, Tuvans and Buryats shows an affinity between the Yakuts and South Siberian populations.

Analyses of directionality point towards the sharing of haplotypes from the South Siberian populations into the Yakuts. This indicates a South Siberian origin for the Yakuts likely placed around Lake Baikal which is in agreement with previous mtDNA work discussed previously as well as genome wide autosomal SNP data (Pugach et al., 2013).

There is a larger amount of sharing between the Yakuts and the Evens and Evenks, in both directions, though skewed towards the direction from the Tungusic groups to the Yakuts. This is likely due to the Yakut men marrying Even or Evenk women as they expanded into their territories. However, as was also seen in Duggan et al. (2013), the majority of the sharing is mostly either between the Yakuts and Evens or Yakuts and Evenks rather than the same haplotypes being shared amongst all three ethnolinguistic groups. This suggests that rather than shared ancestry, we are seeing the effects of population contact when the Yakuts migrated northward and expanded and made contact with the already separated Evens and Evenks. Given that the Yakuts likely expanded in the 17th and18th centuries (Dolgikh 1960) and the data here show that the populations had likely split by then but also show Even and Evenks present together in nodes shared with South Siberian populations in both haplogroup C and D networks suggesting that any potential contact would have occurred before splitting, it is possible that a late split occurred. Unfortunately we are not able to date this split based on the mtDNA data presented here, but along with the autosomal data and Y data previously discussed (Duggan et al., 2013; Pugach et al., 2015), there is more support for the late split hypothesis proposed by Janhunen (1996).

Although the Kyrgyz populations share no haplotypes with one another, their Φ_{ST} values show them to be not genetically differentiated. Additionally, Kyrgyz B shows an affinity with the Mongolians, Western Buryats and Altai based on Φ_{ST} values, but shares no haplotypes with them. This suggests that the Kyrgyz populations included in this study share a common ancestry with some of the populations found around the region between Lake Baikal and the Altai Mountains prior to their migration to the southwest.

Decline in effective population size in Siberian populations

The most striking finding of the BSP analyses is the decline in N_{ef} beginning around 5,000 to 6,000 years ago across all Siberian populations in this study, which is absent in the Mongolians and Kyrgyz. The different pattern observed in the Kyrgyz could be explained by the presence of a much higher percentage of non-Siberian haplogroups, as can be seen in Figure 4.1 and Table 4.2, presumably coming from admixture with Central Asians after they migrated to where they currently live. Due to their lack of a putative decline in $N_{ef,}$ the ancestors of the Mongolians could have been further south around the time of the apparent decline.

Around 7000 to 6000 years ago, the region around Lake Baikal and the Altai Mountains underwent a shift in climatic and environmental conditions from cooler and wetter to warmer and drier (White and Bush 2010). This change is posited to have affected resource availability, causing existing sedentary peoples to become more mobile, possibly adopting a nomadic hunter-gatherer lifestyle in which larger population sizes are less likely to be sustained. The drop in N_{ef} that is observed in the BSPs could be a signal of a series of bottlenecks as previously sedentary populations split off and became nomadic. Research into the archaeology of the Baikal region has shown that at just under 7000 years ago, a period in which no burial sites are found began and continued for around 1000 years (Weber et al., 2010). After this break in burial sites, burials became common again, though with dissimilar mortuary traditions and a genetic discontinuity. The correlation of the timing of this hiatus in burial sites with the drop in N_{ef} seen in the BSPs of the Siberian populations could suggest that these populations all experienced similar demographic forces. It is unknown as to whether the potential climatic changes affected the entirety of Siberia or only the region around Lake Baikal. If it was the latter, then it is possible that the ancestors of the Siberian populations included in this study were around the same region during this time and only later repopulated Siberia. In regard to the Tungusic populations included here, this proposed origin around Lake Baikal is at odds with the autosomal data as seen in Pugach et al. (2015) so perhaps this phenomenon was more widespread.

Conclusions

The observed maternal homogeneity in these data, combined with the consistent drop in N_{ef} seen in all Siberian populations in the last 6,000 years, the archaeological evidence of a lack of burial sites around Lake Baikal, and the evidence for the origin of the Yakuts around Lake Baikal prior to migrating north (Alekseev 1996; Janhunen 1996; Vasilevich 1969), suggests a common history and a connection with this region. Along with the suggestion of a common origin for some of these Siberian populations, the haplotype sharing analysis, performed on the complete mtDNA genomes, suggests relatively recent admixture for some of these populations, especially in the northeast. This is supported by the differential sharing seen between the Yakuts and Evens and Yakuts and Evenks. Such a pattern of sharing would not be expected to have occurred farther in the past, prior to these populations migrating north.

Chapter 5

Origins of the Dolgans

Origins of the Dolgan

Introduction

The Dolgans are a recently formed ethnolinguistic group, who potentially developed out of the admixture of the Yakuts, Evenks, Samoyedic-speakers and, to a lesser extent, Russians (Dolgikh, 1963 as cited in Ziker, 1998, and Stapert 2013). When the Russians arrived in the 17th century, there were no Turkic-speaking groups in this area, and no people calling themselves Dolgan (Stapert, 2013).It has not clear, however, to what degree these different groups contributed genetically to the formation of the Dolgan peoples and when this formation occurred from analyses on uniparental markers.

The Dolgans speak the most northerly Turkic language, which is a dialect of the Yakut language, and currently live both on the Taimyr Peninsula in the Taimyr Municipal District and just below and to the east in the Anabar district of the Republic of Yakutia. They are relative newcomers to these areas. Original occupation of the Taimyr dates back to 7000 years and it is thought that the original inhabitants of this region were relatives of Yukaghirs, and other hunter-gatherer paleo Siberians currently found in the northwest (Ziker, 1998). It has been proposed that later migrations into this region include that of Samoyedic-speaking populations (in the 2nd to 4th centuries) followed by Tungusic speakers (Ziker, 1998). Finally, Russians arrived in the 17th century. These waves of migrations of peoples speaking different languages is reflected in the populations currently inhabiting this region including the Samoyedic-speakers (Nenets, Enets, and Nganasan), Turkic-speaking Dolgans, and Tungusic-speaking Evenks.

Although the language of the Dolgans is Turkic and considered a dialect of Yakut, they utilize reindeer herding, a mode of subsistence more similar to the Evenks than to the horse and cattle pastoralism of the Yakuts. Additionally, the name ‘Dolgan’ is a Tungusic clan name, rather than Turkic (Stapert, 2013). Their current identity is thought to have formed from admixture largely between Tungusic reindeer-herding clans and Yakut traders with additional input from Samoyedic groups and Russians (Dolgikh, 1963 as cited in in Ziker, 1998).

Although uniparental markers from Dolgan populations have been analyzed in molecular anthropology research, most recently in Federova et al. (2013), investigating their origins was not a primary focus and in terms of exploring maternal population histories only HV1 has been used. However, it was enough to show expansions to the north from southern Siberia of the Turkic-speaking and Tungusic speaking populations. Additionally, previous studies have relied heavily on haplogroup lineage analyses without a lack of population-wide comparisons similar to those shown in Chapter 4.

Federova et al. (2013) also contains genome wide autosomal SNP data (>500,000 SNPs), but again it does not explore the issue of Dolgan formation or contact with neighboring populations. A subsequent publication on genome-wide autosomal SNP data (Pugach et al., 2015) utilized published data in combination with newly generated data to look at the origins and admixture of Siberian populations, including that of the Dolgans. These analyses show an affinity to south Siberian populations suggesting that at least a portion of the Dolgan ancestors come from the Yakuts who themselves can be traced back to a region around Lake Baikal (Pugach et al. 2015) and confirm that the closest relationships of the Dolgans are to the Yakuts and

Evenks with some limited “European-like” recent input and potential, but undetermined Samoyedic addition.

Given the historic, linguistic and genetic data, there three likely scenarios. First, that the ancestors of the Dolgans were Yakuts that migrated northward and changed their lifestyle. Second, their ancestors were Evenks who underwent a language shift to that of the Yakuts. Some of their linguistic features point towards just such a shift (Stapert, 2013). However, it is more likely that there was a combination of these two situations with majority input from the Yakut and Evenk and minor input from Samoyedic speakers.

In order to better understand the origins of the Dolgan it is necessary to look more closely into the uniparental markers so complete mtDNA genomes were sequenced and are analyzed below from two geographically separated Dolgan populations. Comparative populations of neighboring Taimyr Evenks from Duggan et al. (2013) and Samoyedic-speaking Nenets from both the Taimyr and Yamal (further west) Peninsulas were also included in combination with the data from populations in Chapter 4. The focus of this chapter is on those analyses shown in Chapters 3 and 4 to be beneficial for exploring maternal contact and histories using complete mtDNA genomes

Materials

In addition to the samples from Chapter 4, 149 samples were analyzed in this study. The Dolgan samples were collected from villages in both the Taimyr Peninsula and the Anabar district in the Republic of Yakutia. The Nenets samples were collected from the Taimyr and Yamal Peninsulas.

The additional Evenk population was collected from the Taimyr Peninsula and was first published in Duggan et al. (2013). The remaining populations of Turkic, Tungusic, and Mongolic speakers and the linguistic isolate, Yukaghir, used for comparison in the following analyses were described in Chapter 4 with sample sizes and affiliations given in Table 5.1. Approximate sampling locations of the new populations are shown in Figure 5.1.

Table 5.1. Linguistic affiliation and sample sizes of the populations included in this study. The samples from five populations not included in Chapter 4 are in italics.

