INFERENCES OF RECENT AND ANCIENT HUMAN POPULATION HISTORY USING GENETIC AND NON-GENETIC DATA

By

ANDREW KITCHEN

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2008

To my Parents and Family.

ACKNOWLEDGMENTS

I would first like to thank my Ph.D. advisor, Professor Connie Mulligan, for her guidance as my committee chair and for her superb mentoring. I would also like to thank Prof. Michael Miyamoto for his support, advice, and insight while performing this research; and Profs. John Moore and Marta Wayne for their guiding hands as committee members. I also thank my collaborator on the Semitic language project at the University of California, Los Angeles, Dr. Christopher Ehret. I also thank the members of Prof. Mulligan's lab for insightful feedback and useful discussions about my projects. Finally, I thank my parents for providing constant support and encouragement.

TABLE OF CONTENTS

page

LIST OF FIGURES

Figure page

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

INFERENCES OF RECENT AND ANCIENT HUMAN POPULATION HISTORY USING GENETIC AND NON-GENETIC DATA

By

Andrew Kitchen

August 2008

Chair: Connie J. Mulligan
Major: Anthropology

I have adopted complementary approaches to inferring human demographic history utilizing human and non-human genetic data as well as cultural data. These complementary approaches form an interdisciplinary perspective that allows one to make inferences of human history at varying timescales, from the events that occurred tens of thousands of years ago to ones that occurred in the most recent decades and centuries. I used slow-evolving human DNA to study the peopling of the Americas tens of thousands of years ago; fast-evolving lexical data to address the origin of Semitic several thousand years ago; and fast-evolving JC virus genomes to investigate human demography within the most recent decades and centuries. In the first study, I used human mtDNA to infer the demographic history of Amerind populations and analyzed a multi-locus human DNA dataset to confirm important parameter estimates of the peopling of the Americas. My analyses produced a three-stage model for the peopling of the Americas that includes a long occupation of Siberia and a rapid expansion ~16,000 years ago into the Americas from a founding population of ~1000 to 5000 individuals. Second, I analyzed lexical data from 25 Semitic languages using computational phylogenetic techniques borrowed from evolutionary biology. Using the sampling dates of extinct languages, I was able to date events in the history of Semitic and place the origin of Semitic ~5900 years ago in what is now Syria. My final project

entailed the analysis of a large dataset of JC virus genomes (>400) and their associated sampling dates to infer the mutation rate and demographic history of JC virus. I estimated a surprisingly fast evolutionary rate for JC virus from the viral sampling dates, and confirmed this fast rate from Bayesian model tests. Ultimately, I was able to use the fast JC virus rate to infer a recent expansion of JC virus in regional populations that correlates with events in human population history of over the most recent decades and centuries.

Genetic anthropologists are uniquely equipped to appreciate the utility of combining genetic data with non-genetic or non-human data to investigate human history. This perspective allows genetic anthropologists to address questions that cannot be answered by human DNA alone by extracting as much information as possible from genetic analyses. Here, I provide examples of this integrative approach to studying human population history.

CHAPTER 1
INTRODUCTION

Demographic change has been a constant in human evolution since the expansion of modern humans out of Africa approximately 50,000 years ago (Fagundes et al. 2007). Genetic anthropologists have generally attempted to infer events in this history of change by examining the genetic diversity of modern human populations. Early genetic studies of human history used variation in proteins, such as the ABO blood groups (Landsteiner 1901; Cavalli-Sforza and Edwards 1964) and allozymes (Pauling et al. 1949; Cavalli-Sforza and Edwards 1964), as markers of underlying genetic variation at the DNA level. These data proved very useful in initial attempts to reconstruct patterns of population affinities and migration (for a summary see: Cavalli-Sforza, Menozzi, and Piazza 1994), but the advent of fast and easy DNA sequencing techniques made it possible for genetic anthropologists to directly assess human genetic diversity, both extant (e.g., Cann, Stoneking, and Wilson 1987; Kolman, Sambuughin, and Bermingham 1996; Underhill et al. 2000; Venter et al. 2001) and ancient (e.g., Paabo 1985; Handt et al. 1994; Stone and Stoneking 1998).

Despite the advance from the study of proteins to studying DNA sequences, modern human genetic diversity remains a complicated product of mutation, natural selection, genetic drift and gene flow. The interaction of these processes has important implications for the utility of human DNA as a marker of past human population events. First, the relatively slow rate of mutation for human DNA, which varies from ~$1x10^{-7}$ substitutions per nucleotide per year in non-coding mitochondrial DNA (mtDNA; Hasegawa et al. 1993; Ingman et al. 2000; Howell et al. 2003) to ~$1x10^{-9}$ substitutions per nucleotide per year in non-coding nuclear DNA (Kaessmann et al. 1999), limits human demographic inference to events that occurred thousands to hundreds of thousands of years ago. Second, natural selection can deterministically shape

diversity by increasing ("positive selection"), decreasing ("purifying selection"), or stabilizing ("balancing selection") allele frequencies according to selective pressures that are, with some important exceptions (i.e., 'near neutrality'; Ohta 1992), independent of the effects of genetic drift (Fisher 1930). Finally, genetic drift, which causes random changes in allele frequencies and has a strong affect in small populations but a weak affect in large ones, and gene flow work to stochastically alter the frequencies of selectively neutral alleles through the complex demographic processes of migration, admixture, population bottlenecks, and population expansions (Wright 1931; Kimura 1968). Thus, our ability to make inferences concerning demographic change, which directly relate to the combined effects of genetic drift and gene flow, are limited by the rate of mutation and the effects of selection acting upon human genomic regions (i.e., the other fundamental evolutionary processes). The use of neutral (or nearly neutral) variation limits the bias of selection upon inferences of demography, but the mutation rate still places a temporal limit on inferring past events. So while neutrally evolving DNA is abundant in the human genome (Payseur, Cutter, and Nachman 2002; Zhang 2007), the mutation rates of human DNA limit the utility of these data to questions about human demography extending from human speciation into the history of anatomically modern humans. For this reason, researchers have begun to use fast-evolving human parasites or pathogens, such as viruses (e.g., Holmes 2004), as complementary markers of the population history of their human host.

In my dissertation, I have adopted complementary approaches to inferring human demographic history utilizing human and non-human genetic data as well as cultural data. These complementary approaches form an interdisciplinary perspective that allows one to make inferences of human history at varying timescales, from the events that occurred tens of

thousands of years ago to ones that occurred in the most recent decades and centuries. This perspective incorporates data evolving at different rates from diverse disciplines to produce insights about human history that human DNA alone might not be able to provide. It also rests upon the understanding that while human demographic history can be investigated from human DNA data directly, other lines of evolutionary data (e.g., pathogens and languages) can be used to augment or provide independent assessments of human population history at different times in the past. Specifically, I have developed a framework in which I use human DNA sequences to infer events thousands to tens of thousands of years ago, language data to investigate human migrations centuries and millennia in the past, and virus genomes to assess the demographic history of human host populations in the most recent decades and centuries.

Over the past six decades, analysis of human DNA has provided many insights into human population history. The genetic material most often used for inferring human demographic history are mtDNA (e.g., Sherry et al. 1994; Richards et al. 1998; Macaulay et al. 2005), the male Y-chromosome (e.g., Hammer et al. 1997; Underhill et al. 2000; Quintana-Murci et al. 2001), and autosomal short tandem repeats (STRs; e.g., Rosenberg et al. 2002; Wang et al. 2007) and *Alu* elements (e.g., Hammer 1994; Athanasiadis et al. 2007). As neutral markers that are not affected by selection (Knight et al. 1996; Shen et al. 2000; Rosenberg et al. 2002; Mulligan, Kitchen, and Miyamoto 2006), the genetic variation at these loci is thought to reflect the demographic history of the populations from which they were sampled. The mutation rates of these markers vary from ~$1x10^{-7}$ to ~$1x10^{-9}$ substitutions per nucleotide per year, so they are useful to infer events thousands or tens of thousands of years in the past. Mutation rates in this range are not fast enough to provide a sufficient number of mutation events to infer past events with fine resolution, and thus estimates made using human DNA markers alone have large

variances. For example, Tishkoff et al. (2001) were able to estimate that two malaria resistance alleles of the glucose-6-phosphate dehydrogenase gene arose ~3,300 (95% credibility interval: 1600 – 6640 years ago) and ~6,400 (95% credibility interval: 3840 – 11760) years ago, but were unable to determine which of several factors (e.g., wetter climate, increased human population size, or higher population density near water) was the causative agent as the credibility intervals were too large (from 50% to 200% of mean estimates) to exclude many alternatives. Newly developed coalescent techniques that employ Bayesian probability allow for the incorporation of independent evidence, such as archaeological or climatological data, into genetic analyses (e.g., Drummond et al. 2005; Hey 2005; Beerli 2006). The Bayesian probability framework weights the likelihood of a model (including parameter values) by a prior probability that the model is correct to produce a posterior probability of the model given the data. This allows researchers to combine information from multiple interdisciplinary lines of evidence as prior probabilities to increase the power of genetic analyses to both accurately infer events and do so with greater precision. For example, archaeological evidence for the first occupation of an island might be used as a prior with zero probability placed on any models in which the colonization occurs more recently than the archaeological evidence. Unfortunately, despite the advantages provided by Bayesian techniques to integrate multiple lines of evidence into genetic analyses, many studies often neglect to do so (e.g., Hey 2005; Garrigan et al. 2007). For example, Hey employed a Bayesian coalescent technique without applying informative priors to estimate that the peopling of the New World occurred ~7,000 years ago with a founder effective population size (N_e) ~70 and large, continuous migration between Asia and the America (Hey 2005). These estimates are contradicted by the extensive archaeological evidence for human populations in the Americas by

~14,000 years ago (for a review see Goebel, Waters, and O'Rourke 2008), and this evidence could have been incorporated as prior information in Hey's Bayesian coalescent analysis.

While human DNA provides the best evidence for ancient events in human history, the focus has expanded to include markers other than human DNA to infer more recent human population history. The two most promising examples are the use of languages (e.g., Gray and Jordan 2000; Holden 2002; Gray and Atkinson 2003; Atkinson and Gray 2005) and genetic diversity of human pathogens (e.g., Falush et al. 2003; Holmes 2004; Gilbert et al. 2007) as data for understanding demographic events in human populations, specifically migration and changes in population size. These markers (languages and pathogens) typically evolve at faster rates [e.g., $\sim 10^{-4}$ substitutions per word per year for languages (Swadesh 1955) and as fast as $\sim 10^{-2}$ substitutions nucleotide per year for lentiviruses (Fu 2001)] than human DNA ($\sim 10^{-7}$ to 10^{-9} substitutions per nucleotide per year), and thus provide greater resolution for inferring population events that have occurred more recently in time, from decades to several thousands of years ago (Campbell 2000; Holmes 2004).

Languages encode genealogical information in their vocabularies (Swadesh 1955) that extends backwards in time for hundreds to thousands of years (Campbell 2000). The statistical analysis of wordlists has progressed from using simple matrices of pairwise differences and principle components to produce hierarchical clustering patterns (e.g., Rabin 1975), to the use of advanced phylogenetic techniques (such as parsimony, maximum likelihood, and Bayesian methods) to produce bifurcating phylogenies with branch lengths representing real time (e,g, Gray and Jordan 2000; Holden 2002; Gray and Atkinson 2003). Though the "linguistic clock hypothesis" is controversial (e.g., Renfrew, McMahon, and Trask 2000), accurate estimates of linguistic rates of change are necessary to make inferences in time. However, the linguistic clock

has been recently supported with strong empirical evidence (Pagel, Atkinson, and Meade 2007), and recent phylogenetic studies incorporated linguistic clocks calibrated with archaeological and historical information to estimate language divergence dates on the order of hundreds to thousands of years (e.g., Gray and Jordan 2000; Gray and Atkinson 2003). Together, the topologies and divergence dates of the inferred phylogenies have been used to combine genetic and historical evidence with linguistic evidence to provide an interdisciplinary understanding of the Bantu expansion (Holden 2002), the colonization of island south East Asia (Gray and Jordan 2000), and the expansion of agriculture into Europe from Anatolia (Gray and Atkinson 2003). The ability of languages to be transferred horizontally, in addition to vertically, between human populations also provides unique opportunities to test hypotheses of gene flow and migration in the presence or absence of substantial cultural connections and exchange.

Human pathogens, and viruses in particular, are considered good markers of recent human history due to their larger population sizes and generally faster mutation rates (Holmes 2004). This combination means that virus populations generate genetic diversity at a far faster rate than humans do, which allows for estimates of population events to be made with both higher resolution and more recently in time. A fine example of this is a study of hepatitis C evolution in Egypt, which was able to trace the progress of an epidemic arising from contaminated antischistomiasis treatments first administered in the 1920s and the effect of public health policies in curtailing the epidemic in the following decades (Pybus et al. 2003; Drummond et al. 2005). A complicating factor of using viruses as markers of human populations is that horizontal transmission of viruses could produce geographic patterns of diversity that are the result of recent dispersals instead of ancestral vicariance (e.g., Twiddy, Holmes, and Rambaut 2003; Holmes 2004; Gilbert et al. 2007). Generally speaking, fast-evolving RNA viruses cause

acute infections associated with ‘emerging’ disease (i.e., a dispersal pattern) and slow-evolving DNA viruses cause persistent infections associated with endemic disease [i.e., a vicariance pattern (Holmes 2004)], though there are exceptions to this general rule [e.g., persistent HIV infections and fast evolving polyomaviruses (Chen et al. 2004)]. Slow-evolving viruses (DNA or RNA) that cause persistent infections are thus more likely to exhibit vicariance with their human hosts and be useful for inferring ancient human population events, whereas the diversity of fast-evolving viruses is most likely the product of recent dispersal and may be useful in investigating recent migration events or connections between populations.

For my dissertation research, I chose to analyze three different types of data to study human demographic history at three different windows of time in the past. In each case, I applied advanced computational techniques to analyze the data from an evolutionary perspective and used my results to make inferences about aspects of human population history. In my first study, I analyzed human DNA and incorporated independent information from other disciplines to investigate the peopling of the Americas, which occurred more than ten thousand years ago. The first colonization of the Americas by the Amerind ancestors has been extensively studied from archaeological, climatological, and genetic perspectives (see Goebel, Waters, and O'Rourke 2008 for a review). Despite the number of genetic studies of this major event in human history, there is little consensus as to the timing, the duration and the size of the initial migration. The dates for the timing of the migration extend from ~40,000 (Bonatto and Salzano 1997) to ~7,000 years ago (Hey 2005), with many dates scattered between these outliers. These studies also postulate a migration that occurred immediately after the proto-Amerind population separated from the central East Asian gene pool, but recent analysis has suggested that the proto-Amerind population spent a considerable length of time isolated in Beringia (Tamm et al. 2007). There is

also great variation in the estimated size of the migrating population, with estimates ranging from ~10,000 (Bonatto and Salzano 1997a) to ~70 (Hey 2005) effective individuals.

I endeavored to develop a comprehensive model for the peopling of the Americas that also narrowed the estimates of the important demographic parameters of migration date and founder population size by incorporating the extensive archaeological, climatological, and historical evidence into my analysis of extant human genetic data. This was the first study to explicitly incorporate the extant physical evidence into a genetic analysis of the peopling of the Americas. For this study, I analyzed a large dataset of mtDNA coding genomes using Bayesian skyline plot analyses (Drummond et al. 2005) to infer historical changes in the effective population size (N_e) of Amerinds and their ancestors. Bayesian skyline plots are a coalescent technique used to estimate historical changes in N_e from the distribution of coalescent events in time without requiring *a priori* assumptions about how the N_e of a population should change over time (e.g., no growth vs. exponential growth). I then used a Bayesian isolation-with-migration (IM) model (Hey 2005) to analyze a combined dataset of mtDNA and nuclear sequences to estimate the Amerind founder population size and infer the effect of migration on these estimates. The IM model is a coalescent technique used to estimate the population parameters (i.e., divergence time, population sizes and growth rates, and migration rates) when a single ancestral population diverges into two daughter populations with subsequent population growth and migration between them. From these analyses, I was able to propose a model for the peopling of the Americas with three distinct stages: (i) a separation of the proto-Amerind gene pool from the Central/East Asian gene pool ~43,000 years followed by a population expansion lasting 6,000 years; (ii) a long period of stable population size in greater Beringia lasting ~20,000 years; and (iii) an expansion into the Americas ~16,000 years ago in which the

Americas were successfully and permanently colonized. This three-stage theory for the peopling of the Americas serves as an example of how to incorporate all pertinent information in a genetic analysis to produce a comprehensive hypothesis that researchers can use as a framework to test future data. This strategy appears to have been successful, as my research has been cited in a review article on the peopling of the Americas (Goebel, Waters, and O'Rourke 2008), as well as requests to use my figures in papers about the extinction of large mammals in Beringia and for a book chapter on studying human history from human DNA.