Population	**Ethnolinguistic Affiliation**	**Acronym**	**n**	**Linguistic Affiliation**
Mongolian	Mongolian	Mongol	56	Mongolic
Eastern Buryat	Buryat	E_Bur	27	Mongolic
Western Buryat	Buryat	W_Bur	48	Mongolic
Kyrgyz B	Kyrgyz	Kyr_B	30	Turkic
Kyrgyz L	Kyrgyz	Kyr_L	27	Turkic
Altai	Altai	Altai	33	Turkic
Tofalar	Tofalar	Tofa	23	Turkic
Tuvan	Tuvan	Tuvan	59	Turkic
Central Yakut	Yakut	C_Yak	89	Turkic
CEPH Yakut	Yakut	C_Yak	24	Turkic
Vilyuy Yakut	Yakut	V_Yak	55	Turkic
Northeastern Yakut	Yakut	NE_Yak	33	Turkic
Yakut-speaking Evenk	Yakut-speaking Evenk	YSE	32	Turkic
Anabar Dolgan	*Dolgan*	*Ana_Dol*	*27*	*Turkic*
Taimyr Dolgan	*Dolgan*	*Ta_Dol*	*51*	*Turkic*
Taimyr Nenets	*Nenets*	*Ta_Nen*	*13*	*Samoyedic*
Yamal Nenets	*Nenets*	*Ya_Nen*	*34*	*Samoyedic*
Taimyr Evenk	*Evenk*	*Ta_Evk*	*24*	*Tungusic*
Stony Tunguska Evenk	Evenk	ST_Evk	39	Tungusic
Nyukzha Evenk	Evenk	Ny_Evk	45	Tungusic
Iengra Evenk	Evenk	Ie_Evk	23	Tungusic
Central Even	Even	C_Evn	26	Tungusic
Western Even	Even	W_Evn	24	Tungusic
Yukaghir	Yukaghir	Yukag	22	Isolate

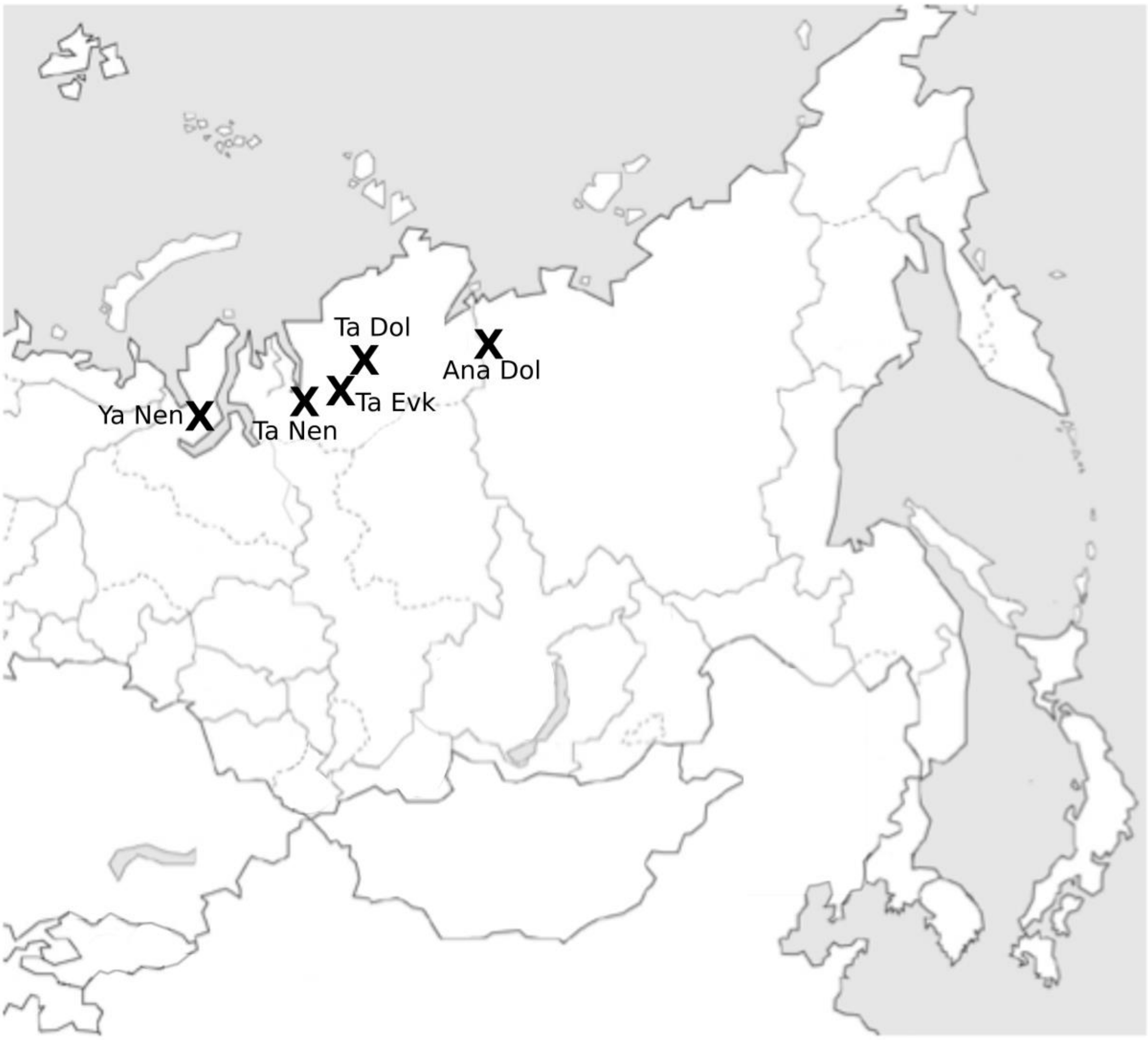

Fig. 5.1. Map showing approximate sampling locations for those populations not included in Chapter 4.

Results

Figure 5.2 shows a map of the populations included in this study and their respective haplogroup frequencies. The Dolgans are predominately haplogroup C and D (63.0% in Anabar Dolgans and 72.5% in Taimyr Dolgans) and have a much lower amount of the D5a sub-haplogroup (only in the Taimyr Dolgans at 5.9%) typical of most Yakut populations and Iengra Evenks. The Taimyr Evenk have a similar pattern to the Dolgans, in terms of their shares of haplogroups C and D though, as is the case in the other Evenk populations (except the Iengra Evenks), there is a larger share of Haplogroup C over D (45.8% to 33.3%). The Dolgans have a more even split at 34.6% to 30.8% for haplogroups C and D overall and the sub-populations show a similar pattern.

The Samoyedic-speaking Nenets combined from both the Taimyr and Yamal Peninsulas have a much lower percentage of haplogroup D (8.8%) than is found in Tungusic populations. Conversely, they have a relatively high percentage of other haplogroups not typical of Central and Northeastern Siberian populations, but rather indicative of Eurasian population influence. Haplogroups H, J and U are present at 14.9%, 6.4%, and 19.1%, respectively, in the Nenets and while U is relatively similar in both Nenets sub-populations, haplogroups H and J are only found in the Yamal Nenets (the population found further to the west).

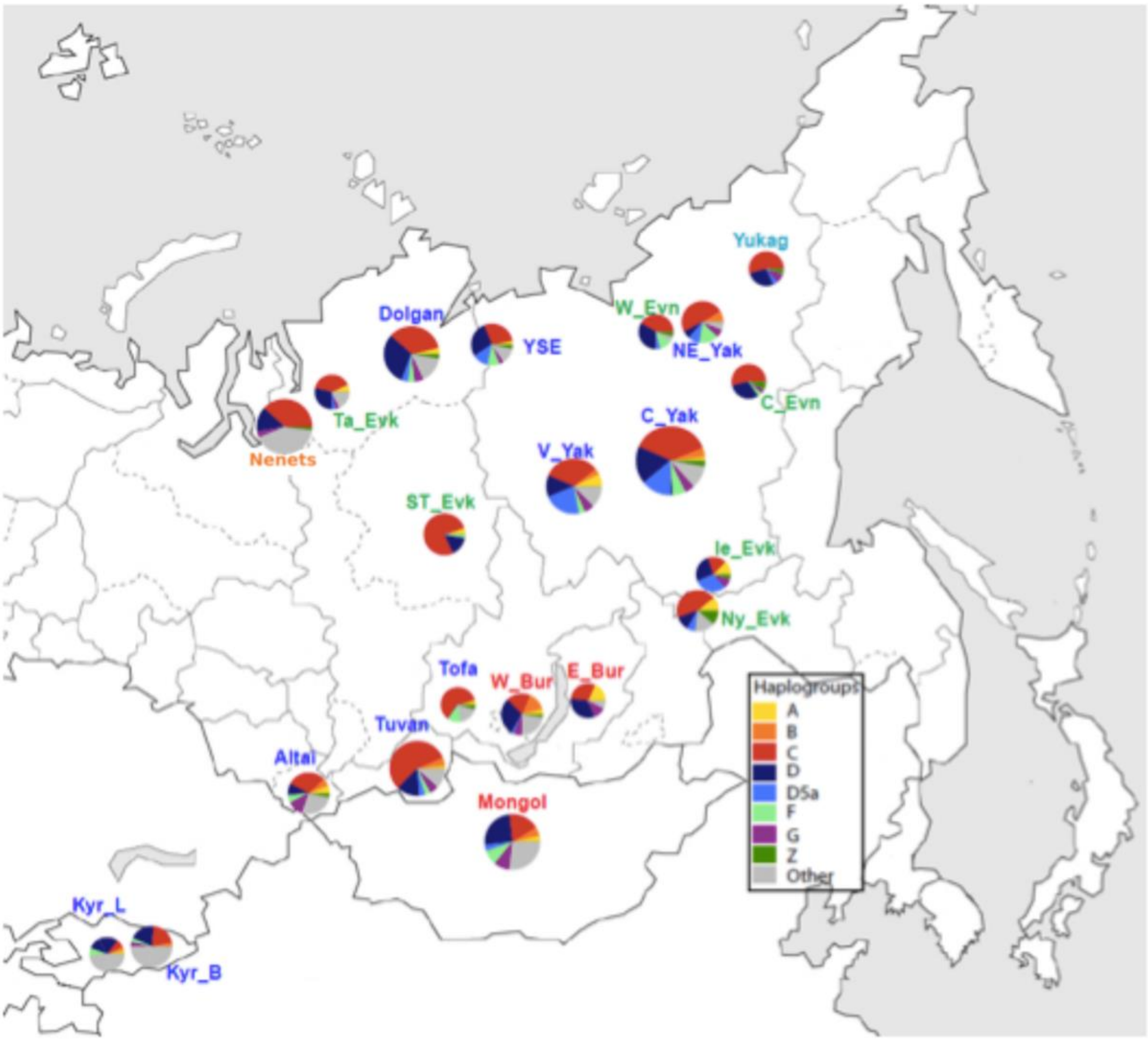

Fig. 5.2. Map showing locations of populations and major mtDNA haplogroup frequencies along with relative sampling sizes corresponding to the size of the pie charts. Linguistic affiliations are shown by population name labels (Turkic, blue; Tungusic, green; Yukaghir, aqua; Mongolic, red; Samoyedic, orange).