In my second study, I used wordlist data to investigate the evolution of the Semitic language family and the history of Semitic speaking populations from a time depth of several centuries to several millennia in the past. The Semitic language family is the only branch of the African language phylum Afroasiatic that is spoken in both Africa and Asia. The distribution of Semitic is the product of migration between Africa and Asia across the Red Sea within the past several millennia. However, despite a strong archaeological record attesting to the oldest Semitic languages in Mesopotamia [Akkadian and Eblaite (Buccellati 1997; Gordon 1997)] and the presence of several Semitic populations in the Levant during the Middle Bronze Age [e.g., Aramaic, Hebrew, and Phoenecian (Lloyd 1984; Richard 2003; Nardo 2007)], as well as in southern Arabia (Kogan and Korotayev 1997) and the Horn of Africa (Connah 2001), the genealogical relationships between these populations remain uncertain. The existing genetic evidence for the genealogical relationships of Semitic speaking populations is not strong, as it was primarily collected to study relationships at different levels of resolution than that between populations speaking different Semitic languages. Specifically, previous studies primarily focused on comparisons between Jewish populations and their non-Jewish neighbors (Hammer et al. 2000; Nebel et al. 2001; Rosenberg et al. 2001), which is too focused, or the contribution of

migration from the Middle East on the genetic diversity of Europe (Barbujani, Bertorelle, and Chikhi 1998; Chikhi et al. 1998; Xiao et al. 2004), which is both too broad and too ancient. Thus, understanding the genealogy of the Semitic languages themselves would shed light on the historical relationships of Semitic speaking populations. I analyzed wordlists from 25 extant and extinct Semitic languages using a Bayesian phylogenetic technique to infer their genealogical relationships. I used the extinction dates for five extinct Semitic languages (Akkadian, Aramaic, Ge'ez, Hebrew, and Ugaritic) to calibrate a linguistic clock and estimate the divergence dates of the Semitic languages, and tested alternative hypotheses of Semitic history using Bayes Factor model comparisons. I estimated that Semitic first diversified ~3900 before the common era (B.C.E.) near Mesopotamia, before expanding into the Levant and Arabia during the Early Bronze Age, and finally migrating from southern Arabia to the Horn of Africa ~1000 B.C.E. More generally, this research demonstrates how computational phylogenetic models [e.g., relaxed clocks calibrated with sampling dates (Rambaut 2000)] and model testing [i.e., log Bayes factor tests of alternative phylogenies (Suchard, Weiss, and Sinsheimer 2001)] borrowed from evolutionary genetics and molecular evolution can be used to investigate language history. Furthermore, this analysis provides a model of population history in the Middle East and Horn of Africa that incorporates linguistic and archaeological evidence that will be tested in future genetic studies of populations in these regions.

For my final study, I analyzed JC virus (JCV) genomes and estimated the JCV mutation rate so JCV evolution could be connected in time to events in the history of its human host. JCV has been often used as a marker of ancient human population history. JCV shows substantial regional sub-structure (Sugimoto et al. 1997), with apparent vicariance with some human populations (e.g., Sugimoto et al. 2002; Zheng et al. 2004; Ikegaya et al. 2005), and as a DNA

virus JCV was considered to have a slowly evolving genome. However, the phylogeny of worldwide JCV populations does not reflect the known relationships of its human hosts (Shackelton et al. 2006), and the mutation rate of JCV had never been independently estimated despite evidence that closely related polyomaviruses evolved much faster than previously thought. Accurate estimates of mutation rates are necessary to discriminate between dispersal in and vicariance with host populations (Holmes 2004) causing the geographic distribution of JCV diversity, as well as correlating events in JCV history with the history of its human host. I performed a maximum likelihood phylogenetic analyis of the complete set of available JCV genomes to confirm the sub-structure of the worldwide JCV population and then used Bayesian skyline plot analyses (Drummond et al. 2005) to estimate an independent mutation rate for JCV from viral sampling dates. I assessed the support for the independent rate using a Bayesian model testing technique, and used the independently estimated mutation rate to infer human population history from events in the history JCV. I found that JCV is an example of a fast evolving DNA virus evolving ~300 times faster than previously thought (at 3.642×10^{-5} substitutions per site per year), and that its evolution reflects changes in the demographic history of its human host over the most recent decades and centuries. More broadly, this research not only highlights the utility of fast-evolving DNA viruses as potential markers of recent human history, but also addresses the evolving nature of long-held assumptions about the relative substitution rates of RNA (fast) and DNA (slow) viruses (Duffy, Shackelton, and Holmes 2008).

In sum, this dissertation demonstrates how genetic anthropologists can investigate human demographic history from distinct types of data evolving at different rates. This study shows that taking an evolutionary approach to the study of linguistic and virus data can provide insights into human population history that are complementary to those made from human genetic diversity.

Specifically, by using data that evolve at different rates, I demonstrated how genetic anthropologists can make inferences about human demography across different timescales, from tens of thousands of years ago to just decades ago. Fundamentally, this approach relies upon the appreciation of evidence from disparate disciplines (human genetics, linguistics, and virology) and embodies the advantage of taking an anthropological perspective to the use of interdisciplinary data. Critically, interdisciplinary research utilizing different forms of data and evidence requires rigorous hypothesis testing to accurately assess the information contained within the data as well as to make results easily accessible to future research endeavors. The natural goal of research programs that synthesize multiple lines of evidence is to produce well defined hypotheses that encourage the future incorporation of additional lines of evidence, and are hopefully robust to inquiry by new datasets.

CHAPTER 2
THREE-STAGE COLONIZATION MODEL FOR THE PEOPLING OF THE AMERICAS[1]

Introduction

For decades, intense and interdisciplinary attention has focused on the colonization of the last habitable landmass on the planet - the peopling of the Americas. The first comprehensive, interdisciplinary model for New World colonization incorporated linguistic, paleoanthropological, and genetic data and generated great controversy, which was due at least in part, to the uniquely broad scope of the research (Greenberg, Turner, and Zegura 1986). Since that time, more focused studies have resulted in agreement on the general parameters of the colonization process, such as a single migration in contrast to the original three-migration model that distinguished Amerinds, Na-Dene, and Eskimo-Aleuts (Greenberg, Turner, and Zegura 1986). However, a full understanding of the complex and dynamic nature of the timing and magnitude of the colonization process remains elusive.

The majority of the genetic literature supports a single migration of Paleoindians into the New World from an East Asian source population (Schurr 2004). Specifically, the reduced variation and ubiquitous distribution of mitochondrial and Y chromosome haplogroups and microsatellite diversity throughout the New World relative to Asia argue strongly for a single migration (Mulligan et al. 2004; Wang et al. 2007). However, a great many models have been proposed that differ significantly in the timing and size of this migration event (Schurr 2004; Bonatto and Salzano 1997a; Bonatto and Salzano 1997b; Forster et al. 1996; Santos et al. 1999; Schurr and Sherry 2004; Shields et al. 1993 Silva et al. 2002; Szathmary 1993; Tamm et al. 2007; Torroni et al. 1993a; Torroni et al. 1993b). Different migration dates have been proposed

[1] Reproduced with permission from Kitchen A, Miyamoto M, Mulligan C. 2008a. A three-stage colonization model for the peopling of the Americas. PLoS ONE. 3:e1596.

ranging from ≈13 thousand years ago (kya) to ≈30-40kya (Schurr 2004; Bonatto and Salzano 1997a; Bonatto and Salzano 1997b; Forster et al. 1996; Santos et al. 1999; Schurr and Sherry 2004; Shields et al. 1993 Silva et al. 2002; Szathmary 1993; Tamm et al. 2007; Torroni et al. 1993a; Torroni et al. 1993b). Numerical estimates of the founder effective population size (N_e) are infrequent in the literature but vary substantially, from a high of ≈5000 (Bonatto and Salzano 1997b) to a low of ≈70 Paleoindian founders (Hey 2005). These dates and population sizes have been proposed to accommodate a wealth of scenarios including ancient, recent, and/or additional migrations responsible for the peopling of the Americas.

Archaeological data provide clear support for a widespread human presence in the Americas by ≈13kya (all calendar dates are recalibrated radiocarbon dates as reported in the cited literature), the time by which the Clovis complex was established across the interior of North America (Hamilton and Buchanan 2007; Waters and Stafford 2007). Older archaeological sites, e.g., the Nenana Complex in Alaska (Hamilton and Buchanan 2007), the Monte Verde site in Chile (Dillehay 1997), and the Schaefer, Hebior and Mud Lake sites in Wisconsin (Joyce 2006; Overstreet 2005), document an earlier chronology possibly 2,400 years before Clovis (Waters and Stafford 2007; Joyce 2006; Overstreet 2005). Additionally, very old radiocarbon dates have been obtained from sites in Asian Beringia suggesting that human populations had reached the north of western Beringia by ≈30kya (Goebel 2007; Pitulko et al. 2004).

The geological and paleoecological records for Beringia and northwestern North America provide further constraints on the timing for the peopling of the Americas. Beringia was a continuous landmass that connected Asia and North America roughly 60kya until ≈11-10kya (Pitulko et al. 2004; Elias et al. 1996; Hopkins 1982). However, Beringia was isolated from continental North America until ≈14kya when an intracontinental ice-free corridor opened up

between the Laurentide and Cordilleran Ice Sheets (Hoffecker, Powers, and Goebel 1993). Paleoecological data indicate that Beringia was able to sustain at least small human populations. Fossil pollen and plant macrofossils from ancient eastern Beringia are indicative of a productive, dry grassland ecosystem (Zazula et al. 2003) and paleontological evidence from Alaska and Siberia demonstrates that large mammals roamed Beringia (Guthrie 1990).

After 11-10 kya, Late Pleistocene sea levels rose sufficiently to re-inundate Beringia (Eilias et al. 1996; Hopkins 1982), creating the Bering strait that now separates the New World from Siberia by at least 100 kilometers (km) of open frigid water. Studies of human settlement throughout the Pacific Islands indicate that open water distances of >100 km constitute significant barriers to human migration, possibly because ancient people were unlikely to travel further than one day out of sight of land (Jobling, Hurles, and Tyler-Smith 2004). Similar constraints (if not worse) would apply to early humans in Alaska and Siberia, thereby severely reducing the migration rate between the New and Old World once Beringia was re-inundated. Reduced migration due to the Bering Strait remains valid even as recent rates of short-range migration have increased between Siberia and Alaska (Tamm et al. 2007). In effect, the two continents were essentially geographically isolated from 11-10 kya until modern times.

No detailed, unified theory of New World colonization currently exists that can account for the breadth and complexity of these interdisciplinary data. We analyze Native American mitochondrial DNA (mtDNA) coding genomes plus non-coding control region sequences as well as a combined nuclear and mitochondrial coding DNA dataset from New World and Asian populations. Mitochondrial DNA data represent the 'gold standard' of genetic data types and provide the most extensive comparative database for human populations worldwide (Pakendorf and Stoneking 2005). Furthermore, it has been proposed that mtDNA may be more sensitive to

demographic changes, such as population bottlenecks, due to its smaller effective population size (Wilson et al. 1985). The combined nuclear and mtDNA dataset was recently used to propose an unusually small N_e for the Amerind founders (Hey 2005), and thus investigation of this dataset is of much interest when attempting to reconcile the existing genetic evidence. We use two complementary coalescent methods to develop a comprehensive scenario of New World colonization, with a focus on the timing and scale of the migration process. Bayesian skyline plot analyses use data from a single population to provide an unbiased estimate of changes in N_e through time, and thus are a powerful means for estimating past population growth patterns when the nature of the growth (e.g., exponential or constant) is unknown (Drummond et al. 2005). The isolation-by-migration (IM) structured coalescent model uses data from sister populations to jointly estimate population divergence time, migration rates and a founder N_e, with an assumption of exponential growth (Hey 2005). Importantly, we explicitly incorporate archaeological, geological, and paleoecological constraints into both analyses. Our goal is to provide a comprehensive model for the initial settlement of the Americas that generates new testable hypotheses and has high predictive power for the inclusion of new datasets. In light of our results, we propose a three-stage model in which a recent, rapid expansion into the Americas was preceded by a long period of population stability in greater Beringia by the Paleoindian population after divergence and expansion from their ancestral Asian population.

Results

Skyline Plot Analyses

Our alignment of 77 full mitochondrial coding genomes is one of the largest published alignments of Native American mtDNA coding genomes (Figure 2-S1[2]). It includes genomes from the four major mtDNA haplogroups in the Americas (haplogroups A, B, C, and D are each represented by 17-31% of the entire sample), as well as the minor haplogroup X (2%). Correspondingly, this set of 77 complete coding mtDNA genomes represents geographically and linguistically diverse populations distributed throughout the New World (Mulligan et al. 2004). Bayesian skyline plots (Drummond et al. 2005) were used to visually illustrate changes in Amerind female effective population size (N_{ef}) over time. Bayesian skyline plots assume a single migration event, which makes the approach ideal for questions concerning the peopling of the Americas since it is generally agreed that there was a single migration (Mulligan et al. 2004). Our skyline plot of the coding genomes describes a three-stage process in which there are two distinct increases in N_{ef} at ≈40kya and ≈15kya that are separated by a long period of little to no growth (Figure 2-1). Specifically, N_{ef} increases from ≈640 [95% credible interval (CI) = 148-9,969] to ≈4,400 individuals (95% CI = 235-18,708) at the first inflection point, and from ≈4,000 (95% CI = 911-13,006) to ≈64,000 individuals (95% CI = 15,871-202,990) at the second inflection point. There is also an apparent decrease in N_{ef} prior to the second inflection point in which median N_{ef} drops to ≈2700 (95% CI = 404–36,628). We define a significant change in population size as the occurrence of non-overlapping 95% CIs at the beginning and end of an

[2] Figure 2-S1 (available online at *PLoS ONE* journal website): Multiple sequence alignment for the 77 Amerind mtDNA coding genomes used in this study. Here, "coding" refers to both protein and structural RNA genes following Pakendorf and Stoneking (2005). Gaps are represented by "-." Position 1 of this alignment corresponds to site 546 of the Anderson Reference Sequence (ARS; Anderson et al. 1981). The final position of this alignment (15,500) corresponds to site 16,042 of the ARS. Sequences starting with "Herrn," "Ing," "Kiv," and "Mis" follow the naming conventions of Herrnstadt et al. (2002), Ingman et al. (2000), Kivisild et al. (2006), and Mishmar et al. (2003), respectively.

increase (see shading in Figure 2-1). Thus, we interpret the recent ≈16-fold increase in N_{ef} over the interval ≈16-9kya as significant. The earlier ≈7-fold increase at ≈43-36kya is suggestive but not significant, although the increase is significant when compared over a much longer time period, e.g., from ≈25kya to the coalescent. Overall, the recent increase is consistent with a rapid, large-scale expansion into the Americas while the older increase is suggestive of a gradual expansion within Asia or Beringia.

The dataset of 812 concatenated mtDNA hypervariable region (HVR) I and II sequences is one of the largest published alignments of Native American HVRI+II sequences (Figure 2-S2[3]). It includes all major New World haplogroups, and represents geographically and linguistically diverse populations distributed throughout the Americas. The HVRI+II dataset was randomly divided into ten non-overlapping alignments of 81 HVRI+II sequences, which allowed for ten independent trials for parameter estimation with a sample size similar to the coding genome alignment. The HVRI+II skyline plot analyses (Figure 2-2) produce estimates for median time to coalescence (55.5kya, 95% CI = 33.5–87.2kya) and N_{ef} at coalescence (820, 95% CI = 26–3,979) and the present (66,200, 95% CI = 9,839–346,289) that are similar to the coding genome analyses (Figure 2-1). However, in contrast to the coding genome skyline plot, the HVRI+II skyline plot traces a very gradual increase in N_{ef} over ≈40,000 years with no clear inflection points. The HVRI+II plot does show a significant increase in N_{ef} but only when measured over the past 35,000 years. The fine detail evidenced in the coding genome skyline plot likely reflects the greater phylogenetic signal in the mitochondrial coding genome relative to the HVR (Non,

[3] Figure 2-S2 (available online at *PLoS ONE* journal website): Multiple sequence alignments for the ten, randomly selected, non-overlapping sets of 81 HVRI+II sequences used in this study. In these alignments, positions 1-403 correspond to HVRI, whereas sites 404-781 refer to HVRII. In turn, these alignment positions correspond to sites 16003-16400 and 30-399 of the ARS, respectively. Gaps are represented by "-." The HVRI+II sequences follow the naming conventions of HRVBase (Handt, Meyer, and von Haeseler 1998).

Kitchen, and Mulligan 2007). In general, estimates of the time to the most recent common ancestor are less sensitive to reductions in the historical signal in mtDNA sequence data than phylogenetic estimation (Non, Kitchen, and Mulligan 2007) , a result consistent with our ability to recover similar coalescence times but not the changes in N_{ef} seen when comparing the coding vs. HVRI+II skyline plots.

Isolation-with-Migration Coalescent Analyses

Bayesian IM coalescent analyses were performed on a set of nine coding nuclear and mitochondrial loci that had been previously analyzed by Hey (2005) in support of an extremely small New World founder N_e of ≈70 individuals. Thus, we performed our analysis on his identical dataset and used the same coalescent and substitution models and model parameters with the exception of new priors on the divergence time and on migration rates between Asian and Amerind populations ($m_{\text{Asia}\rightarrow\text{NW}}$ and $m_{\text{NW}\rightarrow\text{Asia}}$). The lower bound on divergence time was set to15kya, which corresponds to the period immediately preceding the earliest archaeological evidence for human habitation in the Americas (Waters and Stafford 2007; Dillehay 2007; Joyce 2006; Overstreet 2005). We also instituted serial constraints on *m* in order to gauge the effect of changing migration rates on founder N_e estimates. We interpret the various *m* values in comparison to an empirical estimate of *m* for modern Europe (*m* = 4.3; see Materials and Methods). In contrast to modern Europe, migration between the New World and Siberia from 15kya to more recent times would have become increasingly limited as Late Pleistocene sea levels rose sufficiently to inundate the Bering land bridge (Elias et al. 1996; Hopkins 1982). Thus, we expect *m* for modern Europe to be much higher than ancient migration rates between Asia and the Americas, especially after the inundation of Beringia.

Constraining divergence time by applying a lower bound of 15kya results in an estimate of ≈200 for the Amerind founding N_e. Serially constraining $m_{Asia \rightarrow NW}$ and $m_{NW \rightarrow Asia}$, in conjunction with the constrained divergence time, produces increasingly larger estimates of N_e (Figure 2-3). Specifically, as both *m* parameters are simultaneously forced to lower and more biologically realistic values, estimates of N_e steadily increase from ≈200 to ≈1,200, especially after their priors are constrained to be <5. Regardless of the specific priors on the *m* parameters, estimates for the Amerind divergence/expansion event are consistently ≈15kya (data not shown), which is very close to the lower bound of our prior established with known archaeological sites in the New World. Our results demonstrate that smaller estimates of N_e depend upon a substantial level of migration from Asia to account for present-day levels of Amerind genetic diversity, e.g., Hey's (2005) estimate of ≈70 founders is associated with a $m_{Asia \rightarrow NW} > 9.0$, which is twice the migration rate for contemporary Europe ($m = 4.3$). Eliminating all migration between Asia and the New World ($m = 0$) results in the largest estimate of N_e for the Amerind founding population of ≈1,200 individuals.