As there was no differentiation between the Nenets sub-populations of Taimyr and Yamal based on Φ_{ST} analyses, these groups were combined to form one Samoyedic-speaking Nenets population for this analysis. Additionally, based on Φ_{ST} analyses, the Kyrgyz sub-populations were grouped together. Because the focus here is on the Taimyr, for the majority of the analyses the Even sub-populations and the Yakut sub-populations were grouped respectively. In Figure 5.3, the Dolgan sub-populations are most closely positioned next to their nearest geographic neighbors. The Anabar Dolgans are geographically very close to the YSE and the Taimyr Dolgans are geographically most close to the Taimyr Evenks which is reflective of what is seen in the MDS plot. The Dolgans, in particular those from the Taimyr are closer to the Yakut in the first dimension. However, the predominant pattern is rather one of geographical proximity. The Nenets are quite differentiated from their geographic neighbors. And, as seen in Chapter 4, the Tofa, ST Evenks, and Iengra Evenks fall out of the main cluster.

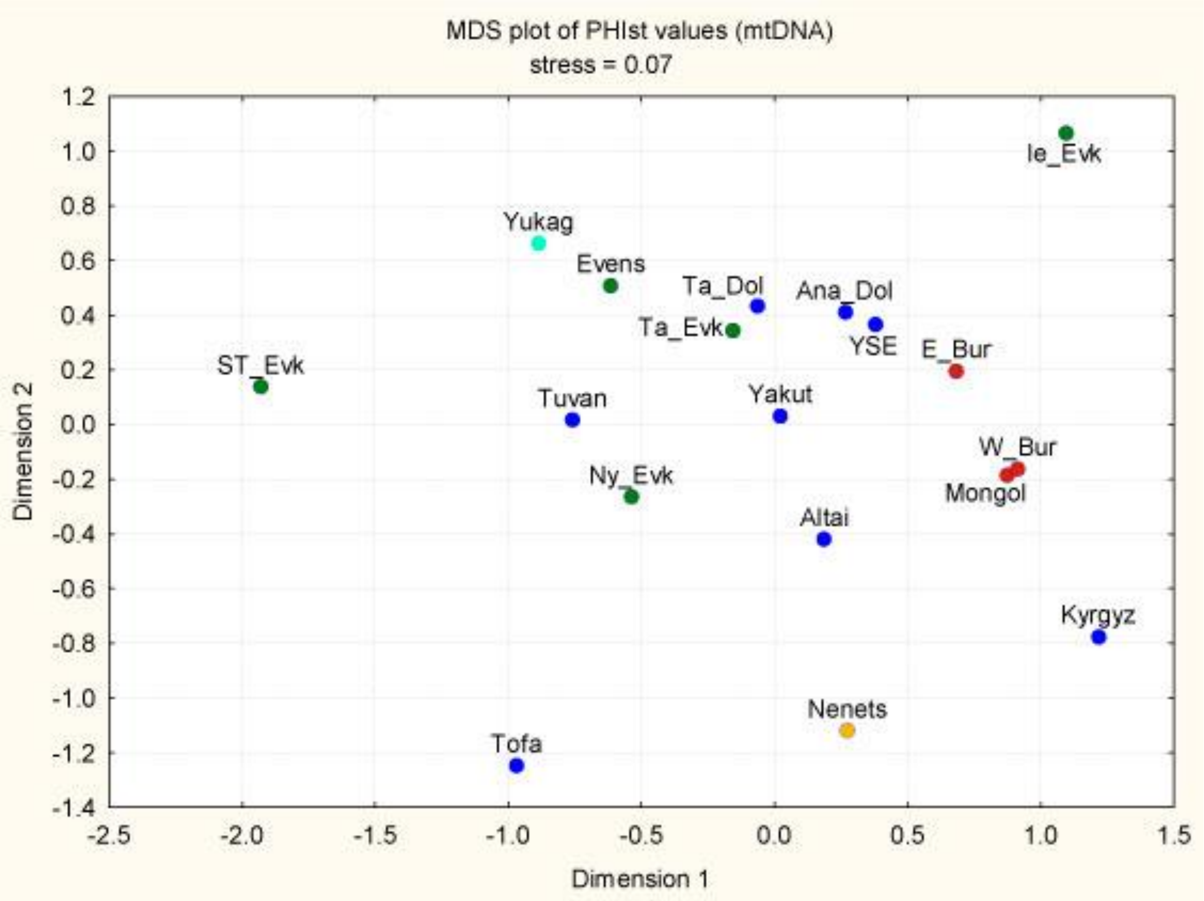

Fig. 5.3. A 2D MDS plot based on pairwise Φst values for dimensions 1 versus 2.

An analysis of haplotype sharing shows some obvious grouping based on geography (Figure 5.4). Because this type of analysis can be informative for questions on recent admixture, the Nenets sub-populations were treated separately here. The pairs of populations with the highest amounts of haplotype sharing are the Taimyr Dolgans / YSE, and the Taimyr Evenks / YSE. This is the highest level of complete mtDNA haplotype sharing seen between any Siberian ethnolinguistic populations examined to date. It is notable that haplotype sharing is greater between the Taimyr Evenk and their geographic neighbors than between them and any other Evenk sub-population. These highest levels of sharing are closely followed by the Anabar Dolgans / YSE and the Taimyr Evenks / Taimyr Dolgans. These four populations (both Dolgan populations, Taimyr Evenks, and YSE) group closely in the MDS plot and share the most haplotypes with one another.

Both Dolgan groups share haplotypes with all Evenk populations, though at a lower level than with their geographic neighbors. There is also a geographic pattern apparent in the plot whereby the Taimyr Dolgans exhibit more sharing with the Taimyr Evenks than with the other Evenks and even with the Anabar Dolgans. Similarly, the Anabar Dolgans share the largest amount of haplotypes with the YSE, The Taimyr Dolgans also share more with Yakuts than they do with the non-Taimyr Evenks. The Anabar Dolgans, however, exhibit less sharing overall but do show more sharing with the Evenks than they do with the Yakuts and, interestingly, they share as much with the Nyukzha and Iengra Evenks as they do with the Taimyr Evenks.

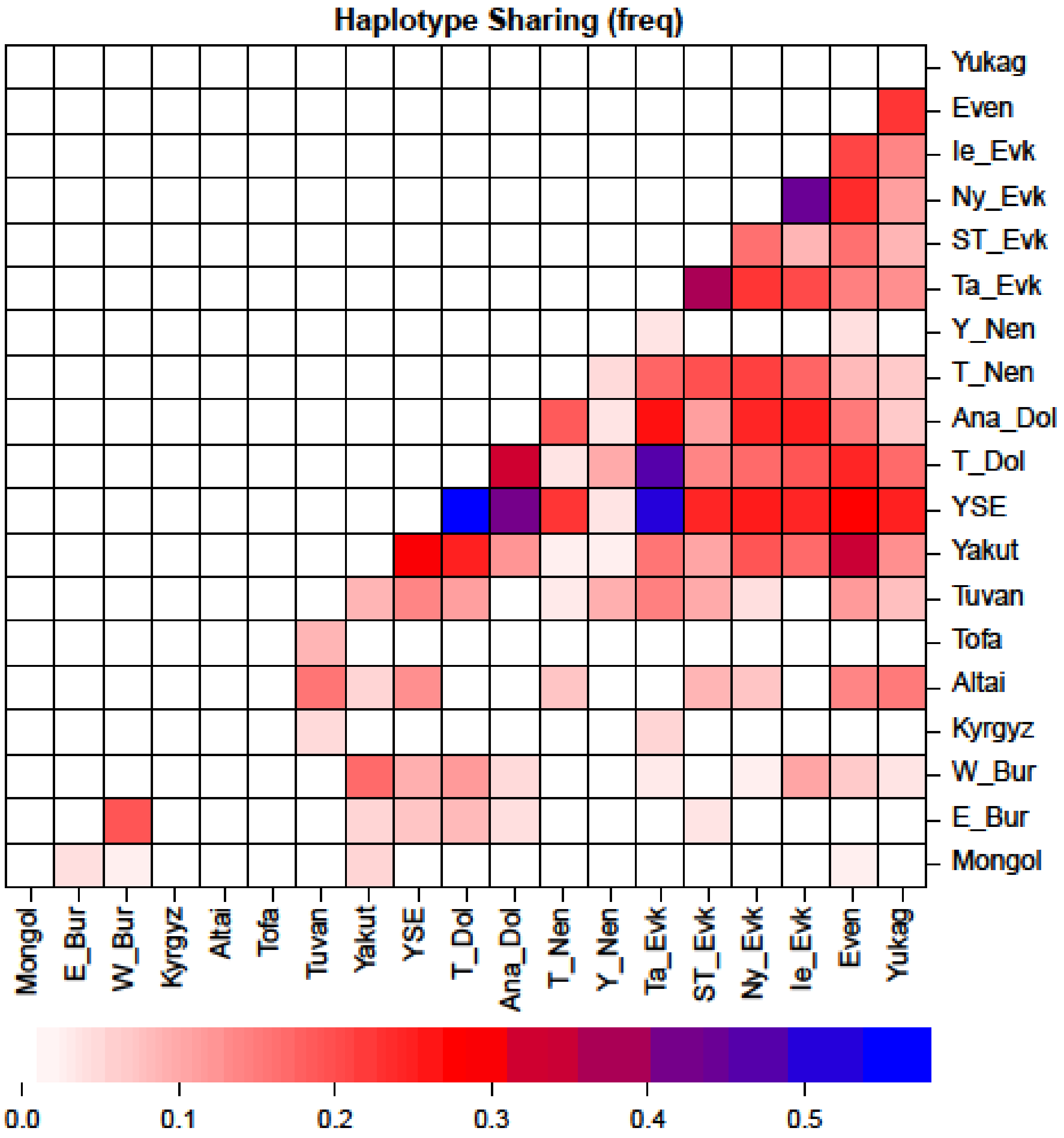

Fig. 5.4. Haplotype sharing heat plot.

The Yamal Nenets exhibit a very low amount of haplotype sharing with any population, though there is some with Taimyr Dolgans and Tuvans. The Taimyr Nenets, on the other hand, share a moderate amount with all Evenk populations including the YSE and also with the Anabar Dolgans. The Taimyr Nenets also share a small amount with the Altai while the Yamal share nothing with the southern populations. Neither share any haplotypes with the Yakuts. This pattern of primary haplotype sharing based on geography between the Taimyr Nenets and Tungusic populations suggests recent admixture from the Tungusic populations into the Taimyr Nenets.