Discussion

When studying complex colonization scenarios, the interpretation of genetic data can benefit substantially from the incorporation of non-genetic material evidence. In our study, we do this in three ways. First, we interpret the skyline plot (see Figure 2-1) to reflect archaeological evidence that places Amerinds in the Americas by ≈15kya and human populations in Beringia ≈30kya, as well as geological and paleoecological evidence that Beringia was habitable yet isolated from the Americas from ≈30kya to 17kya. Second, we use archaeological radiocarbon dates to constrain the divergence time prior in our IM analyses to 15kya as the latest possible date for both the divergence of the Amerind and Asian gene pools and the Amerind expansion

into North America (Figure 2-3). Since the IM model assumes that divergence and expansion occur simultaneously, constraining the time of the expansion also requires identical constraint of the divergence date. Third, in our IM analyses we serially constrain the migration rate parameters to smaller values and deduce likely migration rates between Asia and the New World based on empirical estimates of current migration rates within Europe versus the greatly reduced migration rates of ancient people across the Bering Strait starting ≈11-10kya.

Based on our results, we propose a three-stage colonization process for the peopling of the New World, with a specific focus on the dating and magnitude of the Amerind population expansions (Figure 2-4). We propose that the first stage was a period of gradual population growth as Amerind ancestors diverged from the central Asian gene pool and moved to the northeast. This was followed by an extended period of population stability in greater Beringia. The final stage was a single, rapid population expansion as Amerinds colonized the New World from Beringia.

The initial stage of the colonization process involved the divergence of Amerind ancestors from the East Central Asian gene pool (Figure 2-4A). Based on previous studies that included Asian mtDNA sequences, this divergence likely occurred prior to ≈50kya (Bonatto and Salzano 1997a, 1997b). Our coding skyline plot (Figure 2-1) indicates that the divergence was followed by a period of gradual growth, during which the proto-Amerind population experienced a 7-fold increase from ≈640 to ≈4,400 females over ≈7,000 years, from ≈43–36kya. The migrating founder population (N_{ef} ≈640) was a small subset of the ancestral Asian population, as evidenced by the low levels of variation in New World populations relative to Asians (e.g., Mulligan et al. 2004) as well as the larger effective size of the ancestral Asian population (Hey 2005). Thus, divergence from the Asian gene pool was the time at which a severe population bottleneck

occurred that reduced the genetic variation in Amerind populations. The lack of archaeological sites in Siberia and Beringia that date to ≈43-36kya (Kuzmin and Keates 2005) suggests that this first stage of slow population growth left a light "footprint" on the landscape because of relatively rapid and continuous movement. Consistent with this hypothesis are the younger coalescent dates for modern Siberian populations relative to modern New World populations (Torroni et al. 1993; Derenko et al. 2007), which indicate that the New World migrants passed through Siberia before other East Central Asian population(s) settled permanently in this region at a later date. Such relatively rapid and continuous movement would leave few archaeological sites, which have not yet been discovered due to the vast expanse and harsh conditions of Siberia and the current inundation of Beringia. Thus, an important prediction of the first stage of our model is that older archaeological sites dating to ≈43-36kya await discovery in these regions.

The proposed second stage (Figure 2-4B) consisted of an extended period of little change in population size from ≈36-16kya (Figure 2-1). It is difficult to assign a precise geographic location to this population, but it may have occupied the large region from Siberia to Alaska, most of which is currently underwater. Our N_{ef} estimates of ≈4,000–5,000 (equivalent to N_e of ≈8,000-10,000, assuming an equal sex ratio) indicate that the proposed human presence would have been minor when compared to the size of greater Beringia. Nevertheless, the presence of this population in Beringia for ≈20,000 years would have afforded sufficient time for the generation of new mutations. Indeed, the existence of New World-specific variants that are distributed throughout the Americas indicate that substantial genetic diversification occurred during the Beringian occupation (e.g., Tamm et al. 2007; Torroni et al. 1993a; Torroni et al. 1993b; Malhi et al. 2002; O'Rourke, Hayes, and Carlyle 2000). The proposed period of Beringian occupation coincides with archaeological evidence for the first Arctic inhabitation of

western Beringia (≈30kya; Pitulko et al. 2004) and pre-dates archaeological evidence for occupation of the New World (Waters and Stafford 2007; Dillehay 1997; Joyce 2006; Overstreet 2005). This period also coincides with geological evidence for restricted access to North America because of the impenetrability of the Cordilleran and Laurentide ice sheets (≈17-30kya; Hoffecker and Elias 2003; Mandryk et al. 2001). Botanical remains, such as macrofossils and ancient pollen, indicate that Beringia was a productive grassland ecosystem rather than an exceedingly harsh Arctic desert environment (Zazula et al. 2003). Paleontological evidence from Alaska and Siberia demonstrates that large mammals such as steppe bison, mammoth, horse, lion, musk-oxen, sheep, wholly rhinoceros, and caribou inhabited this area (Guthrie 1990).Thus, the paleoecological data are consistent with a human presence in Beringia although the carrying capacity of Beringia and technological limitations of the human population may have restricted growth until the population could expand into new and fertile lands in the Americas. The rapid expansion of the population only after an ice-free corridor into North America opened (see below) suggests that the population may have departed Beringia as soon as a viable alternative presented.

The final colonization stage (Figure 2-4C) was a rapid geographic expansion into the New World resulting in a significant population increase (≈16-fold; Figure 2-1). The rapid population increase occurred over the period ≈16-9kya according to the coding skyline plot or over the past 15,000 years based on the IM analyses (the latter results supported only the most recent and largest expansion, most likely because IM analyses assume a single, simultaneous divergence/expansion event). The geological record indicates that North America became accessible from Beringia between ≈17-14kya, when the ice sheets covering what is now Canada began to retreat (Hoffecker, Powers, and Goebel 1993; Mandryk et al. 2001). The coincident

timing of an ice-free corridor into North America and the rapid expansion of the Amerind population suggests that a land route may have been the preferred entry into the New World. However, the northwest Pacific coast of North America also may have been deglaciated by ≈17kya, thus presenting a viable coastal route to continental North America (Wang et al. 2007; Mandryk et al. 2001). This period also coincides with the initial inundation of the Bering land bridge, after which migration with Asia would have been severely limited. The first unequivocal evidence for human occupation of the New World occurs in the form of Clovis sites dating to ≈13kya (Waters and Stafford 2007) and pre-Clovis sites in both North and South America dating to ≈14-15kya (Dillehay 1997; Joyce 2006; Overstreet 2005). Our datasets do not include typings from the Na-Dene or Esk-Aleut, so we limit our scope to the largest, initial migration of Amerinds into the New World. However, Na-Dene and Esk-Aleut genetic diversity represents a subset of Amerind diversity (e.g., Kolman et al. 1995; Kolman, Sambuughin, and Bermingham 1996; Merriwether, Rothhamer, and Ferrell 1995) suggesting that Na-Dene and Esk-Aleuts are derived from the same Beringian source population as Amerinds. As stated above, extensive archaeological evidence supports the presence of multiple distinct Native American material cultures by ≈13kya [e.g., Clovis, Nenana and pre-Clovis lithic technologies (Waters and Stafford 2007)]. Our results suggest that these distinct cultures derive from a single New World founder population and are most likely the product of an extensive and complex process of post-peopling migrations within the Americas, possibly through a combination of coastal and/or riverine routes (Wang et al. 2007; Fix 2005).

Determination of the size of the Amerind founding population has received considerable attention. Based on the coding Bayesian skyline plot (Figure 2-1), there is a slight decrease in population size preceding the increase seen at ≈15kya. This decrease is consistent with a

secondary founder effect in which a subset of the Beringian population seeded the proto-Amerind expansion into the Americas. Assuming the apparent decrease in N_{ef} is the result of such a founder effect, the upper bound on the founder population size is ≈5,400 individuals (N_{ef} ≈2,700). Our IM analyses suggest that the founder population size could be lower depending on prior assumptions about the over water migration rates between the Americas and Asia (see Figure 2-3). Migration rates (*m*) within Europe today based on census data have been determined to be 4.3, which can be taken as an extreme upper bound of possible ancient migration rates between the Americas and Asia, especially after the appearance of the Bering Strait ≈11-10kya. Restricting migration rates to <1 results in founder N_e estimates between ≈1,000 and ≈1,200, with ≈1,200 serving as an asymptotic upper bound (see Figure 2-3). Taken together, our Bayesian skyline plot and IM analyses suggest that a founder population with N_e = 1,000-5,400 colonized the New World in a process characterized by a rapid geographic and population expansion. The range of N_e values can be translated into an approximate census population size by applying a scale factor estimated from large mammal populations (scale factor = 5; Templeton 1998), which suggests that the founder population consisted of ≈5,000-27,000 people.

Our three-stage model now awaits further critical testing with new datasets of independent nuclear loci and more sophisticated methods of coalescent analysis. The extensive dataset of ≈700 autosomal microsatellites, compiled by Wang et al. (2007) for both Native American and worldwide populations, offers the opportunity to evaluate critically the size, timing, and duration of each step in our model at essentially a population genomics level. Future versions of BEAST will incorporate a structured coalescent where migration as well as population growth will be allowed to occur among populations from both the New World and Asia (http://evolve.zoo.ox.ac.uk/beast/manual.html). In these BEAST analyses, the microsatellites can

be modeled under a stepwise "ladder process," whereby alleles are inter-related according to their repeat lengths. One can then summarize over these microsatellite loci by assuming independence, which thereby allows for the multiplication of their separate posterior distributions and final estimations of their combined Bayesian skyline plot. In these ways, we fully anticipate that such critical testing will lead to many important refinements of our three-step model, including a further narrowing of our proposed range for the size of the founding population as well as new details about post-peopling expansions within the New World.

Materials and Methods

Datasets

Three datasets were collected for analysis including: (*i*) 77 mtDNA coding genomes; (*ii*) 812 mtDNA HVRI+II sequences; and (*iii*) combined nuclear and mitochondrial coding DNA dataset. The 77 mtDNA coding genomes were collected from publicly available resources (Herrnstadt et al. 2002; Ingman et al. 2000; Kivisild et al. 2006; Mishmar et al. 2003) and aligned using ClustalX (Thompson, Higgins, and Gibson 1994). The resultant 15,500 base pair (bp) multiple alignment was edited by hand to minimize the number of unique gaps and to ensure the integrity of the reading frame (Figure 2-S1). A total of 812 combined HVRI+II sequences were collected from HVRbase (http://www.hvrbase.org; Handt, Meyer, and von Haeseler 1998). These sequences were aligned following the coding mtDNAs, resulting in a multiple alignment of 771 bps (Figure 2-S2). The complete dataset of 812 HVRI+II sequences was randomly divided into ten non-overlapping alignments of 81 sequences that approximate the sample size for the coding mtDNA dataset. Skyline plot analyses of larger datasets (up to 200 HVRI+II sequences) gave the same results as the 81 sequence datasets (data not shown). Thus, the smaller datasets of 81 sequences each were emphasized here since they avoided the likelihood rounding errors that can occur when using large, heterogeneous datasets in Bayesian skyline plot analyses.

The coding nuclear and mtDNA dataset from Asian and Native American populations of Hey (available at http://lifesci.rutgers.edu/~heylab/; Hey 2005) consisted of two autosomal coding loci, five X-chromosome coding loci, one Y-chromosome coding locus, and the complete mtDNA coding genome (totaling 28,454 aligned bps). The sample sizes for these nuclear loci and mitochondrial genome varied from 12 - 108 sequences.

Bayesian Skyline Plot Analyses

Bayesian skyline plots (Drummond et al. 2005) were used to estimate changes in Amerind N_{ef} over time by providing highly parametric, piecewise estimates of N_{ef}. This approach produces serial estimates of effective population size from the time intervals between coalescent events in a genealogy of sampled individuals, and utilizes a Markov chain Monte Carlo simulation approach to integrate over all credible genealogies and other model parameters. It thereby differs from previous approaches (e.g., Polanski, Kimmel, and Chakraborty 1998) in that Bayesian skyline plots fully parameterize both the mutation model (including relaxed clock models) and the genealogical process, whereas prior methods relied on generating estimates from summary statistics (e.g., the use of pairwise differences by Polanski, Kimmel, and Chakraborty 1998).

In these analyses, estimates of τ (N_{ef} x generation time) were converted to N_{ef} by dividing by a generation time of 20 years, following convention (Hey 2005). Skyline plots were generated for the 77 mtDNA coding genome sequences and the ten datasets of HVRI+II sequences using the program BEAST v1.4 (http://beast.bio.ed.ac.uk). These BEAST analyses relied on the same coalescent and substitution models and run conditions of Kitchen, Miyamoto, and Mulligan (2007) except as noted below. Plots were generated using the established mutation rates (μ) for coding mtDNA ($\mu = 1.7 \times 10^{-8}$ substitutions / site / year; Ingman et al. 2000) and HVRI+II

mtDNA ($\mu = 4.7 \times 10^{-7}$; Howell et al. 2003). Markov chains were run for 100,000,000 generations and sampled every 2,500 generations with the first 10,000,000 generations discarded as burn-in. Three independent runs were performed for all coding and HVRI+II Bayesian skyline plot analyses. Markov chain samples from the three independent mtDNA coding replicates and from the 30 HVRI+II analyses were separately combined using the LogCompiler program (distributed with BEAST) and analyzed using Tracer v1.3 to produce the final Bayesian skyline plots.

Isolation-with-Migration Coalescent Analyses

Bayesian IM coalescent analyses were performed using the program IM (Hey 2005) to estimate Ne for the Amerind founder population (males + females) and the divergence time for Amerind and Asian populations. We used the same combined nuclear and mtDNA dataset, same coalescent and substitution models, and same model parameters as Hey (2005) with the exception of new priors on the divergence time and on the migration rates between Asian and Amerind populations. All IM analyses were performed using a flat uniform prior for the divergence time of Amerind and Asian populations set to the interval 15-40kya. The lower bound of this prior is based on accepted archaeological and climatological evidence for the first presence of humans in the Americas (Waters and Stafford 2007; Dillehay 1997; Joyce 2006; Overstreet 2005). The upper bound of the flat uniform priors on the migration rates per mutation per generation between the Amerindian and Asian populations ($m_{\text{Asia}\rightarrow\text{NW}}$ and $m_{\text{NW}\rightarrow\text{Asia}}$) was set to 12 different values (0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, and 50). To help interpret these results, we relied on an estimate of the migration rate in modern Europe as obtained from census data (Weale et al. 2002). Specifically, we converted their migration rate estimate of 0.0004 migrations per gene copy per generation (recalculated assuming a generation time of 20 years based on Hey

2005) to our units of migrations per mutation per generation (*m*) by dividing the former by the geometric mean of the mutation rates for the nine loci in this dataset (9.32×10^{-5} mutations per locus per generation; Hey 2005). These calculations resulted in $m = 4.3$ for modern Europe. In contrast, the ancient migration rates between the New World and Asia would have been significantly less, especially after their geographic separation due to the re-inundation of Beringia starting at ≈11kya (see Introduction). Ten independent replicates were performed for each of the 12 upper bound values on the migration rates, for a total of 120 IM analyses. All Markov chains were run for 100,000,000 generations without heating.

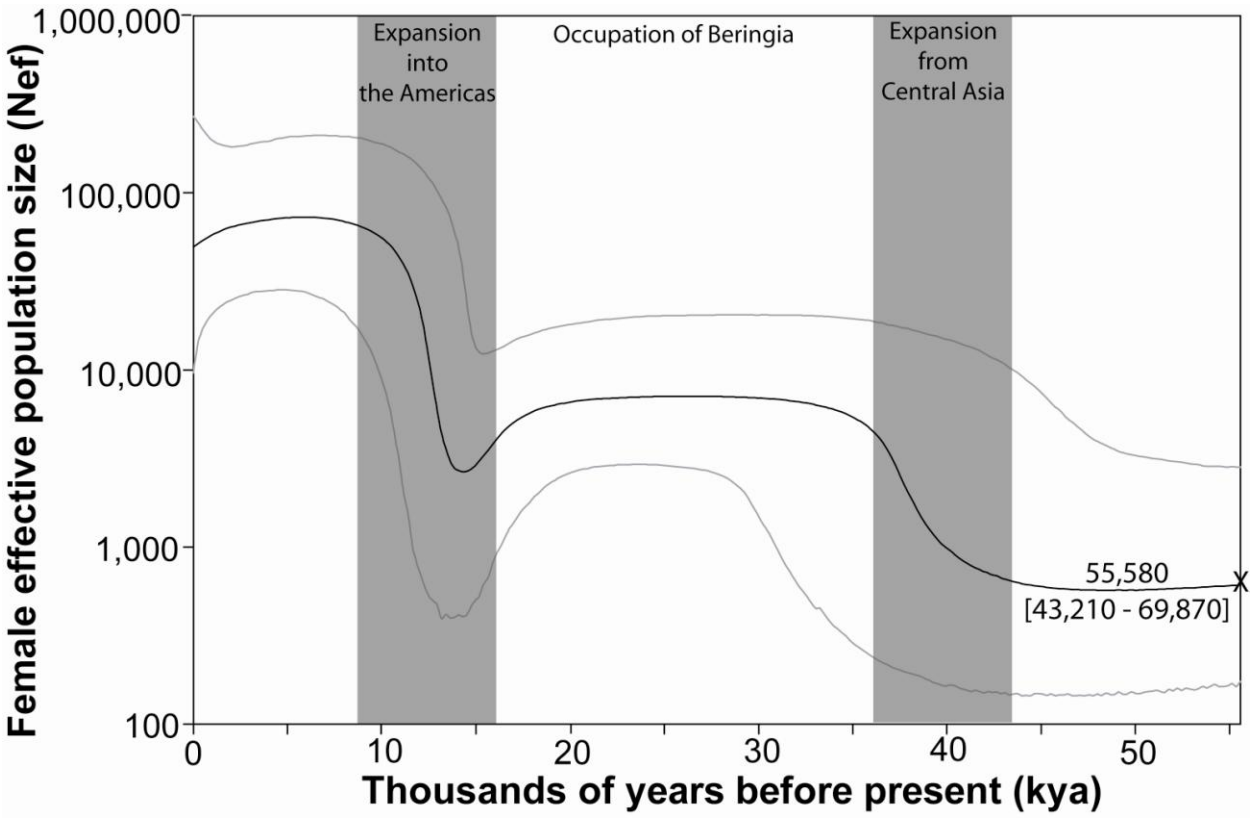

Figure 2-1. Bayesian skyline plot for the mtDNA coding genome sequences. The curve plots median Nef with its 95% CI indicated by the light gray lines. The calculated Nef assumes a generation time of 20 years following Hey (2005); alternatively, using a generation time of 25 years (Fenner 2004) would uniformly decrease all estimates of Nef by 20%. "X" marks the median coalescent time with its 95% CI given in brackets. The shaded regions highlight two periods of substantial population growth. This skyline plot provides the principal evidence for our three-stage model of New World colonization, i.e., the three stages that are depicted and labeled here.