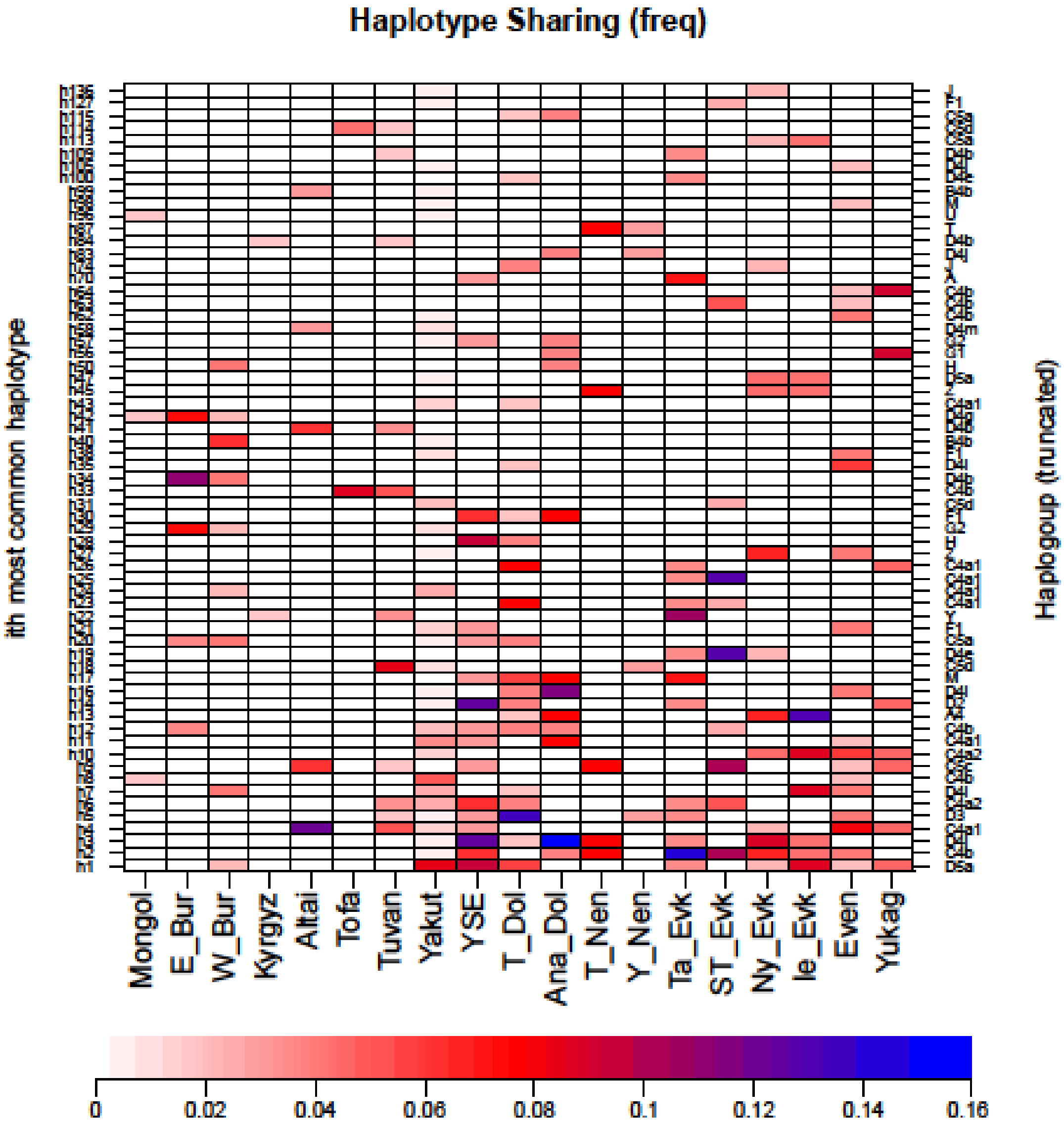

Fig. 5.5. Heat plot of haplotype sharing between at least two populations, decreasing in frequency of sharing from the bottom to the top of the plot. Haplogroups provided on the right side of the figure.

The haplotype sharing analysis shown in Figure 5.4 can be broken down into specific haplotypes shared by multiple populations (Figure 5.5). One noticeable feature of this graph is the high amount of sharing in the YSE as well as the Dolgans and Taimyr Evenks. To explore potential differential sharing, the haplotype frequencies underlying the heat plot shown in Figure 5.5 are provided in Table 5.2. In total the Dolgans share 26 haplotypes with other populations. Of those, only 1 haplotype is shared exclusively between the Anabar and Taimyr Dolgans (h115) and 6 others are shared between these sub-populations and at least one other group (h3, h12, h13, h16, h17, h30). The most commonly shared of these haplotypes is h3 which has the highest frequencies in the Anabar Dolgans and YSE (15.38% and 12.5%, respectively) with very low amounts in the Taimyr Dolgan and Evenk and the lowest in the Yakut (only 0.5%). Given intermediate levels in the Nyukzha Evenks and Iengra Evenks (8.89% and 4.35%) this is likely originating from Tungusic populations who then took the Yakut/Dolgan language and stayed in this region, though having some contact with other Taimyr populations. There are three haplotypes that Dolgans share with Evenks (excluding the Taimyr Evenks) and seven that are shared between YSE and Dolgans but not the Taimyr. The remaining Dolgan haplotypes shared by both sub-populations and at least one other show a mixture of sharing with Yakuts and Evenks but in all cases, the Anabar Dolgans have a higher frequency.

Of the sharing that occurs between south Siberian populations and the Dolgan, this is largely biased towards the Taimyr population. Two haplotypes (h5, h6) are shared between the Tuvans, the Taimyr Dolgans and Taimyr Evenks as well as the Yakuts and YSE, with h6 also being shared with the ST Evenks and h5 being shared with the Yamal Nenets. H5 is the haplotype with the highest frequency in the Taimyr

Dolgans so this lineage could have spread as the Dolgan population expanded. The Buryats also share some haplotypes with the Dolgans. Primarily this is with the Taimyr Dolgans (3 haplotypes which are also shared with the Yakuts) but also 1 shared with both Dolgan populations and another with only the Anabar.

The Taimyr Evenks most frequent haplotype (h2, 14.29%) is shared with all other Tungusic-speaking populations as well as the Taimyr Nenets and YSE as well as one Yakut. Their next highest frequency haplotype (h22) is only shared with the Tuvan and Kyrgyz.

The Nenets sub-populations only share one haplotype (h87) with each other and with the other seven haplotypes that are shared by one Nenets population and another group, there is not a distinct pattern but there is sharing with the south Siberian Altai and/or Tuvan in three haplotypes. While there is very little signal of admixture with the Yakuts, there is sharing with multiple Tungusic-speaking groups.

Table 5.2. Percentages of haplotypes in the populations

Haplotypes	Mongol	E_Bur	W_Bur	Kyrgyz	Altai	Tofa	Tuvan	Yakut	YSE	T_Dol	Ana_Dol	T_Nen	Y_Nen	Ta_Evk	ST_Evk	Ny_Evk	Ie_Evk	Even	Yukag
h136								0.50								2.22			
h127								0.50							2.56				
h115										1.92	3.85								
h114						4.35	1.69												
h113																2.22	4.35		
h109							1.69							3.57					
h105								0.50										2.00	
h100										1.92				3.57					
h99					3.03			0.50											
h98								0.50										2.00	
h96	1.79							0.50											
h87												7.69	2.94						
h84				1.75			1.69												
h83											3.85		2.94						
h74										3.85						2.22			
h70									3.13					7.14					
h64																		2.00	9.09
h63															5.13			2.00	
h62								0.50										4.00	
h58					3.03			1.00											
h57								0.50	3.13		3.85								
h56											3.85								9.09
h50			4.17								3.85								
h47								0.50								4.44	4.35		
h45												7.69				4.44	4.35		
h43								1.49		1.92									
h42	1.79	7.41	2.08																
h41					6.06		3.39												
h40			6.25					0.50											
h38								1.00										4.00	
h35										1.92								6.00	
h34		11.11	4.17																
h33						8.70	5.08												
h31								1.99							2.56				
h30									6.25	1.92	7.69								
h29		7.41	2.08					1.00											
h28									9.38	3.85									
h27								0.50								6.67		4.00	
h26										7.69				3.57					4.55
h25														3.57	12.82				
h24			2.08					2.49											
h23										7.69				3.57	2.56				
h22				1.75			3.39							10.71					
h21								1.49	3.13									4.00	
h20		3.70	4.17					0.50	3.13	3.85									
h19														3.57	12.82	2.22			
h18							8.47	1.00					2.94						
h17									3.13	5.77	7.69			7.14					
h16								0.50		3.85	11.54							4.00	
h14								0.50	12.50	3.85				3.57					4.55
h13										1.92	7.69					6.67	13.04		
h12		3.70						1.99	3.13	3.85	3.85				2.56				
h11								3.48	3.13		7.69							2.00	
h10								1.49								4.44	8.70	6.00	4.55
h9					6.06		1.69		3.13			7.69			10.26			2.00	4.55
h8	1.79							4.98										2.00	
h7			4.17					2.49		1.92							8.70	4.00	
h6							3.39	2.49	6.25	3.85				3.57	5.13				
h5							1.69	0.50	3.13	13.46			2.94	3.57				4.00	
h4					12.12		5.08	1.49	3.13							2.22		8.00	4.55
h3								0.50	12.50	1.92	15.38	7.69		3.57		8.89	4.35		
h2								0.50	6.25		3.85	7.69		14.29	10.26	6.67	4.35	4.00	
h1			2.08					8.46	9.38	5.77				3.57		2.22	8.70	2.00	4.55
total % ht share	**5.37**	**33.33**	**31.25**	**3.50**	**30.30**	**13.05**	**37.25**	**46.83**	**93.81**	**82.69**	**84.63**	**38.45**	**11.76**	**78.55**	**66.66**	**55.54**	**60.89**	**68.00**	**45.48**

The BSPs for the new populations included in this chapter are shown in Figure 5.6. Within the past 5,000 years, there began a steep decline in the effective population sizes in all populations which was very steep in the past few thousand years. The Nenets both share a similar 'bump' prior starting before 10,000 years ago prior to the potential decline not present in the other Siberian populations. Otherwise these graphs are similar to the non-Yakut Siberian populations seen in Chapter 4 and the discussion regarding this common decline potentially suggesting similar geographic origins can be extended to the Samoyedic-speaking Nenets.

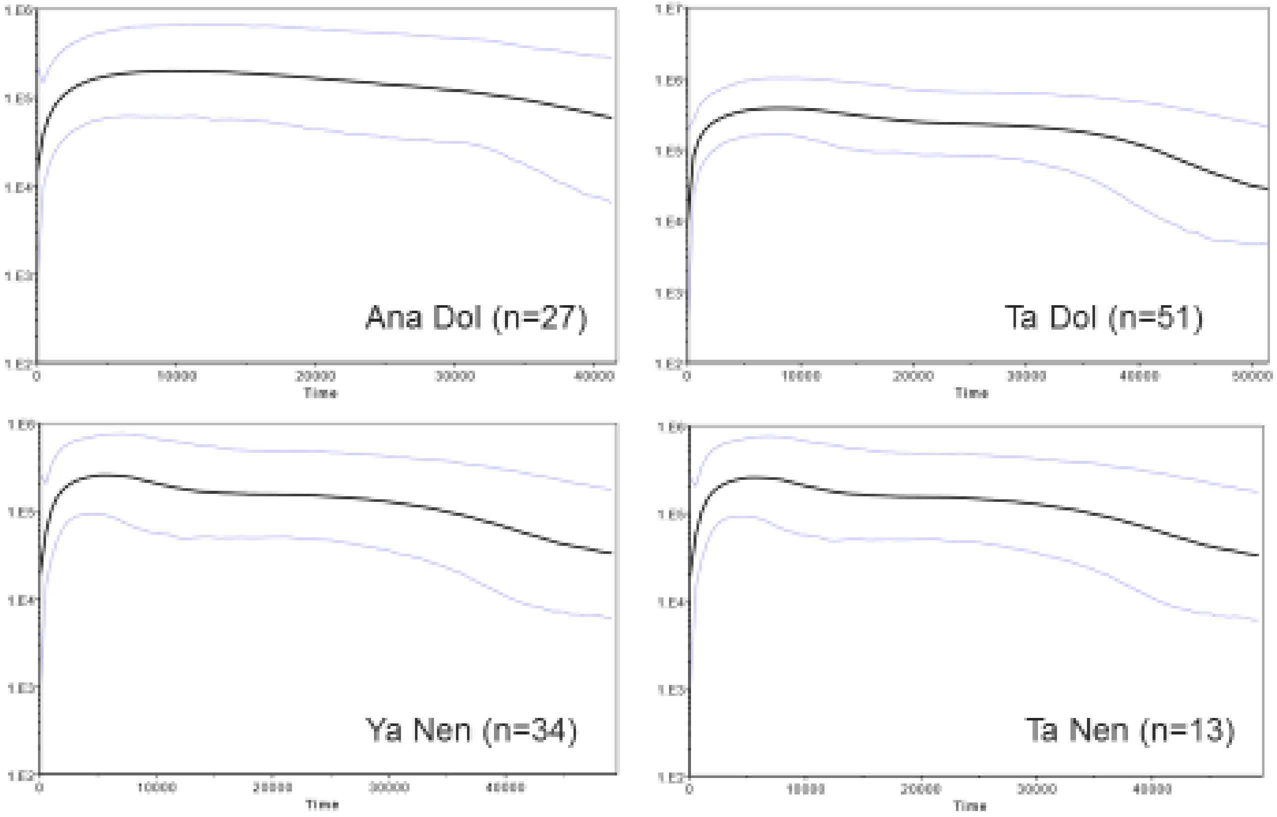

Fig. 5.6. Bayesian skyline plots showing the sub-populations of Dolgans and Nenets

Discussion

The resulting maternal picture is one of high amount of contact between the Dolgans, the Taimyr Evenks and the YSE. These populations are geographically close, being located in and southeast of the Taimyr Peninsula, but linguistically differentiated. The fact that both the mtDNA MDS plot based on Φst values and the haplotype sharing plots showed such similarities suggests that, for the populations who migrated to this region, there was a combination of both shared ancestry and recent admixture in the Taimyr. Given the very high levels of full mtDNA haplotype sharing between the Dolgans, Taimyr Evenks and the YSE as compared to within group sharing observed between the Yakut sub-population or Evenk sub-populations, this suggests a very high level of recent admixture between geographically close populations. The limited haplotype sharing between Dolgan sub-populations suggests more recent admixture with geographic neighbors than linguistic relatives.

The fact that there are multiple haplotypes present in the Dolgans and one of the non-Taimyr Evenks, especially those farther away such as the Nyukzha or Iengra (or, indeed, the Evens) suggests Tungusic input. This would likely not be due to recent admixture, rather a shared ancestry is more probable rather than a simple spreading of Taimyr Evenk haplotypes.

The sharing seen between Taimyr populations and the south Siberian populations is most likely indicative of those lineages being brought up through waves of migrations into the region rather than by recent admixture given the vast geographic distances involved. This points to lineages being carried through the Yakuts as they moved north. However, the Dolgans show more affinity to the Tungusic-speaking populations than they do to the Yakuts. There is significant input from both the Tungusic peoples and the Yakuts. Similarly in the paternal side there is this mixture when looking at haplogroups, with a relatively even split between Yakuts and Evenks, but when looking at MDS plots the Yakuts are outliers and the Dolgan fall much closer to the Tungusic populations (based on data from populations included in this chapter combined with those from Karafet et al., 2002). This is in line with the historical and linguistic evidence suggesting a complex formation process involving the ‘Dolgan’ and other Evenk clans shifting from their Tungusic language to the Turkic language, Yakut, while retaining their lifestyle and culture (Stapert, 2013).

Chapter 6

Conclusions and Future Outlook

Conclusions and Future Outlook

The overall goal of this work was to improve on methods for sequencing complete mtDNA genomes and investigate the putative benefits of using these larger datasets in the exploration of human population histories and contact. The impetus for conducting this study was the need to more closely examine the complex maternal population histories in Siberia given that the historical use of only mtDNA HV1 presented difficulties in addressing questions of maternal population contact and prehistories of the region due to high levels of HV1 sequence type sharing (Bregel, 2003, Fedorova et al., 2003, Pakendorf et al., 2006, Pakendorf et al., 2007, Zlojutro et al., 2008). At the time of this study's inception, there was a lack of thoroughly tested, and relatively inexpensive, methodologies that could exploit 2nd generation sequencing technologies for the generation of large numbers of complete mtDNA genomes. A consequence of this that I identified from my research into the molecular anthropology literature that included complete mtDNA genome sequences was the introduction of potential ascertainment biases in the publically available data due to researchers sub-selecting samples based on haplogroup of interest or HV1 identity, which is still ongoing (Achilli et al., 2005; Derenko et al., 2014; Derenko et al., 2007; Fedorova et al., 2013; Starikovskaya et al., 2005; Volodko et al., 2008). Therefore, the first objective of this thesis was met by identifying and modifying existing laboratory methods and help to create new protocols for sequencing large numbers of complete mtDNA genomes of sufficient quality and coverage. These novel methods were then compared against the reigning 'gold standard' of HV1 Sanger sequencing.

The second objective of this thesis was to explore whether the reliance on traditional methods was affecting our inference of maternal population histories and population contact. This was done by analyzing the effects of both the increase in sequence data per individual (HV1 to complete mtDNA genomes) and the increase in number of individuals sequenced per populations (shifting from using only a biased subset of collected samples to using all samples in a collection).

Finally, the third objective was to put these new methods and learnings into practice by sequencing the complete mtDNA genomes of full collections of samples from multiple Siberian populations which allowed me to shed more light on their maternal prehistories and potential contact.

Novel methodologies

Chapter 2 explains the methodologies modified and developed in collaboration with colleagues from the Max Planck Institute for Evolutionary Anthropology for the purpose of sequencing complete mtDNA genomes at sufficient coverage and quality. Initial testing on the Roche 454 system was problematic so, after extensive testing, the focus was switched to the Illumina Genome Analyzer IIx, The combination of a modified version of the Meyer and Kircher (2010) method described in Chapter 2, combined with our protocol on the multiplexed hybridization enrichment of mtDNAs for the purposes of building libraries to sequence on the Illumina Genome Analyzer IIx (Maricic et al., 2010), was determined to be the most effective in terms of cost and quality.

After successfully generating initial sequences with sufficient coverage, testing was done to compare the HV1 sequences produced using Sanger methods to this novel method. A subset of the samples used in this thesis were sequenced using 2nd generation sequencing as described above and then compared to 379 published

HV1 sequences from Siberian populations that had been generated by Sanger sequencing (Pakendorf et al., 2006, Pakendorf et al., 2007). Nineteen percent of these samples (N=72) exhibited sequence differences between the two methods. However, in these 72 individuals, there were only between one and six differences per sample and almost all of them were in the poly-C regions which are not included in mtDNA population history analyses anyway.

After these protocols were finalized and verified against the established methods, I used them to generate both the sequences analyzed in this thesis and those described in our previous publication (Duggan et al., 2013). These are the first studies to use these methods on Siberian populations and, therefore the first to produced mtDNA genomes for Siberian populations that were not sub-selected for sequencing based on HV1 identity or specific haplogroups and therefore not subject to ascertainment bias. In addition to the work on Siberian populations, I also used these methods to assist in sequencing the complete mtDNA genomes from samples I collected in Burkina Faso (Barbieri et al., 2012) and those collected by others in Zambia (Barbieri et al., 2013a) and a collection from South America yet to be published. There was a high level of interest in these methods that I presented at conferences (Whitten et al., 2009; Whitten et al., 2010) so I trained other researchers from both inside and outside our facility thus resulting in a significant expansion in complete mtDNA genome sequencing and analysis, not only in the field of molecular anthropology in regions throughout the world (Barbieri et al., 2013b; Delfin et al., 2014; Duggan et al., 2014; Duggan et al., 2013; Gunnarsdóttir et al., 2011a; Gunnarsdóttir et al., 2011b; Vyas et al., 2016), but also in the field of forensic genetics (Irwin et al., 2011).

Sample selection biases and benefits of complete mtDNA genome sequencing

As discussed in Chapter 3, prior to the methods described herein, the complete mtDNA genomes publically available were largely not representative of the ethnolinguistic groups from which they were sampled. This was due to the cost and difficulty of sequencing complete mtDNA genomes using traditional Sanger methods which caused researchers to sub-select samples from collections based on either an interest in specific haplogroups or sequence identity in the HV1. Either of these reasons reduced diversity thus rendering the collections unsuitable for population wide analyses and relegated them to discussions of haplogroup histories.