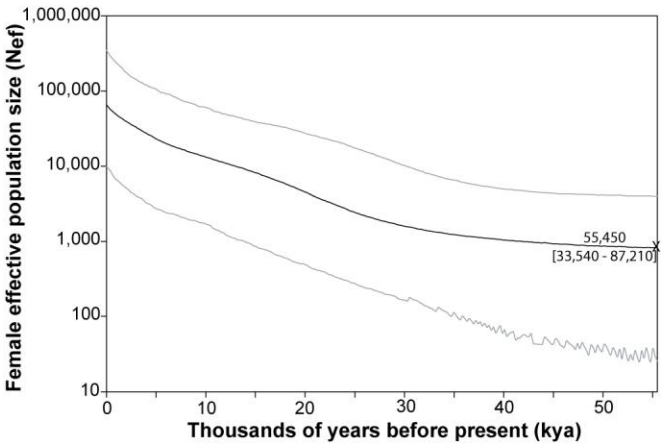

Figure 2-2. Bayesian skyline plot for the mtDNA HVR I+II datasets.

This plot follows the conventions of Figure 2-1. Its estimates of coalescent time and N_{ef} at the coalescence and today are in agreement with the coding mtDNA skyline plot (Figure 2-1). In contrast, this HVRI+II plot provides little resolution for other population size changes, most likely because of mutational saturation in the non-coding control region (see text).

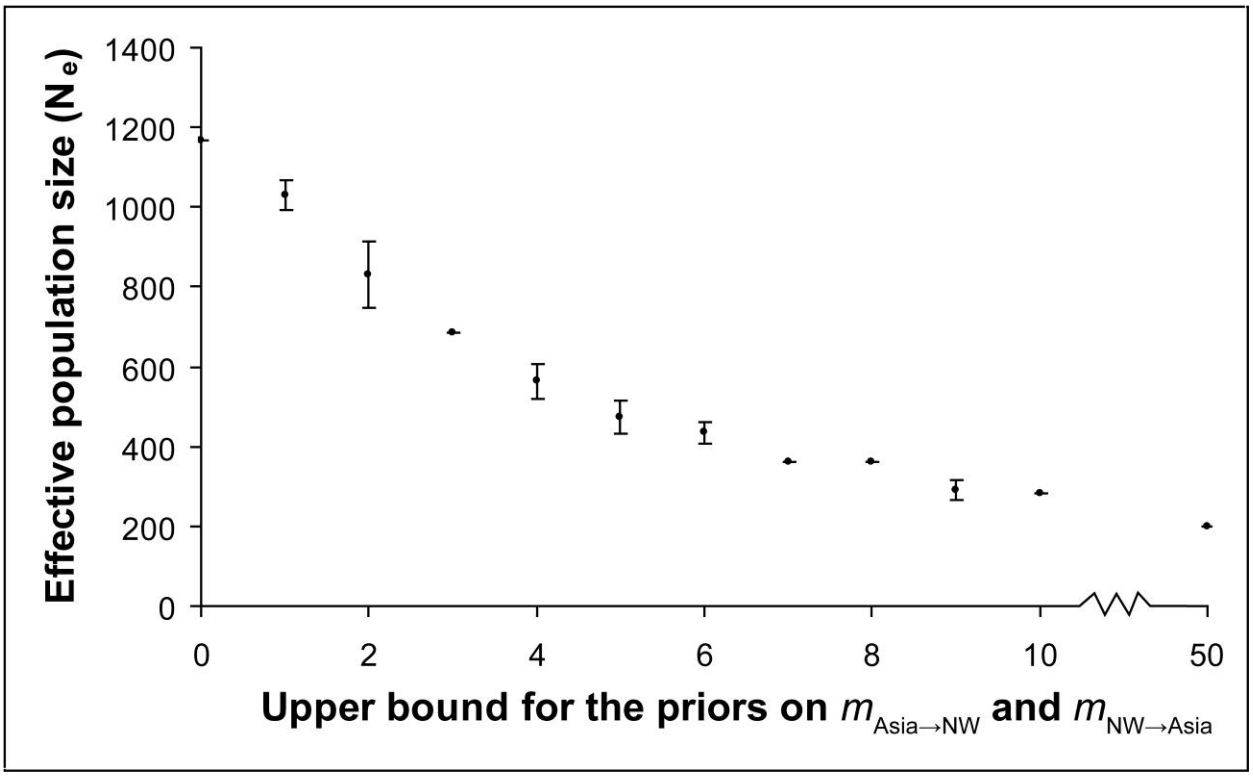

Figure 2-3. Graph of IM results for the combined nuclear and mitochondrial coding DNA dataset. The plot depicts mean Ne for the Amerind founder population (y-axis) as a product of increasing the constraint on the upper bound of the priors for the migration rates (x-axis). In these analyses, the prior on the lower bound of the divergence time was uniformly set to 15kya on the basis of known archaeological materials for human occupation in the New World (see text). Each point is based on the average of the estimated medians for ten independent replicate analyses, with the bars corresponding to ± 1 standard deviation. These standard deviations are often small (with coefficients of variation less than 0.01), since their Markov chains were run for 100 million generations each.

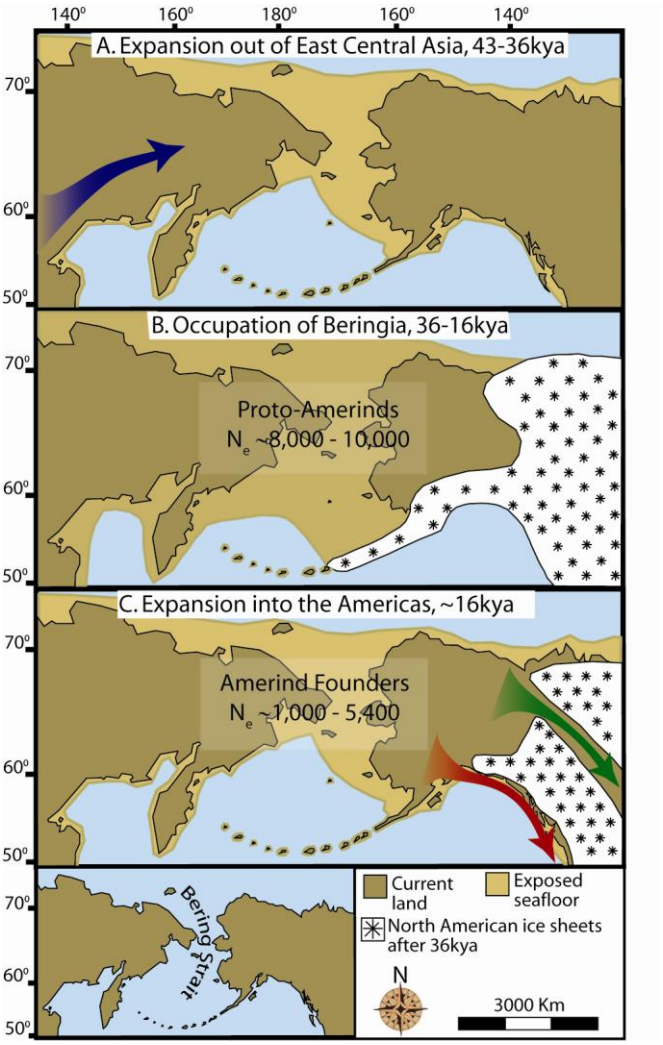

Figure 2-4. Maps depicting each phase of our three-step colonization model for the peopling of the Americas. (A) Divergence, then gradual population expansion of the Amerind ancestors from their East Central Asian gene pool (blue arrow). (B) Proto-Amerind occupation of Beringia with little to no population growth for ≈20,000 years. (C) Rapid colonization of the New World by a founder group migrating southward through the ice free, inland corridor between the eastern Laurentide and western Cordilleran Ice Sheets (green arrow) and/or along the Pacific coast (red arrow). In (B), the exposed seafloor is shown at its greatest extent during the last glacial maximum at ≈20-18kya (Hopkins 1982). In (A) and (C), the exposed seafloor is depicted at ≈40kya and ≈16kya, when prehistoric sea levels were comparable (Elias et al. 1996; Hopkins 1982). Because of the earth's curvature, the km scale (which is based on the straight line distance at the equator) provides only an approximation of the same distance between two points on these maps. In addition, a scaled-down version of Beringia today (60% reduction of A-C) is presented in the lower left corner. This smaller map highlights the Bering Strait that has geographically separated the New World from Asia since ≈11-10kya.

CHAPTER 3
BAYESIAN PHYLOGENETIC ANALYSES OF SEMITIC LANGUAGES IDENTIFY AN EARLY BRONZE AGE ORIGIN OF SEMITIC AND A SINGLE IRON AGE MIGRATION TO AFRICA FOR ETHIOSEMITIC

Introduction

The Semitic languages comprise one of the most studied language families in the world. Semitic is of particular interest due to its association with the earliest civilizations in Mesopotamia (Lloyd 1984), the Levant (Rendsburg 2003), and the Horn of Africa (Connah 2001), which gave rise to several of the world's first great literary (e.g., the Akkadian poem *The Epic of Gilgamesh*) and religious (e.g., Judaism, Christian, and Islamic) traditions. This association dates back at least 4350 years before present (historical sources date events relative to the common era that began 2000 years ago, but for consistency we will use years before present [ybp] as our dating standard) to ancient Sumer in Mesopotamia, where Akkadian replaced Sumerian (the first known written language) and adopted the Sumerian cuneiform script (Buccellati 1997). From this time forward, archaeological evidence for Semitic amongst the Hebrews and Phoenicians in the Levant (Rendsburg 2003) and the Aksumites in Horn of Africa (Connah 2001) suggests that Semitic and Semitic-speaking populations underwent a complex history of geographic expansion and migration tied to the emergence of the earliest urban civilizations in these regions. However, though Semitic-speaking populations are well represented in the archaeological record (Lloyd 1984; Connah 2001; Richard 2003a; Nardo 2007), their origins and relationships to each other remain uncertain. Without knowledge about the history of Semitic populations, our understanding of the ancient civilizations and lasting cultural traditions of the Middle East and the Horn of Africa remains uncertain.

Despite multiple genetic studies of extant Semitic-speaking populations (e.g., Nebel et al. 2002; Capelli et al. 2006), much is still unknown about the genealogical relationships of these

populations. Most previous genetic studies focused on time frames that are either too recent [e.g., the origin of Jewish communities in the Middle East and Africa (Hammer et al. 2000; Nebel et al. 2001; Rosenberg et al. 2001)] or too ancient [e.g., the original out-of-Africa migration (Passarino et al. 1998; Quintana-Murci et al. 1999)] to provide insight about the origin and dispersal of Semitic and Semitic-speakers.

Linguistic studies can aid our understanding of the origin and dispersal of Semitic and Semitic-speakers. Previous historical linguistic studies of Semitic languages have utilized the 'comparative method' to infer the genealogical relationships of Semitic [for review, see (Faber 1997)]. The comparative method is a technique that uses the pattern of shared, derived changes in language (e.g., vocabulary, syntax, grammar), termed innovations, to assess the relative relatedness of languages but cannot date the divergences between languages (Campbell 2000). Simply stated, two languages that share more innovations in common are more closely related via a common ancestor than if they shared fewer innovations, and the pattern of pairwise comparisons of innovation sharing are used to infer relative relationships between many languages. Cognates, which are words that share a common form and meaning through descent from a common ancestor (e.g., the English word 'night' is a cognate with the German word for night, 'Nacht'), serve as the data used most often in comparative analyses.

Several alternative sub-groupings of Semitic have been proposed using the comparative method, but the field has generally coalesced around a model that places the ancient Mesopotamian language Akkadian at the root of Semitic (Hetzron 1976; Faber 1980; Faber 1997). This standard model divides Semitic into East Semitic, composed of only the extinct Akkadian and Eblaite languages, and West Semitic, consisting of all remaining Semitic languages, distributed from the Levant to the Horn of Africa. West Semitic is in turn divided into

South [consisting of Ethiosemitic, 'Epigraphic' South Arabian (ESA), and Modern South Arabian (MSA)] and Central geographic groups, but the genealogical relationships of the languages within these two groups are ill-defined (Huehnergard 1990; Huehnergard 1992; Rodgers 1992; Faber 1997) and require further investigation to clarify their roles in the ancient history of the Middle East and Horn of Africa. Additionally, no consensus exists for placing Arabic in either the Central or South Semitic groups (Hetzron 1976; Blau 1978; Diem 1980; Huehnergard 1990; Huehnergard 1992; Faber 1997), which makes Arabic's genealogical location simultaneously uncertain and interesting, as Central and South Semitic are geographically and genealogically distinct entities.

Dating language divergences has been controversial, especially when linguistic clocks are involved (see Renfrew, McMahon, and Trask 2000 for discussion). The linguistic clock is controversial because it is assumes that languages evolve in a probabilistic manner with a fixed rate (Ehret 2000), whereas there is evidence for variation in rates between languages and words and no reason why languages should evolve with predictable rates (e.g., Blust 2000). However, recent studies show that variation in the rates of linguistic change follow generalized rules that apply across language families (Pagel, Atkinson, and Meade 2007; Atkinson et al. 2008). This suggests that variation in rates between words and languages can be appropriately modeled by applying techniques used in molecular evolution (e.g., gamma distributions of site variation and relaxed clock models) to produce a relaxed linguistic clock that accounts for rate variation. Since the comparative method does not provide time estimates for language divergences, relaxed linguistic clocks coupled with phylogenetic methods also borrowed from evolutionary biology provide a statistical alternative to accurately date language divergences.

In this study, we analyzed lexical data for 25 Semitic languages distributed throughout the Middle East and Horn of Africa (see Fig. 3-1 for the geographic distribution of languages) using a Bayesian phylogenetic method to simultaneously infer the genealogical relationships and estimate the divergence dates of the Semitic languages. We used epigraphic and archaeological evidence for the sampling dates of the lexical information (i.e., the time at which the materials were written and the lexical data stopped changing) from extinct Semitic languages (Akkadian, Aramaic, Hebrew, Ge'ez, and Ugaritic) to calibrate a relaxed linguistic clock and increase the accuracy of our divergence date estimates. We employed a log Bayes Factor (log BF) model-testing technique to statistically test alternative Semitic histories and verify the information content of our lexical data. Finally, we combined our divergence date estimates with the epigraphic and archaeological evidence to form an integrated model of Semitic history.

Results

Genealogy of Semitic Languages

The phylogenetic analysis of the Semitic languages produced the phylogenetic tree shown in Fig. 3-2. A brief summary of the tree highlights (1) the greater age of the non-African Semitic languages (5925 ybp vs 2925 ybp), (2) the presence of Akkadian followed by Central Semitic near the root of Semitic, (3) the relatively poor resolution of non-African languages in comparison to the well resolved relationships of the African languages and (4) the well-resolved and recent divergences of the Ethiosemitic languages in a monophyletic (i.e., single origin) clade. Branches with posterior probability estimates (the probability that a group of languages is more closely related to each other than to other languages) less than 0.70, which is generally considered the benchmark for statistically well supported clades, were considered to be unresolved (i.e., the relative pattern of divergence amongst the taxa could not be ascertained) and collapsed to reflect this uncertainty. Long branches are indicative of long intervals between

divergences or the presence of un-sampled languages, such as the long branch leading to the two Arabic languages in our study, whereas short branches indicate bursts of diversification that are a hallmark of linguistic evolution (Nettle 1999; Atkinson et al. 2008).

Semitic Language Divergence Dates

In addition to illustrating the relationship between different Semitic languages, our phylogenetic analysis produced dates for the divergence of related languages. The mean divergence times and 95% credible intervals (CI) of all language divergences are depicted on the tree in Fig. 3-2, with all times in years. The tree displayed a primary division between Akkadian and the remaining Semitic languages, which supported an estimated origin of Semitic ~5900 ybp (CI = 4300 – 7750 ybp) during the Early Bronze Age (Ehrich 1992). A secondary division occurred between Central and South Semitic groups (node **A** in Fig. 3-2) with an estimated divergence time of ~5425 ybp (CI = 3850 – 7525 ybp). The Central Semitic clade (Arabic, Aramaic, Hebrew, and Ugaritic) had strong posterior support (0.82; node **B**) and a weakly supported internal divergence of Ugaritic from an unstructured group of Arabic, Aramaic, and Hebrew dated at ~4100 ybp (CI = 3400 – 5925 ybp). The Arabic clade (node **C**) had 100% posterior support and an estimated divergence time of Moroccan and Ogaden Arabic of ~540 ybp (CI = 110 – 1375 ybp).

On the other half of the tree, the South Semitic clade showed an ancient divergence (node **D**) Ethiosemitic and MSA language groups dated to ~4525 ybp (CI = 2700 – 6825) and over-lapping with the transition from the Early to Middle Bronze Ages. The most recent common ancestor of strongly supported MSA clade (1.0; node **E**) was estimated to have been extant ~1300 ybp (CI = 475 – 2550), and the narrow geographic distribution of MSA along the southern

coast of Arabia facing the Gulf of Aden suggests that the diversification of MSA occurred in this area.

The lone, strongly supported branch (posterior = 0.94) leading to Ethiosemitic indicates a single origin for the Semitic languages in the Horn of Africa with their diversification into North and South clades (node **F**) occurring ~2975 ybp (CI = 1850 – 4500 ybp), or during the second period of the Iron Age. The large number of small internal branches in the Ethiosemitic group indicates a rapid diversification of these languages. Our phylogenetic analysis divided the South Ethiosemitic languages into three well-supported clades with estimated divergences occurring approximately a millennia ago: Outer Gurage (node **G**; divergence = 1175 ybp, CI = 460 – 2175), East Gurage (node **H**; divergence = 1125 ybp, CI = 400 –2175), and Amharic-Argobba-Gafat (node **I**; divergence = 1000 ybp, CI = 300 – 2100).

Log Bayes Factor Tests

The validity and usefulness of our interpretations rests on the accuracy of our phylogeny. Thus, we assessed the robustness of our phylogenetic analysis by statistically testing alternative histories of Semitic. This was done using log Bayes factor (BF) model tests, which compare the probability that each model produced the observed data (i.e., the wordlist data). Log BF values (all values in log base 10 units) in the intervals 0-1, 1-2, and >2, are considered to be "substantial," "strong," and "decisive", respectively, evidence for the primary model (Kass and Raftery 1995). We tested four alternative Semitic histories in the following three comparisons. The comparison between the Standard model [which constrains Akkadian to the root and constrains the sub-division of West Semitic into Central and South Semitic clades] and a phylogenetic tree with no constraints showed a "strong" preference for the Standard model (log BF = 1.1,), consistent with Semitic genealogies estimated using the comparative method (Faber

1997). We next compared the Standard model to a model of Semitic history that placed Ethiosemitic at the root (i.e., an African origin for Semitic), and the log BF test showed "substantial" support for the Standard model of Semitic history (log BF = 0.793), consistent with a non-African origin of Semitic. Our final comparison, between the Standard model and the Old Arabic model (a version of the Standard model with the most recent common ancestor of Ogaden and Moroccan Arabic constrained to the period prior to the expansion of Arabic ~1400 ybp.), resulted in "strong" support for the Standard model (log BF = 1.197) and demonstrated the accuracy of our rate estimates and the ability of our analysis to reject unrealistic divergence dates (i.e., the divergence of Arabic prior to Arabic's geographic expansion in the 7th century).

Discussion

The Semitic language family is unique in that it is the sole member of the Afroasiatic language phylum to have been historically spoken both outside Africa as well as in Africa (Hayward 2000). Semitic is associated with some of the oldest urban states in ancient Mesopotamia and the world, such as the first Akkadian Empire of Sargon in the third millennium B.C.E. (Lloyd 1984). Furthermore, Semitic societies in the ancient Levant, particularly the Canaanites, Phoenicians, and Israelites, played important roles in the second and first millennia B.C.E. by developing new kinds of long-distance commerce in the Mediterranean and ancient Middle East and creating great literary and religious traditions (Rendsburg 2003). The oldest states in southern Arabia (e.g., Saba'; von Wissman 1975) and the Ethiopian Highlands [Yeha, Aksum (Connah 2001)] are also associated with Semitic and underscore both the antiquity and geographic extent of the influence and importance of Semitic-speaking populations in the history of the Middle East and Horn of Africa. However, important aspects of the complex history underlying this distribution of Semitic remain unresolved, such as the timing and location of

Semitic origins (e.g., non-African vs. African origin), the age of Semitic languages in the Levant, and the timing and provenience of the Semitic languages in Africa.

Historical linguists have traditionally used the comparative method to produce language trees that depict degrees of relatedness to a common ancestor. The comparative method uses shared linguistic innovations (i.e., derived linguistic features, analogous to DNA mutations) to infer patterns of hierarchical relatedness, but can not produce estimates of divergence times. Conversely, a branch of historical linguistics called lexicostatistics (or glottochronology) employs the percentage of cognates shared between languages as a measure of evolutionary distance that is used to estimate language divergence dates based on a strict linguistic clock. This method is controversial and seldom used today because many exceptions have been found to its underlying assumption that linguistic change is constant in time and across languages (Ehret 2000). Recent research, however, has identified generalized mechanisms underlying rate variation that act across language groups (Pagel, Atkinson, and Meade 2007; Atkinson et al. 2008) suggesting that variation in rates of linguistic evolution can be accurately modeled and accounted for using relaxed clocks analogous to those used in molecular evolution.

Several recent studies have taken a statistical approach to inferring language trees via the application of model-based, computational phylogenetic methods borrowed from evolutionary biology to the analysis of cognate lists traditionally used in historical linguistics (Gray and Jordan 2000; Holden 2002; Gray and Atkinson 2003). These analyses employed computational phylogenetic techniques that allowed the evolution of the linguistic data to be modeled, the strength of alternative language trees to be assessed by objective criteria, the confidence of evolutionary models and parameters to be estimated, and alternative linguistic histories to be statistically tested. The first of these studies used maximum parsimony phylogenetic criteria,

which favor trees that minimize the number of evolutionary changes necessary to explain the data, to investigate the Austronesian (Gray and Jordan 2000) and Bantu (Holden 2002) language expansions but did not estimate divergence dates for either language family. A similar study of Indo-European (Gray and Atkinson 2003) employed a Bayesian phylogenetic method (which estimates the probability that a model is correct given the data and some independent, prior knowledge) to infer an Indo-European language tree and then applied a likelihood dating technique (which maximizes the probability that the model produces the data), calibrated with *a priori* constraints on the times of specific sub-family divergences, to date events in the history of Indo-European. In contrast to the likelihood method used to estimate divergence times of Indo-European, which does not account for uncertainty in both the tree and the rate, we employ a Bayesian phylogenetic technique that allows for the co-estimation of Semitic language trees and divergence dates while fully accounting for the uncertainty of both (Drummond et al. 2006). We also provide the first use of language sampling dates (as opposed to constraints on dates of specific nodes within the language tree used by (Gray and Atkinson 2003)) drawn from the archaeological record to calibrate the rate of linguistic evolution and date events in Semitic history. Furthermore, we use Bayes factor model tests to provide quantitative support for distinguishing between alternative hypotheses of Semitic language evolution and confirm the information content of our dataset.

Semitic Origins

Our phylogenetic analysis of Semitic produced a language tree with dates that establish Akkadian as the deepest branch in the Semitic family tree and estimates the origin of Semitic to ~5900 ybp (Fig. 3-2). This date places the Semitic origin a surprising ~1500 years before the first Akkadian inscriptions, which were written using Sumerian cuneiform script (Daniels 1997) and

appear in the archaeological record of northern Mesopotamia (Buccellati 1997). The city-states of Sumer were established and flourishing in Mesopotamia with their own indigenous languages ~5900 ybp (Lloyd 1984), so it is unlikely that Akkadian was spoken in Sumer for the entirety of the 1500-year interval between its divergence from ancestral Semitic and initial appearance in the archaeological record of Sumer. Furthermore, the closest relative of Akkadian and the only other member of East Semitic, Eblaite (unsampled in our study), was spoken in northwestern Syria, in a region adjacent to where some of the oldest West Semitic languages were spoken. The presence of ancient members of the two oldest Semitic groups (i.e., East and West Semitic) in the same area suggests their divergence from ancestral Semitic occurred there, in what is today Syria. This, combined with the long interval between the origin of East Semitic and the appearance of Akkadian in Sumer, suggest that East Semitic originated in present-day Syria and Akkadian later spread from Syria eastward into Mesopotamia and Sumer (see Fig. 3-1 for a map of Semitic dispersals).

Our Semitic tree indicates that within several centuries of the initial divergence of ancestral Semitic into East Semitic (represented in our study by Akkadian) and West Semitic branches, West Semitic in turn diverged ~5425 ybp (Fig. 3-2, node **A**) to form Central and South Semitic. The short interval between this divergence and the origin of Semitic in Syria (~500 years) suggests that this divergence occurred in the same region of the interior Levant, and that Central Semitic (found throughout the Levant) spread to the west while South Semitic spread into the southern Levant. Thus, the early emergence ~4525 ybp (Fig. 3-2, node **D**) of a South Arabian lineage of South Semitic may reflect an early Bronze Age expansion of Semites from the southern Levant southward through the Arabian Peninsula. The Central Semitic sub-branch of West Semitic was characterized first by a divergence into northern (Ugaritic) and southern

(Arabic, Aramaic, and Hebrew) lineages ~4100 ybp (Fig. 3-2, node **B**), and later by the divergence of Arabic, Aramaic and Hebrew from each other ~3200 ybp (Fig. 3-2). The expansion of Central Semitic ~4100 ybp was likely part of the migration process that was definitive of the transition from the Early to the Middle Bronze Age in the Levant (Ehrich 1992; Ilan 2003; Richard 2003b). This period in the Levant involved the devolution of many urban societies at the tail end of the Early Bronze Age (Richard 2003b) and their replacement with new urban societies that were culturally and morphologically distinct at the start of the Middle Bronze Age (Ilan 2003). Our analysis suggests that the shift in urban populations from the Early to Middle Bronze Age is associated with the wider expansion of Semitic in the Levant.

The recurrent spread of early Semitic peoples and their languages, first South Semitic and later Arabic, into the marginal, desert lands of the Arabian Peninsula, combined with the Biblical testimony on early Hebrew subsistence, suggest that the earliest West Semitic society had a largely pastoralist economy particularly adapted to such conditions. Furthermore, this is consistent with the ancestral Semitic society occupying the easternmost Early Bronze Age urban developments in the Levant. This placement provides a plausible setting for the initial movement of Akkadian into Mesopotamia, as well as the subsequent expansion of South Semitic through the Arabian Peninsula and Central Semitic throughout the Levant. The expansion of Arabic into the Arabian Peninsula is the result of a more recent migration of Central Semitic southward from the Levant.

Recent Arabic Divergence

The Arabic languages, or dialects, represent the largest group of extant Central Semitic languages (Gordon 2005). The Arabic languages originated in north Arabia and expanded along with Islam in the 7th century to occupy a geographic range that extends from Morocco to Iran

(Kaye and Rosenhouse 1997). Our phylogenetic analysis indicated that the two studied Arabic languages (Moroccan and Ogaden) diverged ~540 ybp (Fig. 3-2, node **C**), or ~800 years after the expansion of Arab populations associated with Islam. We were able to employ a log BF test to asses the ability of our linguistic clock to reject unrealistic historical scenarios by comparing the effects of different constraints on the divergence times of these Arabic languages. Specifically, we compared our Standard model (Arabic divergence date estimate = ~540 ybp) with a model that constrained the divergence of Arabic to occur prior to the spread of Islam (i.e., > 1400 ybp), and were able to show "strong" support for the Standard model over the model with unrealistic constraints. This test confirms the accuracy of our linguistic clock, while suggesting that in some regions, such as Morocco, Arabic languages became fully established as local indigenous tongues, replacing earlier indigenous languages (Berber in this case), only within the past millennium, long after the initial expansion of Islam.

South Semitic and the Origins of Ethiosemitic

We estimate that after the divergence of West Semitic into Central and South Semitic ~5425 ybp, South Semitic continued to expand southward until ~4525 ybp (Fig. 3-2, node **D**). It was at this time that South Semitic diverged into two lineages with to the south of the Levant. One of these lineages was ancestral to the Modern South Arabian languages, and its speakers likely inhabited the southern coasts and coastal hinterlands of the peninsula. The other lineage leads to the 'Epigraphic' South Arabian language(s), spoken in and around highland Yemen in the later second and the first millennia B.C.E. The founding speakers of Ethiosemitic in the far northern Ethiopian highlands likely came from this Yemeni population. Our estimate of ~2975 ybp for this transfer of Semitic to Eritrea and Ethiopia (which became the Ethiosemitic languages) is contemporaneous with the adoption of Iron Age technologies in the Middle East

(Ehrich 1992; Moorey 1994), but pre-dates the rise of the Aksumite Kingdom in Eritrea and Ethiopia by at least ~600 years (Connah 2001). This migration to Africa most likely reflects a influx of Semitic speaking migrants of unknown size, concurrent with the rise of the first towns and cities in the northern edges of the Ethiopian Highlands during or before the middle of the first millennium B.C.E. (Ehret 1988), and suggests that the introduction of Semitic from Arabia was temporally correlated with the development of the first urbanized states in Eritrea and Ethiopia (Fattovich 1990).

Intriguingly, the estimated date for the Ethiosemitic migration is broadly consistent with the Queen of Sheba (Saba') myth, which posits a 10^{th} century B.C.E. connection of the Ethiopian Highlands and Yemen with the Levant. This myth claims that after visiting King Solomon in Israel, the Queen of Sheba returned home to bear a son who would later found a new state in Eritrea and Ethiopia culturally attached to Semitic populations in the Levant (specifically the Hebrews). Genetic studies have shown that Ethiosemitic speaking populations are genetically similar to Cushitic-speaking populations within Eritrea and Ethiopia (Lovell et al. 2005), indicating that the migration of Semitic to the Horn of Africa was accomplished with little gene flow from the Arabian Peninsula. This process of local adoption, in which Ethiosemitic was first introduced to and later adopted by individuals recruited from existing local populations is also consistent with the word borrowing evidence (Ehret 1988) that a relatively small number of people from Yemen, holding economically and politically strategic positions, introduced Semitic to populations in the Horn of Africa.

Conclusion

We used Bayesian phylogenetic methods to elucidate the relationship and divergence dates of Semitic languages, which we then related to the archaeological and epigraphic record to

produce a comprehensive hypothesis of Semitic origins and dispersals (Fig. 3-1). Our analysis demonstrates the first time that language sampling dates have been used to calibrate the mean rate of linguistic evolution, including variation between lineages. This allowed us to provide dates for important events in Semitic history and place them in context. For example, we estimate that (i) Semitic had an Early Bronze Age origin (~5900 ybp) in the dry interior areas of the Levant from which Akkadian subsequently expanded into Mesopotamia, (ii) Semitic then dispersed, earlier than previously thought, throughout the Levant as part of the Early to Middle Bronze Age transition in the Eastern Mediterranean, and (iii) conclude that Ethiosemitic was the result of a single, early Iron Age (~2975 ybp) migration of Semitic across the Red Sea consistent with 10th century B.C.E. Queen of Sheba myths connecting Ethiopia to non-African Semitic populations. Furthermore, we employed the first use of Bayes factors to statistically test competing language histories and confirm the robustness of our inferences about Semitic history. These inferences shed light upon the complex history of Semitic, answer key questions about Semitic origins and dispersals, and provide important hypotheses to test in future studies with new data.

Methods

Word Lists and Cognate Coding

Wordlists were modified from Swadesh's 100-word list of most conserved words (Swadesh 1955), with the final lists containing 97 words for 25 extant and extinct Semitic languages (Supplementary Figure 3-S1). Wordlists for the Ethiosemitic languages (Amharic, Argobba, Chaha, Gafat, Ge'ez, Geto, Harari, Innemor, Mesmes, Mesqan, Soddo, Tigre, Tigrinya, Walani, Zway) and Ogaden Arabic were drawn from Bender's *The Languages of Ethiopia* (Bender 1971). The lists for Moroccan Arabic, the South Arabian languages (Gibbali, Harsusi,

Mehri, Soqotri), and extinct non-African Semitic languages (Akkadian, Aramaic, ancient Hebrew, and Ugaritic) were constructed from previously published lexicons (Rabin 1975).

Cognate classes were determined for each of the 97 words using a comparative method that emphasizes the similarity of consonant-consonant-consonant roots and known consonant shifts when comparing two words (see Supplemental Figure 3-S2 for a graphical depiction of the cognate coding and loanword treatment). Loanwords were identified using lexical information from distantly related but geographically close language families (such as Cushitic) as well as comparisons with lexicons of languages within the Semitic family. Loanwords were dealt with in two ways: (i) all loanwords were coded as missing data (i.e., '-') to indicate the loss of the ancestral word in the accepting language(s), and (ii) loanwords shared by descent (i.e., the loan event occurred in a common ancestor of multiple languages) were considered to be an extra meaning in which all languages sharing the loanword via descent are coded as cognate (e.g., 'A') while all others are coded as missing data (i.e., '-'). In the second case, the loanword was used as a marker of common descent in languages that diverged after the loan event, while no common ancestry for that meaning was supposed between the language(s) that provided the loanword and the language(s) that accepted the loanword. The coded cognate dataset, including the treated loanwords, is available as Supplemental Figure 3-S3.

Phylogenetic Analysis and Divergence Date Estimation

Phylogenetic trees were constructed under a Bayesian framework using BEAST v1.4.6 (Drummond and Rambaut 2007). An unordered model of cognate class evolution was used to allow transitions between any pair of cognate classes. Rate heterogeneity across meanings was modeled by a gamma distribution on meaning-specific rates. This model accommodates variations in the rate of change across meanings, such that conserved meanings (e.g., a single

cognate class for all languages) were assigned a slower rate than the mean while highly variable meanings (e.g., few shared cognate classes between languages) were assigned a faster rate than the mean. Priors for the unordered rate matrix and gamma shape parameter were flat.