This issue of putative bias was initially presented at a conference on molecular anthropology in the genomic era based on the Siberian sequences described in Chapter 3 and later in Filipino populations (Whitten et al., 2009), Gunnarsdóttir et al., 2011a). It was shown that HV1 sequence identity does not predict identity in the rest of the mtDNA genome. Rather, the majority (83%) of pairs of individuals examined herein with identical HV1 sequences have differences elsewhere (between 1 and 22). Simulations of biased sampling based on randomly selecting one individual out of each group of HV1 identical individuals showed significant differences from unbiased datasets in most of the diversity tests shown in Chapter 3. Gunnarsdóttir et al. (2011a) also tested sub-selection based on selecting only one individual from the haplogroups or lineages within those haplogroups and running a Bayesian Skyline Plot. The results of this differed markedly from those created from the full data sets in terms of population growth and effective population size estimates. Together, these studies show that full collections of samples should undergo mtDNA genome sequencing rather than sub-selecting in order to obtain a more accurate picture of population histories. Also, special care must be taken to avoid combining newly generated data with existing data that has been created in this manner.

In addition to showing the benefits of sequencing all samples in a collection to avoid bias, I also tested whether different inferences of population histories would be made when switching from HV1 sequencing to complete mtDNA sequencing. This was done on samples from the five ethnolinguistic populations described in Chapter 3 by using either the HV1 or complete mtDNA genomes to create pairwise Φst values for use in MDS plots. Indeed, there was a different pattern of distances between populations. Another researcher looked at HV1 as compared to the control region in Siberian populations and found differences as well (Johnson, 2013) but did not push for the switch to sequencing larger regions due to issues such as cost. However, with these new methods, even when a laboratory does not have the sequencing machines, it should be possible to still set up the libraries as per the protocol in Chapter 2 and send them to a facility for low cost 2nd generation sequencing.

Exploring maternal population histories and contact in Siberia

The complete mtDNA sequencing of individuals from the ethnolinguistic groups described in Chapter 4 allowed for better insights into the histories of these peoples than using the HV1 alone. This is primarily evident in the ability to perform more in-depth analyses including Bayesian Skyline Plots and an exploration of haplotype sharing. Also, given the differences between inferring population histories based on HV1 vs complete mtDNA genomes shown in Chapter 3, the remaining analyses should better reflect the populations (with the exception of those based on haplogroups, for which it is acceptable to use HV1 along with some SNPs where necessary).

In looking at the genetic structure of the populations based on standard analyses of diversity, haplogroups, and AMOVAS, very little structure was seen. It is clear that, due in part to the practices of patrilocality and high mobility in many Siberian populations, there is a high level of maternal genetic sharing. Exploring haplogroup frequencies and correspondence analyses did not add much to our understanding of these population histories. The AMOVA results showed little variance between groups for either linguistic or geographical splits.

Where the additional information from the complete mtDNA genomes contributed the most was in the exploration of population contact by analyzing shared haplotypes to identify potential directionality of sharing and viewing this in the light of Φst values. One of the conclusions that can be drawn from this is a confirmation of the shared ancestry of Yakut populations and their affinity to southern Siberian populations which points towards an origin around Lake Baikal and subsequent migration northward. Based on the differential sharing of haplotypes seen between the Yakuts and the Evens and Yakuts and the Evenks, it is likely that these Tungusic-speaking populations were already in this region and had split prior to the arrival and expansion of the Yakuts. Along with the differential nature of this sharing there was also the directionality from the Tungusic-speaking populations into the Yakuts which can be explained by the Yakuts marrying the women from the Even and Evenk groups that they encountered as they expanded. The YSE, while showing relatively high levels of haplotype sharing with both the Tungusic-speaking populations and the Yakuts also showed a common ancestry without recent admixture with the Mongolians and, to a lesser extent, the Buryats. Conversely, the YSE shared the strongest signal of recent admixture without common ancestry with the ST Evenks. Strong support for shared ancestry amongst the southern populations was seen between the Altai, Mongolians and Buryat. It was also possible to tie the Kyrgyz to this region by showing they shared a common ancestry without recent admixture.

Uncovering population demographic histories in terms of growth or decline and estimating effective population sizes in the past is another task made possible by utilizing complete mtDNA genomes. The Bayesian Skyline Plots presented here expand on those presented in Duggan et al. (2013) by including additional populations. Similar patterns of decline are seen across the Siberian populations beginning around 5,000 years ago and, in the Yakuts, a later rapid increase in the last millennium. This pattern is quite different from that of the Mongolians and Kyrgyz suggesting that, taken together with the climatic and archaeological data presented in Chapter 4, there could have been a regional climatic event that caused this signal of a population decline.

Chapter 5 was a more fine scaled investigation into Siberian population histories and contact on a less studied population. The Dolgans are one of the most recently formed ethnolinguistic populations whose origins, at least genetically, have not been fully explored. The data shown herein suggest a complex mixture of shared ancestry with Yakuts and Evenks and strong signals of recent admixture between the Dolgans, Taimyr Evenks, and YSE. The high levels of complete mtDNA haplotype sharing is striking even when compared to other Siberian populations who, as seen in Chapter 4, already exhibit a large amount of sharing. The amounts of sharing are larger between these linguistically and culturally diverse populations than they are between sub-populations of other Siberian ethnolinguistic groups such as the Yakut. This suggests a very high level of intermarriage between these diverse groups.

Just as autosomal work has shown a predominant mixture of Yakut and Evenk ancestry (Pugach et al. 2015), the mtDNA genomes explored here also show these populations to be the largest sources from which the Dolgans were formed. However, from the mtDNA genomes included in this dissertation, it appears that, maternally, the Dolgans are more closely related to Tungusic populations than the Yakuts. Paternally, based on initial results from SNP typing of the populations collected for this study as well as from data presented in Karafet et al., (2002) the Dolgans fall much closer to the Tungusic-speaking populations than to the Yakuts who have been consistently shown to have low diversity and differentiation from other Siberian populations (Karafet et al., 2002, Pakendorf et al,. 2002, Pakendorf et al., 2006). However, there are signs of paternal input from both the Yakuts and the Evenks. Taken together, the autosomal, mtDNA and Y chromosome highlight a complex mixture of populations and reveal a fluidity of marriage between populations around the Taimyr Peninsula despite the different languages and modes of subsistence.

Concluding remarks

As has been shown, it is necessary to utilize complete mtDNA genomes to gain an understanding of maternal population contact and prehistories. These large datasets of complete mtDNA genomes made possible by the methods described herein have helped to elucidate not only the prehistories of Siberian populations, but of many others around the globe. Future studies that include genome-wide data should enable even more detailed insights into the prehistory of these populations. In order to better disentangle the origins of maternal population histories, it would be beneficial to include additional linguistic and geographic neighbors as well as to develop new tools for testing potential demographic scenarios that are able to account for regions where there are high levels of putative admixture.

References

Achilli A, Rengo C, Battaglia V, Pala M, Olivieri A, Fornarino S, Magri C, Scozzari R, Babudri N, Santachiara-Benerecetti AS et al. . 2005. Saami and Berbers—An Unexpected Mitochondrial DNA Link. American Journal of Human Genetics 76(5):883-886.

Alekseev A. 1996. Ancient Yakutia: the Iron Age and the medieval epoch [in Russian]. Novosibirsk: Izdatel'stvo Instituta Arkheologii i Etnografii SO RAN.

Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, and Howell N. 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23(2):147.

Atkinson QD, Gray RD, and Drummond AJ. 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.

Barbieri C, Butthof A, Bostoen K, and Pakendorf B. 2013a. Genetic perspectives on the origin of clicks in Bantu languages from southwestern Zambia. Eur J Hum Genet 21(4):430-436.

Barbieri C, Vicente M, Rocha J, Mpoloka SW, Stoneking M, and Pakendorf B. 2013b. Ancient substructure in early mtDNA lineages of southern Africa. Am J Hum Genet 92(2):285-292.

Barbieri C, Whitten M, Beyer K, Schreiber H, Li M, and Pakendorf B. 2012. Contrasting maternal and paternal histories in the linguistic context of Burkina Faso. Mol Biol Evol 29(4):1213-1223.

Bregel Y. 2003. An Historical Atlas of Central Asia. Leiden: Brill.

Briggs AW, Good JM, Green RE, Krause J, Maricic T, Stenzel U, Lalueza-Fox C, Rudan P, Brajkovic D, Kucan Z et al. . 2009. Targeted retrieval and analysis of five Neandertal mtDNA genomes. Science 325(5938):318-321.

Cann RL, Stoneking M, and Wilson AC. 1987. Mitochondrial DNA and human evolution. Nature 325(6099):31-36.

Crubezy E, Amory S, Keyser C, Bouakaze C, Bodner M, Gibert M, Rock A, Parson W, Alexeev A, and Ludes B. 2010. Human evolution in Siberia: from frozen bodies to ancient DNA. BMC Evol Biol 10(1):25.

de Filippo C, Barbieri C, Whitten M, Mpoloka SW, Gunnarsdottir ED, Bostoen K, Nyambe T, Beyer K, Schreiber H, de Knijff P et al. . 2011. Y-chromosomal variation in sub-Saharan Africa: insights into the history of Niger-Congo groups. Mol Biol Evol 28(3):1255-1269.

Delfin F, Min-Shan Ko A, Li M, Gunnarsdottir ED, Tabbada KA, Salvador JM, Calacal GC, Sagum MS, Datar FA, Padilla SG et al. . 2014. Complete mtDNA genomes of Filipino ethnolinguistic groups: a melting pot of recent and ancient lineages in the Asia-Pacific region. Eur J Hum Genet 22(2):228-237.

Derenko M, Malyarchuk B, Denisova G, Perkova M, Litvinov A, Grzybowski T, Dambueva I, Skonieczna K, Rogalla U, Tsybovsky I et al. . 2014. Western Eurasian ancestry in modern Siberians based on mitogenomic data. BMC Evolutionary Biology 14(1):1-11.

Derenko M, Malyarchuk B, Grzybowski T, Denisova G, Dambueva I, Perkova M, Dorzhu C, Luzina F, Lee Hong K, Vanecek T et al. . 2007. Phylogeographic Analysis of Mitochondrial DNA in Northern Asian Populations. American Journal of Human Genetics 81(5):1025-1041.