Divergence times were estimated using an uncorrelated lognormal relaxed clock model that assumes a single underlying rate for the entire phylogeny but allows for variations in rates across the tree (Drummond et al. 2006). This relaxed clock model accommodated differences in the overall rate of cognate transition between languages by assigning rates drawn from a lognormal distribution to individual branches in the tree. We used heterogeneous tip date information (Rambaut 2000) and constraints on the root of Semitic to calibrate the relaxed clock. Specifically, we included the sampling dates of the five languages in our dataset that are no longer spoken [Akkadian = 2800 ybp, Aramaic = 1800 ybp, Ge'ez = 1700 ybp, ancient Hebrew = 2600 ybp, and Ugaritic = 3400 ybp; (Rabin 1975)] and used a flat prior on the root of the Semitic set to 4,300 [the age of the first evidence of a Semitic language, Akkadian (Buccellati 1997)] to 8,000 ybp to calibrate the clock. (Although the earliest few documents in Akkadian date to ~4300 ybp, the great bulk of Akkadian lexical data comes from late Assyrian materials, 2900-2700 ybp, hence the choice of 2800 ybp for the sampling date of Akkadian.) A flat prior of 0.01 to 0.0001 cognate transitions per meaning per year (roughly, a 0.01% to 1% replacement rate per year) was placed on the mean of the lognormal distributed clock.

The robustness of our results was tested by the application of log BF model tests to comparisons of alternative Semitic histories. In addition to the analysis with no topological constraints described above (the unconstrained model), we first constrained the tree topology to reflect the accepted major divisions of Semitic: East Semitic (Akkadian) vs. West Semitic, with West Semitic into Central (the Arabic languages, Aramaic, ancient Hebrew, and Ugaritic) and

South (the Ethiosemitic and South Arabian languages) Semitic clades (Faber 1997). We called this Semitic history the Standard model, as it represents a general consensus amongst Semitic linguistisc that the primary division of Semitic is between East and West clades. For our second test, we compared the Standard model with a model (termed the 'Old Arabic' model) that constrained the divergence of Moroccan and Ogaden Arabic to occur more than 1400 ybp. This comparison was used to test the accuracy of our relaxed clock model, as the Old Arabic model forces the divergence of the two Arabic languages to have occurred unrealistically before the expansion of Arabic with Islam starting in the 7^{th} century. Our final test compared the Standard model with a model (termed the 'African' model) that forced Semitic to have an African root by constraining Ethiosemitic as the outgroup. The mean clock rate estimated under the Standard model (4.438×10^{-4} transitions per meaning per year) was used for all log BF tests, and Markov chains were run for 10^8 generations. Marginal likelihoods for each model were estimated using the smoothed harmonic mean of the likelihood distribution (Newton et al. 1994; Redelings and Suchard 2005) and all log BF values were calculated by taking the difference in the log of the marginal likelihoods of each model (Kass and Raftery 1995), with log BF values reported in log base 10 units. A log BF > 0 would indicate a preference for model 1, whereas a log BF < 0 would indicate a preference for model 2.

BEAST uses a Markov chain Monte Carlo simulation technique to estimate the posterior distribution of parameters. All Markov chains were run for 20,000,000 generations with samples taken every 1,000 generations (for a total of 20,000 sampled states per run). Burn-in was 1,000,000 generations (1,000 states), and post run analysis of parameter plots in Tracer v1.4 (Rambaut and Drummond 2007) suggested all chains had reached convergence by the end of the

burn-in period (e.g., all estimated sample size values were > 500). The MCMC sampling and run conditions, and all prior distributions, were identical for all analyses unless otherwise stated.

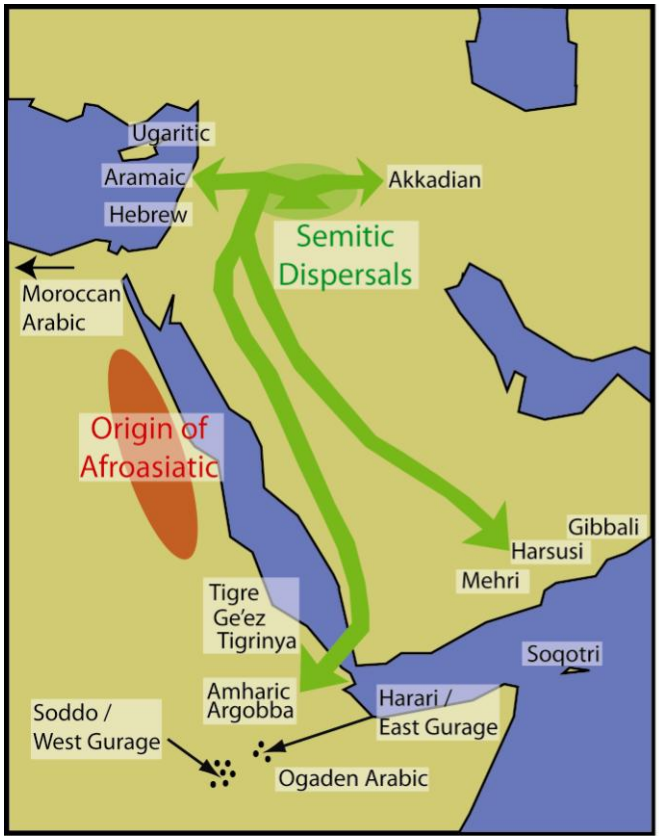

Figure 3-1. Map of Semitic languages and inferred dispersals. The location of all languages sampled in this study are depicted. The map also presents the dispersal of Semitic inferred from our study. The origin of Afroasiatic along the African coast of the Red Sea is indicated in red (Ehret 1995; Ehret, Keita, and Newman 2004), while Semitic migrations are depicted by green arrows. The Semitic dispersal follows a radial pattern expanding from a Semitic origin in eastern Syria. The current distributions of all Semitic languages in Eritrea and Ethiopia (Ethiosemitic) follow Bender (Bender 1971) and the remaining follow Hetzron (Hetzron 1997). The ancient distributions of the extinct languages are indicated, i.e., Akkadian, Aramaic, Ge'ez, Hebrew, and Ugaritic (Bender 1971; Hetzron 1997). The West Gurage (Chaha, Geto, Innemor, Mesmes, and Mesqan) and East Gurage (Walani and Zway) Ethiosemitic language groups in central Ethiopia are listed as two combined groups since the distribution of these languages was too circumscribed to list individually in the map.

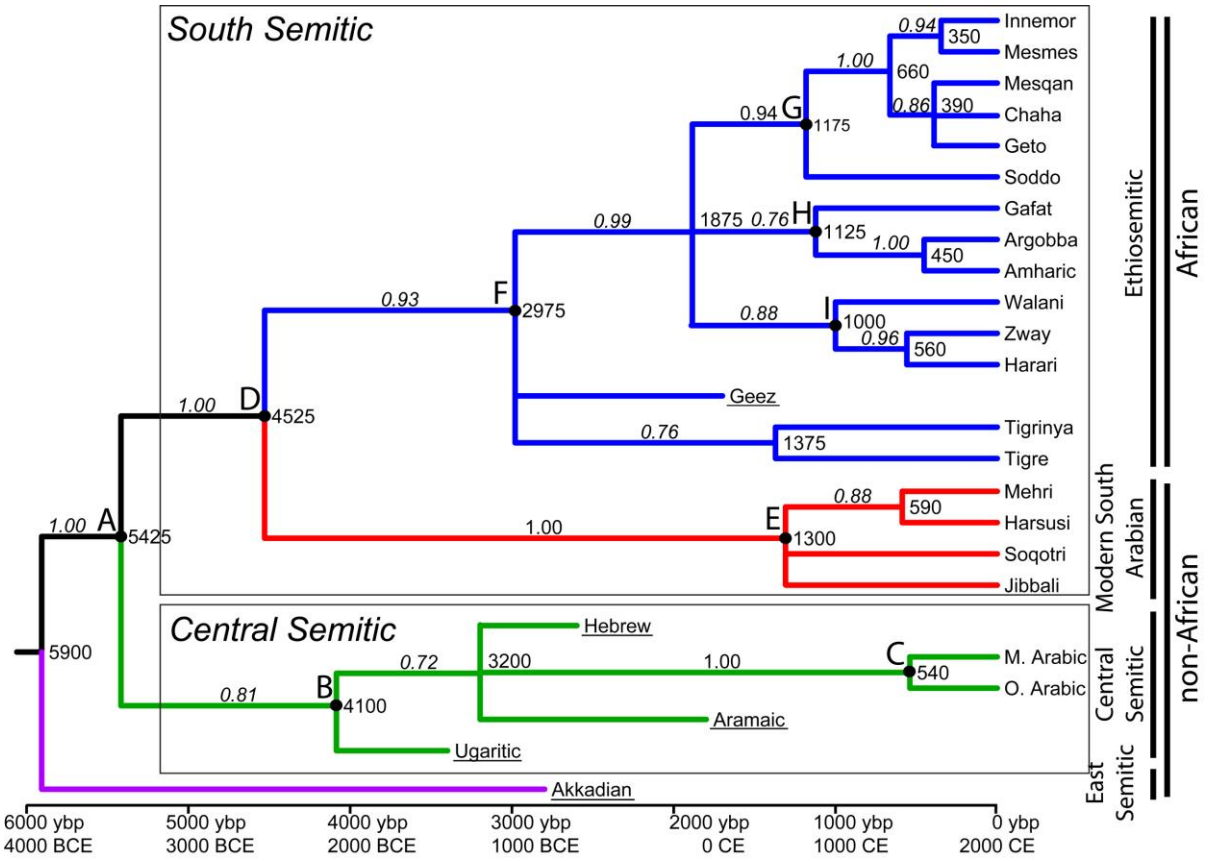

Figure 3-2. Phylogenetic tree of Semitic languages. The language tree produced by the phylogenetic analysis of Semitic wordlists placed the origin of Semitic at ~5900 ybp (CI = 4300 – 7700 ybp). All branch lengths are in years, represented by median estimates from their posterior distributions, and node dates are relative to years before present (ybp). The scale bar along the bottom of the tree presents dates in ybp and relative to the common era (C.E. and B.C.E.). Extinct languages are under-lined, and all other languages are considered to evolve to the present. Sub-groups of Semitic are identified by color (East Semitic = purple, Central Semitic = green, Modern South Arabian = red, and Ethiosemitic = blue), with bars to the right of the tree, and by two boxes (Central Semitic and South Semitic), while the geographic distribution (African vs. non-Africa) is indicated by a bar to the right of the tree. Important groups are indicated by letters **A-I**. These are: **A**, West Semitic; **B**, Central Semitic; **C**, Arabic; **D**, South Semitic; **E**, Modern South Arabian; **F**, Ethiosemitic; **G**, Outer Gurage; **H**, East Gurage; and **I**, Amharic-Argobba-Gafat. Posterior probabilities of internal branches are printed in italics above each branch, with median divergence dates printed to the right of each node. All branches with posterior support < 0.70 are collapsed and considered unresolved. The topology follows the constraints of the Standard model, which is preferred by our log BF analysis.

Gloss	Ge'ez	Tigre	Tigrinya	Amharic	Argobba	Harari	Zway	Walani	Gafat
All	kʷɬllu	kɬllu	kullu	hullu	diyyu	kulluɔlm	hullɛn	ullɔmka	yɬlho
Ashes	ħamɛd	baʾrəħ	ħamuxwusti	amɛd	ħamɛd	ħamɛd	hamɛd	amɛd	--
Bark (tree)	lɛhɛʾ	kʾaraf	lɛħəʾɔ	lɛtʾ	litʾɔ	litʾ	kʾɛlncʾɛ	bɔssɛ	--
Belly	7ɛngɛdɛ7a	kabɛd	kɛbdi	hod	kɛrɛ	kʾɛrfit	dɛl	dɛl	ɛrɛɛ
Big	9abiyy	9abi	9abiyyi	tɛllɛkʾ	nɛham	gidiːr	gɔidiːr	ɛgɛdɛrɛ	ɛmmunɛ
Bird	9ɔf	sarɛːrat	9uf	wɛf	(w)ɔf	7uːf	ɔːmf	unf	yɛfʷɛ
Bite (v.)	nɛkkɛɛɛ	nakɛa	nɛxɛɛɛ	nɛkkɛɛɛ	nɛkkɛɛɛ	nɛxɛɛɛ	nɛkkɛɛɛ	--	nɛkkɛɛɛ
Black	sʾɛllim	sʾallɛn	sʾɛllim	tʾɛkʾʷur	tʾħɛkʾ(ʷ)ur	tʾɛy	gɛmbɛlla	tʾɛm	tʾɛkʾurɛ
Blood	dɛm	dam	dɛm	dɛm	dɛm	dɛm	dɛm	--	dɛmʷħɛ
Bone	9asʾɛm	9asʾim	9atʾɛmi	atʾɛnt	hatʾɛm	aːtʾ	hatʾɛm	atʾɛm	asʾmʷɛ
Breast	tʾɛbb	tʾɛb	tʾub	tʾut	tʾut	tʾu7t	tʾuːb	tʾub	tʾuwwɛ
Burn (tr. v.)	7awɛ9ayɛ	7andada	akʾasʾɛlɛ	akʾatʾtʾɛlɛ	akʾaːtʾɛli	maːʾgɛdɛ	anɛddɔt	--	(tɛ-)kʾatʾtʾlɛ
Cloud	dɛmmɛna	dabnaː	dɛbɛna	dɛmmɛna	dɛmmɛna	daːna	daːbana	dɛvɛnɛ	dɛmmɛnɛ
Cold (air)	bɛrɛd	bɛrɛd	kʾurri	bɛrd	bɛrd	biːrdi	bɛrd	bɛrd	wɛrcʾɛ
Come	mɔsʾɛ7a	masʾa	mɔsʾ7ɛ	mɔtʾtʾa	mɔtʾtʾa	diːjɛ	mɔtʾtʾaːn	--	sɛllɛ
Die	mɔtɛ	moːta	mɔtɛ	mɔt	moːdva	moːtɛ	moːtɛn	--	fɛttɛrɛ
Dog	kɛlb	kalɛb	kɛlbi	wɛssa	wussa	buːci	gɛnni	buco	wussɛ
Drink (v.)	sɛtɛyɛ	sata	sɛtɛyɛ	tʾɛtʾtʾa	ɛscca	sɛːcɛ	sɛːcɛin	--	tʾɛtʾtʾa
Dry (adj.)	yɛbɛɛ	yɛbuɛ	nɛxʾaɛ	dɛrɛkʾ	dɛrɛkʾ	dɛrɛkʾ	dɛrɛkʾ	dɛrɛkʾ	dɛrɛk
Ear	7ɛzɛn	7ɛzɛn	7ɛzni	joro	joro	7uʾzun	ɛzzɛn	ɛzɛn	ɛznɛ
Eat	bɛl9a	bal9a	bɛl9ɛ	bɛlla	bɛlla	bɛlɛ7a	bɛllvaːn	--	bɛllɛ
Egg	7ɛnkʾɔ(ɲ)kʾɛħa	7ɛinkʾɔkʾħa	7ɛnkʾulaliħ	ɛnkʾʷulal	ɛnkʾulal	7akʾuːh	ɛnkʾɛlk	ɛnkʾɛlkʾɔt	ankʾʷɛ
Eye	9ɛyn	9ɛn	9ayni	ayn	ɛːn	7iːn	iːn	in	inɛ
Fat (n.)	sɛbɛħ	sɛbɛħ	sɛbbi	cʾɔmma	cʾɔːmma	lɛzɛ	cʾɔmma	cʾɔmɛ	bussɛrɛ
Fire	7ɛsɛt	7ɛsaːt	ħawwi	ɛsɛt	ɛssaːd	7iːʾsat	jiːra	jira	isatɛ
Fish (n.)	9asɛ	9aːsa	9asa	asa	asaː	tuːlɛm	tʷulʷuːm	asa	asɛ
Fly (n.)	zɛnb	cʾɛncʾaːy	sʾɛnsʾɛya	zɛmb	zɛnb	ziːnbi	zɛnb	--	zinbɛ
Foot	7ɛgɛr	7ɛgɛr	7ɛgri	ɛgɛr	ɛngɛr	7iːgir	ɛngɛr	ɛngɛr	igʷrɛ
Give	mɛtʾɛwɛ	haba	habɛ	ɛtʾtʾɛ	haːbɛ	ɛtʾɛ	aːbɛːnʷa	--	wabɛ
Go	ħorɛ	geːza	kɛdɛ	hɛdɛ	hɛːda	lɛːtʾɛ	hidan	--	(a)horɛ
Good	sɛnnay	gurrum	sʾɛbbuxʾ	tʾɛru	tʾɛra-heː	bɛrɛħ	bɛːza	tʾuri	gunnɛ
Grass	ħamɛlnal	sa9ar	sa9ri	sar	sir	saːr	siːr	sar	sar
Hair (head)	sʾɛgur	cʾɛgar	sʾɛgʷri	tʾɛgur	cʾɛgɛr	cʾiːgɛr	tʾɛgɛr	dumi	sʾigɛrɛ
Hand	7ɛd	7ɛdeː	7id	ɛjj	ɛnji	7iji	ɛnji	7ɛnj	ɛdjɛdjɛ
Head	rɛ7ɛs	ra7ɛs	rɛ7ɛi	ras	dɛmmaː	7urus	ohaːt	dumi	dɛmʷɛ
Hear	sɛm9a	sam9a	sɛm9ɛ	sɛmma	sɛmma	sɛma7a	samma	--	sɛmmɛ
Heart	lɛbb	lɛb	lɛbbi	lɛbb	lɛbb	kʾɛlbi	wɛzɛːna	wɛzɛn	--
Horn	kʾɛrn	kʾar	kʾɛrni	kʾɛnd	kɛnd	kʾɛr	kʾɛr	kʾɛrr	kʾɛndɛ
I	7ɛnɛ	7anaː	anɛ	ɛnɛ	ɛyyɔ	7an	ɛya	ɛhɛ	anɛt
Kill	kʾɛtɛlɛ	kʾatla	kʾɛtɛlɛ	gɛddɛlɛ	gɛddɛli	gɛʾdɛlɛ	kʾɛɔciːn	--	--

Figure 3-S1. Semitic wordlist data for 25 languages. The final Semitic wordlists contained 97 meanings for 25 extant and ‘extinct’ languages. Data for 15 Ethiosemitic languages [Amharic, Argobba, Chaha, Gafat, Geto, Ge’ez, Harari, Innemor, Mesmes, Mesqan, Soddo, Tigre, Tigrinya, Walani, and Zway; (Bender 1971)], Ogaden Arabic (Bender 1971), and 9 non-African Semitic languages [Akkadian, Moroccan Arabic, Aramaic, Gibbali, Harsusi, Hebrew, Mehri, Socotri, and Ugaritic; (Rabin 1975)] were gathered from previously published wordlists. Gaps in the data are indicated by ‘-‘ characters.