Dolgikh BO. 1960. Rodovoi i plemennoi sostav narodov Sibiri v XVII v.

Drummond A, and Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology 7(1):214.

Duggan Ana T, Evans B, Friedlaender Françoise R, Friedlaender Jonathan S, Koki G, Merriwether DA, Kayser M, and Stoneking M. 2014. Maternal History of Oceania from Complete mtDNA Genomes: Contrasting Ancient Diversity with Recent Homogenization Due to the Austronesian Expansion. The American Journal of Human Genetics 94(5):721-733.

Duggan AT, Whitten M, Wiebe V, Crawford M, Butthof A, Spitsyn V, Makarov S, Novgorodov I, Osakovsky V, and Pakendorf B. 2013. Investigating the Prehistory of Tungusic

Peoples of Siberia and the Amur-Ussuri Region with Complete mtDNA Genome Sequences and Y-chromosomal Markers. PLoS ONE 8(12):e83570.

Excoffier L, Laval G, and Schneider S. 2005. Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evolutionary bioinformatics online 1:47-50.

Fedorova S, and Khusnutdinova E. 2010. Gene pool of peoples from the Republic Sakha (Yakutia): Structure, origin, genetic relationships. Russian Journal of Genetics 46(9):1102-1104.

Fedorova SA, Bermisheva MA, Villems R, Maksimova NR, and Khusnutdinova EK. 2003. Analysis of mitochondrial dna lineages in Yakuts. Molec Biol 37(4):544-553.

Fedorova SA, Reidla M, Metspalu E, Metspalu M, Rootsi S, Tambets K, Trofimova N, Zhadanov SI, Kashani B, Olivieri A et al. . 2013. Autosomal and uniparental portraits of the native populations of Sakha (Yakutia): implications for the peopling of Northeast Eurasia. BMC Evolutionary Biology 13(1):1-18.

Fenner JN. 2005. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. American Journal of Physical Anthropology 128(2):415-423.

Gogolev A. 1993. Jakuty. Problemy etnogeneza i formirovanija kul’tury. Yakutsk: Izdatel’stvo JaGU.

Gunnarsdóttir ED, Li M, Bauchet M, Finstermeier K, and Stoneking M. 2011a. High-throughput sequencing of complete human mtDNA genomes from the Philippines. Genome Research 21(1):1-11.

Gunnarsdóttir ED, Nandineni MR, Li M, Myles S, Gil D, Pakendorf B, and Stoneking M. 2011b. Larger mitochondrial DNA than Y-chromosome differences between matrilocal and patrilocal groups from Sumatra. Nat Commun 2:228.

Heyer E, Chaix R, Pavard S, and Austerlitz F. 2012. Sex-specific demographic behaviours that shape human genomic variation. Molecular Ecology 21(3):597-612.

Irwin J, Just R, Scheible M, Loreille O, and Vienna A. 2011. Assessing the potential of next generation sequencing technologies for missing persons identification efforts. FSIGSS Forensic Science International: Genetics Supplement Series 3(1):e447-e448.

Janhunen J. 1996. Manchuria. An Ethnic History. Helsinki: The Finno-Ugrian Society. Helsinki: The Finno-Ugrian Society.

Jobling MA, and Tyler-Smith C. 1995. Fathers and sons: the Y chromosome and human evolution. Trends in genetics : TIG 11(11):449-456.

Johnson SM. 2013. Phylogenetic Resolution with mtDNA D-loop vs. HVS 1: Methodological Approaches in Anthropological Genetics Utilizing Four Siberian Populations [M.A.]: University of Kansas. 120 p.

Kaessmann H, Zöllner S, Gustafsson AC, Wiebe V, Laan M, Lundeberg J, Uhlén M, and Pääbo S. 2002. Extensive Linkage Disequilibrium in Small Human Populations in Eurasia. The American Journal of Human Genetics 70(3):673-685.

Kaluzynski S. 1962. Mongolische Elemente in Der Jakutischen Sprache. Warsaw: Mouton & Co.

Karafet TM, Osipova LP, Gubina MA, Posukh OL, Zegura SL, and Hammer MF. 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.

Katoh K, Misawa K, Kuma K-i, and Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucl Acids Res 30(14):3059-3066.

Kircher M, Stenzel U, and Kelso J. 2009. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. Genome Biology 10(8):R83.

Konstantinov I. 1975. The origins of the Yakut people and their culture [in Russian]. Yakutsk: Yakutskiy filial SO AN SSSR.

Lebeynsky I. 2007. Les Nomades - les peuples nomades de la steppe des origines aux invasions mongoles (IXe siècle av. J.-C. - XIIIe siècle apr. J.-C.). Paris: Wandering. 301 p.

Li M, Schönberg A, Schaefer M, Schroeder R, Nasidze I, and Stoneking M. 2010. Detecting Heteroplasmy from High-Throughput Sequencing of Complete Human Mitochondrial DNA Genomes. The American Journal of Human Genetics 87(2):237-249.

Librado P, Der Sarkissian C, Ermini L, Schubert M, Jónsson H, Albrechtsen A, Fumagalli M, Yang MA, Gamba C, Seguin-Orlando A et al. . 2015. Tracking the origins of Yakutian horses and the genetic basis for their fast adaptation to subarctic environments. Proceedings of the National Academy of Sciences 112(50):E6889-E6897.

Malyarchuk B, Derenko M, Grzybowski T, Perkova M, Rogalla U, Vanecek T, and Tsybovsky I. 2010. The Peopling of Europe from the Mitochondrial Haplogroup U5 Perspective. PLoS ONE 5(4):e10285.

Maricic T, Whitten M, and Pääbo S. 2010. Multiplexed DNA Sequence Capture of Mitochondrial Genomes Using PCR Products. PLoS ONE 5(11):e14004.

Martinez-Cruz B, Vitalis R, Segurel L, Austerlitz F, Georges M, Thery S, Quintana-Murci L, Hegay T, Aldashev A, Nasyrova F et al. . 2010. In the heartland of Eurasia: the multilocus genetic landscape of Central Asian populations. Eur J Hum Genet.

McComb J, Crawford MH, Osipova L, Karaphet T, Posukh O, and Schanfield MS. 1996. DNA interpopulational variation in Siberian indigenous populations: The mountain Altai. American Journal of Human Biology 8(5):599-607.

Meyer M, and Kircher M. 2010. Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. Cold Spring Harb Protoc 2010(6):pdb.prot5448-.

Meyer M, Stenzel U, and Hofreiter M. 2008. Parallel tagged sequencing on the 454 platform. Nature protocols 3(2):267-278.

Mooder KP, Schurr TG, Bamforth FJ, Bazaliiski VI, and Savel'ev NA. 2006. Population affinities of Neolithic Siberians: a snapshot from prehistoric Lake Baikal. Am J Phys Anthropol 129(3):349-361.

Naumov IV. 2010. The History of Siberia. London: Routledge.

Okladnikov A. 1955. The history of the Yakut ASSR. Volume 1: Yakutia before its incorporation into the Russian state [in Russian]. Moscow: Izdatel'stvo Akademii Nauk SSSR.

Oven Mv, and Kayser M. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Human Mutation 30(2):E386-E394.

Pakendorf B. 2007. Contact in the prehistory of the Sakha (Yakuts): Linguistic and genetic perspectives. Utrecht: Leiden University.

Pakendorf B, and Novgorodov IN. 2009. Loanwords in Sakha (Yakut), a Turkic Language of Siberia. In: Haspelmath M, and Tadmor U, editors. Loanwords in the World's Languages. Berlin: De Gruyter Mouton. p 496-524.

Pakendorf B, Novgorodov IN, Osakovskij VL, Danilova AP, Protod'jakonov AP, and Stoneking M. 2006. Investigating the effects of prehistoric migrations in Siberia: genetic variation and the origins of Yakuts. Hum Genet 120:334-353.

Pakendorf B, Novgorodov IN, Osakovskij VL, and Stoneking M. 2007. Mating patterns amongst Siberian reindeer herders: Inferences from mtDNA and Y-chromosomal analyses. American Journal of Physical Anthropology 133(3):1013-1027.

Phillips-Krawczak C, Devor E, Zlojutro M, Moffat-Wilson K, and Crawford MH. 2006. mtDNA Variation in the Altai-Kizhi Population of Southern Siberia: A Synthesis of Genetic Variation. Human Biology 78(4):477-494.

Pugach I, Matveev R, Spitsyn V, Makarov S, Novgorodov I, Osakovsky V, Stoneking M, and Pakendorf B. 2015. The complex admixture history and recent southern origins of Siberian populations. bioRxiv.

Rambaut A, and Drummond AJ. 2003. Tracer

Ségurel L, Martínez-Cruz B, Quintana-Murci L, Balaresque P, Georges M, Hegay T, Aldashev A, Nasyrova F, Jobling MA, Heyer E et al. . 2008. Sex-Specific Genetic Structure and Social Organization in Central Asia: Insights from a Multi-Locus Study. PLoS Genet 4(9):e1000200.

Starikovskaya EB, Sukernik RI, Derbeneva OA, Volodko NV, Ruiz-Pesini E, Torroni A, Brown MD, Lott MT, Hosseini SH, Huoponen K et al. . 2005. Mitochondrial DNA Diversity in Indigenous Populations of the Southern Extent of Siberia, and the Origins of Native American Haplogroups. Annals of human genetics 69(0 1):67-89.

Tarskaia LA, Gray RR, Burkley B, and Mulligan CJ. 2006. Genetic Variation at the Mitochondrial DNA 9-bp Repeat Locus in the Sakha of Siberia. Human Biology 78(2):179-198.

Vasilevich G. 1969. Evenki. Istoriko-etnograficheskie ocherki (XVIII-nachalo XX v.). Leningrad: Izdatel'stvo, Nauka' Leningradskoe otdelenie.

Volodko NV, Starikovskaya EB, Mazunin IO, Eltsov NP, Naidenko PV, Wallace DC, and Sukernik RI. 2008. Mitochondrial Genome Diversity in Arctic Siberians, with Particular Reference to the Evolutionary History of Beringia and Pleistocenic Peopling of the Americas. American Journal of Human Genetics 82(5):1084-1100.

Vyas DN, Kitchen A, Miró-Herrans AT, Pearson LN, Al-Meeri A, and Mulligan CJ. 2016. Bayesian analyses of Yemeni mitochondrial genomes suggest multiple migration events with Africa and Western Eurasia. AJPA American Journal of Physical Anthropology 159(3):382-393.