Gloss	Ge'ez	Tigre	Tigrinya	Amharic	Argobba	Harari	Zway	Walani	Gafat
Know	lɨbbəwɛ	7alməra	fələt'ə	awək'ɛ	wɔlk'a	7aʲk'ɛ	ca:lɛn	--	ɨalɛ
Knee	bɨrk	—	bɨrki	gulbət	gulbət	gɨlɨb	gulbət	bɨrk	--
Leaf	k'ʷɔɛ'ɨl	k'at'af	k'ʷɛɛ'li	k'it'əl	k'ət'əl	k'ut'ɛ	k'ɨt'ɛ́y	k'ut'əl	katalɨɛ
Liver	kəbd	kabdat	t'afiya	gubbət	amfaha	kuld	gubu:t	kɛbt	--
Long	nɛwwix	rayyim	nɛwwiɦ	rəjjɨm	rɛzzɨn	gudurr	guder	guder	gɛddɛrmɛ
Louse	k'ɨmal	k'ɨmal	k'umal	k'ɨmal	k'ɨm(n)al	k'umaɣ	k'ɨmáɣ	k'umal	kɛmalɨɛ
Man	sɛ7ɨb	7addaɨm	sɛb	sɛw	su:	7usu7	sɛb	sɛb	sɛwwɛ
Many	bəzux	bəzuɦ	bəzuɦ	bəzu	bəzzu	bəjiɦ	bəzzɨ(n)	bɛzi	tɛbbɛ
Meat	sɨgɛ	sɨga:	sɨga	sɨga	jo	bɛsɛr	bɛsɛrr	bɛsɛr	brandɨɛ
Moon	wərɨɦ	warɨɦ	wərɦi	c'ərək'a	c'ərək'a	c'ɛ'rɛk'ɛ	t'ɨrk'a	weri	ɛ'ərək'ɛ
Mountain	dəbər	dabər	7ɨmba	tərara	gubb	ɛɛri	ɛɛri	--	amba
Mouth	7[illegible]	7af	7af	af	af	7ɔf	af	af	sɨmɔtɛ
Name	sɨm	sɨm	sɨm	sɨm	sɨm	sum	sɨm	sum	sɨmʷɛ
Neck	kɨsad	ɛ'ɨgal	kɨsad	angət	angɛd	7angɛt	angət	angət	angɛt
New	ɦaddɨs	ɦadɨs	ɦaddɨs	addɨs	haggɨs	ɦajɨs	wɔlyra	ajɨs	addɛɛs
Night	lɛlit	la:li	lɛyti	lɛlit	ɛ:ta7	7ɔrti	aru:t	arɔt	lilɛ
Nose	7anf	7a:nɨf	afɨnc'a	afɨnc'a	affɨnc(c)a	7u:f	nɨ:f(ɨ)t	barbari	a:fʷɛ
One	7aɦad	wɔ:ro	ɦadɛ	and	ha:nd	7aɦad	ha:d	7ɛdd	ɨjjɛ
Rain (n.)	zɨnam	zɦɨla:m	may	zɨnab	zɨnab	zɨnaɓ	zɨnab	zɨla:m	zɨnab
Red	k'ɛyyɨɦ	k'ayyɨɦ	k'ɛyyɨɦ	k'ɛy	k'ɛyyɨɦ	k'ɛɦ	arɔa	--	k'ɛyɛ
Road	fɨnɔt	gabay	mɛngɛddi	mɛngɛd	hɛ:mma	7uga	ɔ:nga	ungɛ	managi-ɛ
Root	sɨrwu	k'ɨrɨd	sur	sɨr	sɨrrɔ'd	sirr	sɨr	sɨr	wadɛrɛ
Sand	xɔɛ'a	ɦɔ:'a:	ɦɔɛ'a	asɛwa	haɛɛwa:	sɨɦ'ɛɛra	ɛrɛɛrɛ	asɛwɛ	--
Say	bəɦɨlɛ	bɛ:la	bɛlɛ	alɛ	a:la	baɣɛ	bal	--	balɛ
See	rɨ7ɨyɛ	ra7a	rɛ7ayɛ	ayyɛ	haɪnja	ri7a	iri:n(ɔ)	--	ajjɛ
Seed	zɛr7ɨ	zɨrɨ7	zɛr?	zɛr	zɛri	ɛɛni	ɛɛnni	zɛr	zɛr
Sit	nəbbərɛ	kaf bɛ:la	kɔf bɛlɛ	tək'ɛmmət'ɛ	a:rrɛggaza	tɛgɛɓɛla	tɛgu:bbi	--	tɛwannɛ
Skin (human)	ma7ɨs	k'arbat	k'ɔrbət	k'ɔda	k'ɔda	gɔ:ga	gu:ga	gɔga	g'ag'ɛ
Sleep (v.)	nɔmɛ	sakba	dɛk'k'ɛɛɛ	tɛnna	tanni	nɛ:7a	tɛnni	--	(tə-)gɛddɛlɛ
Small	nɨ7uɛ	nu7uɛ	nɨ7ɨtɔ(ɣ)	tɨnnɨs	mɛɛɛt	t'i'k'ɛɔ	t'i'ʷt	furt'	ɨnɛɛ
Smoke (n.)	t'ɨs	tanaɨn	tɨkki	t'ɨs	c'ɨ:s	tɛn	ta:n	tɛn	tɛnnɛ
Snake	ɛarwɛ	7arwɛ:	tɛmɛn	ɨbab	hawa:w	ɦubaɓ	ɨmbaɓ	--	ɨuau'ɛ
Stand	k'ɔmɛ	bat'ra	dɛw bɛlɛ	k'ɔmɛ	k'ɔma	k'a:'nɛnɛ	k'a:na	--	k'ɔmɛ
Star	kɔkɛb	kɔ́kab	kɔxɛb	kɔkɛb	kɔkɔ́b	t'u:n	kɔkɛb	kɔkɔb	kɔkɔbɛ
Stone	7ɨbɛn	7ɨbɛn	7ɨnni	dɨngay	dɨngaɣ	7un	umu:n	7un	dɛnja
Sun	ɛ'ɨhay	ɛ'ɨhaɣ	ɛ'ɨhay	ɛ'ɨhay	c'ɨhɛd	7irr	ari:t	air	sɛgɛlɛ
Swim	tɛ-wanɛyɛ	ɦambaɛa	ɦambɛɛɛ	wannɛ	wanna	tɛfu'kaɨka	waɨkan	--	--
Tail	zɛnɛb	zanab	c'ɨra	jɨrat	zɛri	k'ɛ'nɛwɛ	gu:nna	gɔn	c'ɨrɛ
Thin	k'ɛt'ɨn	k'at'ɨn	k'ɛt't'ɨn	k'ɛc'c'ɨn	k'ɛc'c'ɨn	k'ɛ'c'i:n	k'ɛç'ɨn	--	kacɨn
This	zɨntu	7ɨlli	ɨzzuy	yɨh	hul	yi	yi:	--	ɨnnɨ

Figure 3-S1 continued.

Gloss	Ge'ez	Tigre	Tigrinya	Amharic	Argobba	Harari	Zway	Walani	Gafat
Three	ʂʂlʂstu	ʂalaʂ	ʂrlʂstʂ	ʂoʂt	ʂoʂt	ʂi7isti	ʂʂlʂt	ʂeʂt	ʂ"ʊstʂ
Thou	7antʂ	7ʖntə	nʖʂʂʖxə	antʂ	anka	7ahaʔh	atta	atʂ	ant(ʂ)
Tongue	lʂʂan	nʖʂʂʔl	mʔlhʂʂ	mʔlaʂ	mʔlaʔʂ	arrat	aramat	arimʂt	mʔlaʂʂ
Tooth	ʂʖn	7anjaʔb	ʂʖnni	t'ʖrʂ	ʂʖnn	ʂiʔn	ʂʖn	ʖʂʖn	ʂʖnʂ
Tree	9ʖʂ'ʂ	9ʖc'ʂ'at	7ʊm	zaf	zaf	lʂfu	ʖnt'ʊt	ʖnʂ'i	zaf"ʂ
Two	kʔlʔ7ʊttu	kʔl7ʊ:t	kʔlttʂ	hulʂtt	kʔ:t	koʔʊt	hʊ:yt	hʊit	ʔlʔʂtʂ
Warm	mʖwukˈ	ʰifun	mʖwuxˈ	mukˈ	mukˈ	wʊ:y	mukˈ	mukˈ	mʊkˈʂ
Water	may	mʔ:y	may	wuha	ʊhuwwa:	mʔ:y	mʔ:y	mʊ	ʂgʂ
We	nʔhanʂ	ʰʔnʔ:	nʔhna	ʔnna	ʔnna	7innaʔʂ	ʔnnʂ	ʔnnʊ	ʔnni
Wet	rʊtˈub	tˈulul	rʔhuʂ	(ʔ)rtˈʔb	ʔrtˈʔb	ruʔtˈʂ	hurtˈuʔb	--	jibbʂ
What?	mʔnt	mi	7ʔntay	mʔn	mʔnn	miʔn	mʔn	mʔn	--
White	ʂˈʂadʂ	ʂˈʂ9ada	ʂˈʂ9da	nʂʂˈʂˈ	zahh	nʂʂˈiʔh	gʊ:ˈmara	gumʂrʂ	nʔʂˈwʂ
Who?	mʊnnu	man	mʊngʊddi	man	manna	mʊn	maʔnnu	ma	mʔn
Woman	bʔ7ʔʂʂt	7ʔʂʂit	ʂrbʊˈ7ti	ʂ"ʊt	nʔʂʂa	kˈʊhaʔt	sʊ:t	mʔst	ʔmnʔitʂ
Ye	7antʔmu	7ʔntum	nʔʂʂʔxatkum	ʔnnantʂ	ʔnnakum	7axaxaʔʂ	atum	atum	ʔnnantʂ

Figure 3-S1 continued.

Gloss	Soddo	Mesqan	Innemor	Mesmes	Geto	Chaha	Hebrew	Ugaritic	Aramaic
All	kullɨmu	ɨnɛmɔ	ɨnnɨ	ɔt'tɛmɨ	ɨnnɨn	ɨnnɨn	kol	kl	kul
Ashes	amɛd	amed	amɛnd	hawɛnda	amɛd	amɛd	epher	nr	k'ef'n'/rnu'a
Bark (tree)	k'erfit	milɛc'ɛ	xaɨra	haɨnna	xara	xara	k'elip:ah	k'lapthaʔ	k'irfah
Belly	kerɛ	den	kɛɛ	kɔɛɛa	dɛn	dɛn	bet'en	kbd	karɛa
Big	maɨlɛk'	wɛdɛl	nu7	k'ɔk'ɔ	nɨk	nɨk	gadhol	gdl/rb/mid	rabba
Bird	wɔf	af	anfw	ɔɨnfa	anfw	anfw	s'ippor	ɛr	ɛeppra
Bite (v.)	nɛkkɛɛɛ-n	nɛkɛɛɛ	nɛkɛɛɛ	nakɔ-ɛe	nɛkɛɛɛ-n	nɛkɛɛɛ-n	naʃakh	ntk	nkhath
Black	t'ɨk'ur	t'ɛk'ʷɔr	du7ur	gɔmbɔnna	t'ɨk'ur	t'ɨk'ur	ʃahor	---	uk²ama
Blood	dɛn	dʸɛn	dɛn	dɔ	dɛn	dɛn	dam	dm	dma
Bone	at'ɨm	at'ɨm	a7ɨm	haʷwa	at'ɨn	at'ɨm	ɛɛ'ɛn	zm	garma
Breast	t'ɨbuiya	--	t'u	t'uʷiyɛ	t'u	--	hagah	---	hadhya
Burn (tr.v.)	anɛddɛdɛ-n	ak'at'ɛlɛ	nɛkʸɛrɛ	totɔɛɛ-yɛ	nɛkʸɛrɛ-n	nɛkʸɛrɛ-n	ʃaraph	ʃrp	auk'ɛdh
Cloud	dɛmmɛna	dabɛna	dɛnɛra	dɔɨna	dɛmɛra	dɛmɛra	anan	rpt/grpt	nana/arpel²a
Cold (air)	bɛrrɛdɨ-n	anɛdar	anɛdaɨr	ziɨ:-ɛ	anɛdar	anɛdar	k'ar	---	k'ar²ira
Come	nɛɨt'ɨ-n	cɛnɛ	na7a	-nna7a-yɛ	cɛnɛ-n	cɛnɛ-n	ba'	bw'/'tw	etha/nt'a
Die	nɔɨtɛ-n	nʷɔtɛ	nɔɨdɛ	nɔtɔ-	nɔtɛ-n	nɔtɛ-n	meth	mwt	mith
Dog	wuɛɛa	gɨɨyɛ	gɨyɛ	gɨɨyɛ	gɨyɛ	gɨyɛ	kɛlɛbh	klb	kalba
Drink (v.)	ɛɛc'ɨc'ɛ-n	ɛɛc'ɛ(y)	ɛɛc'ɛ	ɛɛ'cɔ-	ɛɛc'ɛ-n	ɛɛc'ɛ-n	ʃathah	ʃ t y	eʃty
Dry (adj.)	dɛrɛk'	t'ɛrɛk'	dɛrɛ7	dɛ'rɔ-ɛ	ɛ'ɛrɛk'	t'ɛrɛk	yabbeʃ	h r b	yabbaɛ
Ear	ɨnzɨn	ɨnzɨn	ɨnzɨr	unˈzuɨra	ɨnzɨr	ɨnzɨr	ozen	u d n	ɛdhna
Eat	bɛlla-	bɛna	bɛn7a	ba7'nɔ'-	bɛna-n	bɛna-n	akhal	k l	ɛkhal
Egg	ank'ɔ	ank'ʷa	ɨnk'ura	k'u'ra	ɨnk'ura	ɨnk'ura	bɔ̈ɣɛah	---	bey'tha
Eye	ɨɨn	ɛʸn	ɛr	ɨɨn	ɛn	ɛn	ayin	'n	aina
Fat (n.)	c'ɔɨma	--	mazɛ	c'ɔmma	fuk'ɨr	manzɛ	deʃɛn	---	meʃha
Fire	wɛat	ɨɛat	ɨɛaɨd	ɨ'ɛaɨdɛ	ɨɛat	ɨɛat	ɛh	iʃ t	nura
Fish (n.)	aɛa	aɛa	aɛa	--	aɛa	aɛa	dagh	d g	nuna
Fly (n.)	zɨmb	zɨmb	zɨmb	taɨyɛ	zɨmb	zɨmb	aph	w p	prah
Foot	wgɨr	egɨr	ɛgɨr	ɨg'gɨrɛ	cɛnba	ɛgir	rɛghel	p' n	rɛghla
Give	aβɛ-	abɛ	amɛ	haˈmɔ-	aβɛ-	aβɛ-	nathan	y t n	yäÑbh
Go	aɨllɛfɛ-	--	wɛɨrɛ	hɔˈrɔ'-	--	wɛyɛ-	---	---	---
Good	feyyan	feyya	wɛxɛ	mɔˈ7ɔ-	wɛxɛ	wɛxɛ	t'obh	t b/d m g	t'abha
Grass	ɛar	ɛer	ɛa7ar	ɛa'7ari	ɛɛr	ɛer	---	---	---
Hair (head)	gun7ɛr	-digɛr	gun7ɛr	duˈguɨra	gunɛr	gunɛr	s'ear	s' r	s'artha
Hand	wj	ɨjɛr	ɛj	ɨjja	ɛj	ɛj	yadh	y d	idha
Head	gunnɛn	g(ʷ)unɛn	gun7ɛr	gunnurɛ	gunɛr	gunɛr	rɔ'ʃ	riʃ	reʃa
Hear	ɛɛmma-	ɛɛma	ɛɛm7a	ɛɔˈ7ma-	ɛɛma-	ɛɛma-	ʃama	ʃ m '	ʃma
Heart	lɨb	lɨb	xʸɨn	nuɨba	xʸɨn	xʸɨn	lɛbˀ	l b	lebˀa
Horn	k'ɛrr	--	k'ɛn	kɔnna	--	k'ɛn	k'eren	k' r n	k'arna
I	ɨdi	ɨya	iya	hiy'ya	ɨya	ɨya	an/anoki	a n /a n k	ena
Kill	gɛddɛlɛ	k'ɛt'ɛrɛ	7ɛtɛrɛ	ɨtɔrɔ-	k'ɛt'ɛrɛ-	k'ɛt'ɛrɛ-	haragh	h r g	k''t'al

Figure 3-S1 continued.