Weber AW, Katzenberg MA, and Schurr TG, editors. 2010. Prehistoric Hunter-Gatherers of the Baikal Region, Siberia: Bioarchaeological Studies of Past Life Ways. Philadelphia: University of Pennsylvania Press. 319 p.

White D, and Bush A. 2010. Holocene Climate, Environmental Change, and Neolithic Biocultural Discontinuity in the Baikal Region. In: Weber AW, Katzenberg MA, and Schurr TG, editors. Prehistoric Hunter-Gatherers of the Baikal Region, Siberia:

Bioarchaeological Studies of Past Life Ways. Philadelphia: University of Pennsylvania Press.

Whitten M, Li M, Filippo Cd, and Pakendorf B. 2009. Investigating potential ascertainment bias in sample selection using complete mitochondrial DNA genome sequences of Siberian populations. Molecular Anthropology in the Genomic Era, 4th International conference of the series DNA polymorphisms in human populations. Rome, Italy.

Whitten M, Li M, and Pakendorf B. 2010. Complete mitochondrial DNA sequencing of Siberian populations. American Association of Physical Anthropologists. Albuquerque, NM, USA. p 242-243.

Wilder JA, Kingan SB, Mobasher Z, Pilkington MM, and Hammer MF. 2004. Global patterns of human mitochondrial DNA and Y-chromosome structure are not influenced by higher migration rates of females versus males. Nat Genet 36(10):1122-1125.

Wrischnik LA, Higuchi RG, Stoneking M, Erlich HA, Arnheim N, and Wilson AC. 1987. Length mutations in human mitochondrial DNA: direct sequencing of enzymatically amplified DNA. Nucleic Acids Research 15(2):529-541.

Zlojutro M, Tarskaia LA, Sorensen M, Snodgrass JJ, Leonard WR, and Crawford MH. 2008. The Origins of the Yakut People: Evidence from Mitochondrial DNA Diversity. Int J Hum Genet 8(1-2):119-130.

Zlojutro M, Tarskaia LA, Sorensen M, Snodgrass JJ, Leonard WR, and Crawford MH. 2009. Coalescent simulations of Yakut mtDNA variation suggest small founding population. Am J Phys Anthropol 139(4):474-482.

Acknowledgements

I would first like to thank all the sample donors for their generous participation in this study. Both the ones for the Siberian work here and all the hundreds of wonderful families I met throughout Burkina Faso for the sample collecting I did there. I was welcomed everywhere I went and it was an amazing experience. I also would like to thank Brigitte Pakendorf for giving me the opportunity to pursue this research in her group on Comparative Population Linguistics. Additionally I am grateful to Anne Butthof, Antje Müller and Serena Tucci for laboratory assistance, Matthias Meyer and Tomislav Maricic for helpful discussions about the laboratory protocols, Marc Bauchet for his haplogroup-calling script, Martin Kircher for processing of the Illumina GAII raw data, Mingkun Li for his scripts to analyze the sequencing reads and call the consensus sequences, the MPI-EVA sequencing group, Cesare de Filippo and Chiara Barbieri for help with R analyses and discussions, Christoph Theunert and Fred Delfin for assistance with simulations, Eugenie Stappert for insights into the Siberian populations, Mark Stoneking and the current and former members of the Human Population History group at the MPI-EVA for helpful discussions, and all the members of the former Max Planck Research Group on Comparative Population Linguistics and. I would also like to thank Connie Mulligan, Evelyn Heyer, Victor Wiebe, and Michael Crawford for the use of their sample collections.

Thank you to the people I've known throughout my education. From my Hillsboro High School friends from Nashville, Tennessee to the wonderful people at the conservation centers and field sites to my primatology masters course mates in England: It's been a great ride.

I could never have done this without the support of my loving mother, Susan Birdwell, and my stepfather, Tony Birdwell who always had my back as well as my late grandmother Mildred Raines for a great deal of patience and financial support to get to the point of starting this PhD work.

I would like to give a heartfelt thanks to all the men that I've loved and lost, especially the ones in Leipzig. Thomas, it has been an adventure and I will always hold on to the good memories. Philipp, I hope we always stay in contact and that you're as happy for me as I am for you.

To my love, the best friend that a little bearmaid could have, Mimi Arandjelovic, there are absolutely no words that can even come close to explaining how much your friendship has meant to me. From the first time I brought you a balloon in the Christmas market until the

day I die, you will have my unconditional love. Also, because of our friendship I've been lucky know Zach, Zoran and Gioia, and of course Zeus and Schumi who can always make me smile.

To my current and past flatmates at the Gruenewaldstrasse flat (Lydia, Mimi, Jesse, Nick, Adam, Gaelle), I thank you for listening to me complain for years and for wonderful distractions and your continued support. It has not just been a place where I sleep, but has also been my home and I've been lucky to share it with you all. Sergio, you never lived here, but I'm glad that it feels like sometimes you do. Peer, you're a good friend and I love you, but we can't get married and live separately.

Tim and Chiara, you are my neighbors, my moral support, my advisors in many issues, and I love the wine and dinner nights that we don't do nearly often enough.

What can I say about the best guys group a guy like me could ask for? Jesse, Alex, Fred, Christoph and Torsten: you are the brothers I never had and I'm a lucky guy to be a part of such a family. Jo and Enrico, you guys are awesome and you keep me young and hopeful.

To all the co-watchers of various tv shows and movie nights, thanks for the great times and friendship. The others mentioned already plus Erin, Jenn, Vlad, Mike, Stephane, Maureen, Jack, and Ammie.

To the Chocolate group as a whole: Nevermind, some things must never be put on paper.

A special thanks to Kathrin Franke and all the past and current team at dii-Healthcare where I've worked as a consultant since 2012. Your flexibility and support has helped me finish this dissertation as well as keeping my education going through our work researching lab technologies.

To all the others not specifically mentioned here who have affected my life in some way, thanks for being a part of my life. I have been lucky to make Leipzig my home for the last years and hopefully for more years to come. So a final thanks goes to the warm and welcoming people I've met around this beautiful city. Thanks!

Curriculum Vitae

Mark Whitten

Grünewaldstraße 3, Leipzig Germany, 04103.
00 49 (0) 176 311 89087
C.Mark.Whitten@gmail.com

Experience

dii Healthcare, GmbH
Consulting firm specializing in the IVD industry with a focus on molecular diagnostics
October 2012 – Present, Senior Consultant, *Leipzig, DE*

Max Planck Institute for Evolutionary Anthropology
August 2007 – 2014 Department of Evolutionary Genetics, *Leipzig, DE*

McDonald Institute for Archaeological Research
June 2006 - August 2007 Visiting Scholar, *Cambridge, UK*

Royal Veterinary College London
January 2006 - August 2007 Research Assistant, *London, UK*

PrIME Lab - University of Cambridge
December 2004 - November 2005 Visiting Student Researcher, *Cambridge, UK*

Education

Max Planck Institute for Evolutionary Anthropology / University of Leipzig
September 2006 PhD
Molecular Anthropology *Leipzig, DE*
Research group on Comparative Population Linguistics. Focus on Siberian population prehistory from a maternal genetic perspective and additional projects including sample collection in Burkina Faso, and laboratory work on populations from South America, Africa, and Siberia

Oxford Brookes University
September 2006 Master of Science
Biology *Oxford, UK*
Genetics work performed at the University of Cambridge PrIME lab under Dr. Knapp
Thesis: The use of non-invasive sampling of mitochondrial DNA in determining female relatedness and identifying nuclear integrations in Japanese macaques

Middle Tennessee State University
May 2004 Bachelor of Science
Biology *Murfreesboro, TN, USA*
Double Major: Biology with a concentration in Genetics / Biotechnology
Anthropology with a minor in Latin American Studies

Invited Talks

Brazil and Mexico: Insights and Challenges of Two Emerging Molecular Markets
AACC Conference International Market Briefing, 2014

Armed Forces DNA Identification Lab, Brown Bag Seminar, Rockville, MD, USA. & Institute of Legal Medicine, Innsbruck Medical University, *Innsbruck, Austria. Complete mtDNA sequencing on the Illumina Genome Analyzer II using a novel method of indexing and hybridization enrichment.*

University of Tennessee, Dept. of Anthropology, *Knoxville, TN, USA*. & Middle Tennessee State University, Dept. of Biology, *Murfreesboro, TN, USA*.
The past, present and future of DNA sequencing technologies.

Vanderbilt University, Dept. of Anthropology, Brown Bag Seminar, *Nashville, TN, USA*. & Middle Tennessee State University, Dept. of Sociology and Anthropology, Murfreesboro, TN, USA.
Population history and population contact in Siberia from a molecular anthropological perspective.

Conference Presentations

September 2011: **Development of Forensic-Quality mtDNA Data using Next-Generation Sequencing**, Rebecca Just**, Mark Whitten**, Mingkun Li, Elizabeth Lyons, Odile Loreille, Jodi Irwin. ISFG, Vienna, Austria

July 2010: Complete mtDNA genome and Y chromosome data from West Africa: demographic history and population contact in Burkina Faso, C. Barbieri, **M. Whitten**, K. Beyer, H. Schreiber, B. Pakendorf. SMBE, Lyon, France.

April 2010: Complete mitochondrial DNA sequencing of Siberian populations, **M. Whitten**, M. Li, M. Stoneking, B. Pakendorf. American Association of Physical Anthropologists, 79th Annual Meeting, Albuquerque, NM, USA.

December 2009: Investigating potential ascertainment bias in sample selection using complete mitochondrial DNA genome sequences of Siberian populations, **M. Whitten**, M. Li, C. de Filippo, M. Stoneking, B. Pakendorf. Molecular Anthropology in the Genomic Era, 4th International conference of the series DNA polymorphisms in human populations, Rome, Italy.

July 2008: SNaPshot(™) assay successfully determines SNP for chestnut coat color in historic Thoroughbred horses, **M. Whitten**, M.G. Campana, A.M. Murphy, C.J. Edwards, M. Binns, M.A. Bower. XXXI Conference of the International Society of Animal Genetics, Amsterdam, The Netherlands.

Declaration of Independence

I declare that I have conceived of and written this thesis without any inadmissible help or material that has not been indicated. I have not previously attempted to complete this or any other PhD thesis.

Leipzig, 25.05.2016,

Christopher Mark Whitten