Gloss	Soddo	Mesqan	Innemor	Mesmes	Geto	Chaha	Hebrew	Ugaritic	Aramaic
Know	ɛalɛ-	harɛ	xarɛ	haˈrɔ-	xarɛ-	xarɛ-	yadha	y d ˈ	idha
Knee	gulbɛt	--	gurmɛnd	gunnɔlda	--	--	berekh	b r k	burka
Leaf	kˈɛtˈɛl	kˈɛtˈɛl	kɛ7ɛr	koˈ7ora	nɛzɛr	kˈɛtˈɛr	aleh	---	tˈarpa
Liver	gɛbbɔt	habɛd	--	fore	xɛbt	zxɛbt	kabhedh	k b d	kabhda
Long	gɛllɛf	gelˈɛf	delka	gudˈdɔr-	geef	geef	arokh	a r k	arˁikha
Louse	kˈɛmal	kɛmal	ɛmaˈr	kˈuˈwaina	kˈɛmar	kˈɛmar	kinˁam	---	kˈaltha
Man	sɛˈb	sɛb	sɛβ	sewˈe	sɛb	sɛb	iš	m t	gabhra
Many	malɛkˈ	bɛze	noˈra	kˈokˈo	nɛ-kˈar	bɛze	rabˁim	m z d	rabˁa
Meat	bɛsɛr	bɛsɛr	bɛsɛr	boˈsora	bɛsɛr	bɛsɛr	baʃar	b ʃ r	besra
Moon	derraʔkˈa	--	dana7a	denˈna7a	--	tˈanakˈa	yareaħ	y r ħ	sahra
Mountain	gɛggɛra	anane	kˈoto	aˈnya	kˈoto	kˈoto	har	g r	tˈura
Mouth	af	af	amfw	anfe	anfw	anfw	peh	p	puma
Name	sɛm	sɛm	su	sum	sɛm	sɛm	ʃem	ʃm	ʃama
Neck	angɛt	angɛt	angɛd	angɔda	angɛt	angɛt	sˈawãˈr	ħ l kˈ ʔ	sˈura
New	addiɛ	gɛdɛr	gɛdɛr	woyˈyamo	gɛdɛr	gɛdɛr	---	---	---
Night	nɛɛɛrman	mɛɛarɛ	mɛɛɛdɛ	haˈˈonɛɛdɛ	mɛɛarɛ	mɛɛarɛ	latlah	l l ʔ	lelya
Nose	anfunna	nanfuna	amfuna	anˈfunna	anfuna	anfuna	aph	ap	nhira
One	atˈɛm	ath	aˈt	haˈti	aˈt	aˈt	ehadh	aħd	ħadh
Rain (n.)	zɛnab	zɛnab	diyɛ	diˈye	zɛraβ	zɛraβ	geʃen/matˈar	g ʃ m/m tˈr	ʃeghme
Red	bɛɛɛa	bɛɛa	bɛɛa	biˈɛa	bɛɛa	bɛɛa	adhon	a d m	sunkˈa
Road	nɔˈcˈɛ	ena(ˈ)	mea	moˈya	mɛya	ema	derekh	n t b	urħa
Root	sɛr	sɛɛr	sɛɛr	kˈɛˈna	sɛɛr	sɛɛr	ʃoreʃ	ʃ r ʃ	ʃerʃa
Sand	aɛɛwa	aɛawa	aɛawa	tˈoˈna	aɛɛwa	aɛɛwa	ħol	---	ħala
Say	balɛ-	nene	barɛ	-benɔ-	barɛ-	barɛ-	amar	r g m	emar
See	azzɛ	aze	aɛɛ	-haˈˈyɔ-	aɛɛ-	aɛɛ-	raˈah	m h/ħ d y/p h	ħza
Seed	zɛr	zena	zɛr	zur(z)ye	zɛr	zɛr	zera	d rˈ	zeru
Sit	tonna-	cˈona	cena	-cˈona-	cona-	cona-	yaʃabh	y t ˈb	ithebh
Skin (human)	goga	goga	golga	golga	goga	goga	or	---	meʃka
Sleep (v.)	ɛnna-	tegedɛre	nɛ7ɛ	-wɔdɔ(7ɔ)-	nɛyɛ-	nɛyɛ-	yaʃen	y ʃ n	dmekh
Small	malɛɛ	kˈel	kˈɛri	uˈnɛɛ	ɛrɛ	ɛrɛ	k/atˈan	ɛˈ g r/d q	zˈora
Smoke (n.)	ten	ten	tɛn	tonna	tɛn	tɛn	aʃan	q tˈ r	tenˁana
Snake	wbab	yɛdɛm wɛdere	cˈerɛ	hawaˈy	cɛrɛ	cɛrɛ	---	---	---
Stand	kˈɔmɛ-	kˈˈomɛ	tɛɛɛkɛβɛ	-tɛɛekkɔ-	kˈwɛmɛ-	kˈwɛmɛ-	amadh	---	akˈim
Star	kɔkɛb	kɛlɛbɛze	xˈɛxɛβ	hohoye	xoxɛβ	xoxɛβ	kokhabh	k b k b	kaukbha
Stone	ɛmmayɛ	ɛmɛn	ɛmɛr	oᵇna	ɛmɛr	ɛmɛr	ebhen	a b n	kepha
Sun	yɛmɛr	cˈeˈˈt	ayɛd	imɛˈ	cˈeyɛt	eyat	ʃ eme]	ʃ p ʃ	ʃ emʃˁ
Swim	wannɛ-	wacɛ	dɛnɛgɛ	-wa7aˈ	dɛnɛgɛ-	dɛnɛgɛ-	ʃaħgh	---	ɛħa
Tail	cˈɛra	faco	juwɛ	ju.7e	jɛwɛ	jɛwɛ	zanabh	eˈn b	dunba
Thin	kˈɛˈcˈcˈɛn	kˈɛcˈɛn	kˈɛcˈɛn	kˈɔcˈcˈɛna	kˈɛcˈɛr	ɛɛɛɛ	---	---	---
This	zɛ-	(ɛ)zɛˈ(h)	walda	wuˈ-	zata	zɛx	hazˁeh	h n d	dna

Figure 3-S1 continued.

Gloss	Soddo	Mesqan	Innemor	Mesmes	Gero	Chaha	Hebrew	Ugaritic	Aramaic
Three	ɕɔɛt	ɕɔˀɛt	ɕɔ7ɔɛt	ɕɔːɛtɨ	ɕɔɛt	ɕɔɛt	---	---	---
Thou	ɗɛhe	ahe	axɛ	ahe	axɛ	axɛ	atˤah	a t	ā ṅ t
Tongue	ɲɛlaɛ	alement	anɛβɛd	anˈnɔːda	anɛβɛt	anɛβɛt	laʃon	l ʃ n	leʃana
Tooth	ɛɨnn	ɕɔˀɛt	ɛɨn	ɛɨːne	ɛɨn	ɛɨn	ʃen	ʃ ñˇplˇ	ʃenˤa
Tree	zaf	ɛɨn	erɛ	yɛ7e	atankɨrt	ɛcˈɛ	eʃ	ɛˈ	ilana
Two	kit	zaf	wɨɾɛlt	wuˈ7ɛːtɨ	xˈˈɛlt	xˈˈɛlt	ʃnayim	tˈ n	trḗɣn
Warm	yɛmɔlkˈe	hˈɛy	mɔˈ	mɛ7ɔy-	mɔːkˈ	mɔːkˈ	ħam	ħ m	ħamˤima
Water	iga	mˈˈakˈ	ɨxa	7uˈha	ixa	ixa	mayim	m y m	mayˤa
We	ɨnna	ega	ina	ɨnna	yɨna	yɨna	anaħanu	---	ħnan
Wet	ychɛɛɛɛbbɛɛɛ	iːna	xːya	irˈramo	ira	ziːza	---	---	---
What?	mɨkˈay-	iːra	mɨr	mun(n)-e	mɨr	mɨr	mah	m h	ma
White	nɛcˈcˈɛ	mɨ(n)	gusaːd	gɛd-e	nɛcˈɛ	nɛcˈɛ	labhan	l b n	ħewar
Who?	maːlɛkˈ	nɛcˈe	maːn	homun-e	mɛwaːn	mwaːn	mɪ	m y	man
Woman	mɨɛt	mɨɛi	mɨɛt	ɛː(n)ɛta	mɨɛt	mɨɛt	ʃah	a t ˈ t	ĩfitˤa
Ye	ɗɛhɨm	ɛɛɨnta	axua	aˈhulwe	axu	axu	---	---	---

Figure 3-S1 continued.

Gloss	Akkadian	Mehri	Gibbali	Soqotri	Harsusi	Moroccan Arabic	Ogaden Arabic
All	kalu	kal	kal	fahre	faxreh/kal	koll	kull
Ashes	turmru/dikmenu	rmi|d	rid	riimid	remed	rmad	ramaɗ
Bark (tree)	kʾulpu	kʾali|r	kʾalor	---	kextr	kʾeʃra	jaza9
Belly	karʃa	hoofel	ɕcfci	me|r	hofel	kerʃ	batin
Big	rabu/maʾdu	loox	ʔib	ʔiikʾer	nob	kbir	kabiːr
Bird	iṣʾṣʾuru	tʾeyriit	ʔeɛfeɛrɛt	ʔeɛfeeroh	ɛʾehher	tʾir	tayr
Bite (v.)	naʃaku	nθtoowek	laʔar	θʾaʔab	kʾebbox	ɛed²	9aɟɟ
Black	ṣʾalmu	ħoor	ħôrm	ħaaher	cʾamel	kḥel	7aswad
Blood	damu	ðoora	ðor	door	eʾore	dem²	dam
Bone	eṣʾemtu	ʔazayr	ʔaylel	laħloh	alayi	ɛden	9aɟim
Breast	irtu	eodi	eado	todi	eodi	bezzula	eari
Burn (tr. v.)	ʃarapu	latakʾ	nhaat	neeħe	ħrokʾ	hṛeq	ħarraq
Cloud	erpetu	ʔafoor	ʔafor	ħihor	afor	shaba	saħab
Cold (air)	kaɛʾu	beruud	ħor	ħebhor	kʾaɛʾm	berd	bard
Come	alaku	nooka	zaha	geedaħ	noka	gay	7aja
Die	muta	moot	xerɔg	saame	mot	mot	maːt
Dog	kalbu	---	kolɔb	kalb	kob	klab	kalb
Drink (v.)	ʃatu	ʔetekʾ	ʃtikʾi//nilil	re/teere	tekʾ	ʃerb	sarab
Dry (adj.)	ablu	---	---	---	kola	yabeɛ/naʃef	yaɓiɛ
Ear	uznu	ħẹhyden	ʔiðe	ʔiðhen	ħeyeen	wednin	yuɟun
Eat	akalu	tuwo	te/the	tee	tewo	kla yakol	7akal
Egg	pelu	biizayt	kʾahizi	kaħihin	beeeh	bidaʔ	beːɟ
Eye	inu	ʔayn	ʔi	ʔayn	ayn	ɛeynin	9eːn
Fat (n.)	lipu/naħu	labh	labeh	labaħ	labħ	ʃehen	saħna
Fire	iʃatu	liwoot	loot	liyaat	lewetʾ	safya	naːr
Fish (n.)	nunu	ɛayd	ɛood/ɛeed	ɛoodeh	ħet	ħuta	samak
Fly (n.)	napruɛu	ħạðbeet	nibboot	ʔedbiiboh	eebbet	dabbana	dubaɓ
Foot	ʃepu	iryan	faʔe	laaf	fam/xef/gedel	regel	xuff
Give	nadanu	wuzun	ezu	tʾeef	adwor/wezom	stʾa	haba
Go	---	ɛeyoor	avad	tʾaaher	ɛeyor	mʃa	raħ
Good	dukʾu	meʃreh	rħi	ʃkaar	ged/taw	nezyan	tayyib
Grass	---	heriim	rvɔ	kʾazha	zera	ribɛ	ħaziːɛ
Hair (head)	ʃaʾr	leffiit	lof	lef	lor	ɛeera	sa9ir
Hand	kʾatu	ħ ayd	ʔed	ʔed	kef	idd	yad
Head	reʃu	ħẹroh	reʃ	rey	ħerih	raɛʾ	raːs
Hear	ʃemu	hiima	ʃiiʔ	ʔiimaʔ	homa	ɛmeɛ	sama9
Heart	libbu	kʾalb	ʔuub	ʔelbeb	kʾalb/helbab	kʾelb	qalb
Horn	kʾarnu	---	---	---	k/on	kʾern	qarn
I	anaku	hoh	he	hoh	hoh	ana	7ana
Kill	daku	letoov	letev	letеʔ	gayyeb	kʾtel	qatal

Figure 3-S1 continued.

Gloss	Akkadian	Mehri	Gibbali	Soqotri	Harsusi	Moroccan Arabic	Ogaden Arabic
Know	idu	varuub	yuvɔreb	ʔeerob	del	sref	ʕaraf
Knee	birku	bark	eerk	erɔ	bark	rokba	rukba
Leaf	aru	wark'a	eevizfoot	eaɔliifoh		werqa	waraq
Liver	kabit²u	ʃubdéét	eebdéét	ʃébdeh	ʃebdet	kebda	kabd
Long	arku	t'wiil	rii	riiyom	t'ewil	t'wil	tawiil
Louse	kalamatu	kinmtit	einiit	kaanim	kenemot	k'emal	qumal
Man	awilu	vay	vég	ʔaaʒ	gayg	rajel	rajul
Many	ma'du	méékon	make	déélak'	mekon/wiyen	bazzaf	katiir
Meat	ʃeru	tiiwi	téñ	téh	tewi	lhem	laħam
Moon	arħu	warx, ʔuurex	ʔeereet	ʔeereh	aret	gemra	qamar
Mountain	ʃadu	ʒibeel	gyel	fidhon	kermaym	gbel	jabal
Mouth	pu	xoh	xoh	ħah	xah	forman	fam
Name	ʃumu	ham	ʃu	ʃem	hme	seem	ʔisim
Neck	kiʃadu	rek'bet	voðo	fak'reeho	goei	senq	raqaba
New	---	---	---	---	---	---	jadiid
Night	muʃu	leeyet	_voʔre	ħte	as'er	lila	leil
Nose	appau	nexeer	naxrer	naħre	nxerir	nif	ʔanf
One	iʃten	t'aad	t'eet	t'ad	ehad/wahad	waħed	waħid
Rain (n.)	zunnu	rahmeet	musa, rhit	maaaa	gee'	ʃta	matar
Red	samu	ʔoofer	afor	ʔafer	afer	ħmer	ʔaħmar
Road	urħu	ħgoorem	ʔvru	ʔoorim	woren	t'irik'	tariiq
Root	ʃurʃu	ʔarok' ʔʔ	ʃirɔox	ʃeriiħoh	arks'	gder	jazar
Sand	baz'u	bat'aħ	aħʃi	ʃiimi	remlet	remla	turaab
Say	k'abu	ʔamoor	herɔg	ʔeemor	amor	k'al	qaal
See	amaru	valook'	ʃini	ʔek'dom	ħeni	ʃaf	saif
Seed	zar	---	---	---	baðr	zerriza	ħabb
Sit	aʃabu	ʃxolool	ekof	izɔem	ixewlol/kefot	gles	jalas
Skin (human)	maʃku	beʃreet	god	gaad	god/mhamis	gela	jiisim
Sleep (v.)	itulu/s'alalu	ʃuquuf	ʃef	ʃeef	ʃewkof	nersa	raqad
Small	s'ehru	k'ano	niisa	k'iinoh	k'ennon	s'gir	saxtir
Smoke (n.)	k'utru	niideex	mindoox	ʔeto	nedéx	doxxan	duxaan
Snake	---	riiʃiit	vozzet	bekeeleh	beddefen	henʃ	ʕuʔbaan
Stand	izuz²u	saar	---	---	s'or	wk'ef	qaam
Star	kak²abu	naʒm	ʃebkeb	kebkeb	kebkib	negma	kokab
Stone	abnu	ʃaykat	fodu	ʔooben	s'ewwer	ħegra	ħajar
Sun	ʃamʃu	ħeyoom	yu, ʃu	ʃaam	ħeyon	ʃems	xams
Swim	abnu	sbuuh	riih, ebrehi	taaabeh	seboħ	sewma	sabaħa
Tail	zib²atu	ðeel	ðonuub	deenub	ðeneb	ʃuwwal	danab
Thin	---	---	---	---	---	rhif	ʒaʔiif
This	annu	ðih	ðenu	ðedha	ðah	had	haʒa

Figure 3-S1 continued.

Gloss	Akkadian	Mehri	Gibbali	Soqotri	Harsusi	Moroccan Arabic	Ogaden Arabic
Three	---	---	---	---	---	---	talaːta
Thou	atˀa	hem, sen	ʃu, se	yaanhen, sen	eton	nta	ʔanta
Tongue	liʃanu	liʃiin	lʃa	ləʃʃin	lêʃen	lsan	lisaːn
Tooth	sinˀ	mazrah	ʃini	ʔaaləh	meθenêt	senna	sinn
Tree	isˁu	ʃiɟreet	heru	ʃerham	herom	ʃeɟra	saʒar
Two	ʃina	θroh	θroot	troh	θeret	guga	tineːn
Warm	emmu	---	---	---	ħarkˁ	dafi	ħaːmi
Water	mu	ha moh	ni imih	riiho	ħeɣyoh	ma	moːya
We	ninu	nħa	inħan	ħan, hen	neħa/ˁeti	ħna	niħin
Wet	---	---	---	---	---	fazeg	ratab
What?	minum	hee	ʔine	nim	haɬen	aʃ	ʔeːs
White	piɛˁu	laboom	lu	saʔbab	elbon	byed	ʔabyaɟ
Who?	manu	moon	mu	man	mon	ʃkun	man
Woman	sinˀiʃtu	ħarmeet	teeθ	ʔaaɟe	tetˁ/dim	mra	ħurma
Ye	---	---	---	---	---	---	ʔantum

Figure 3-S1 continued.

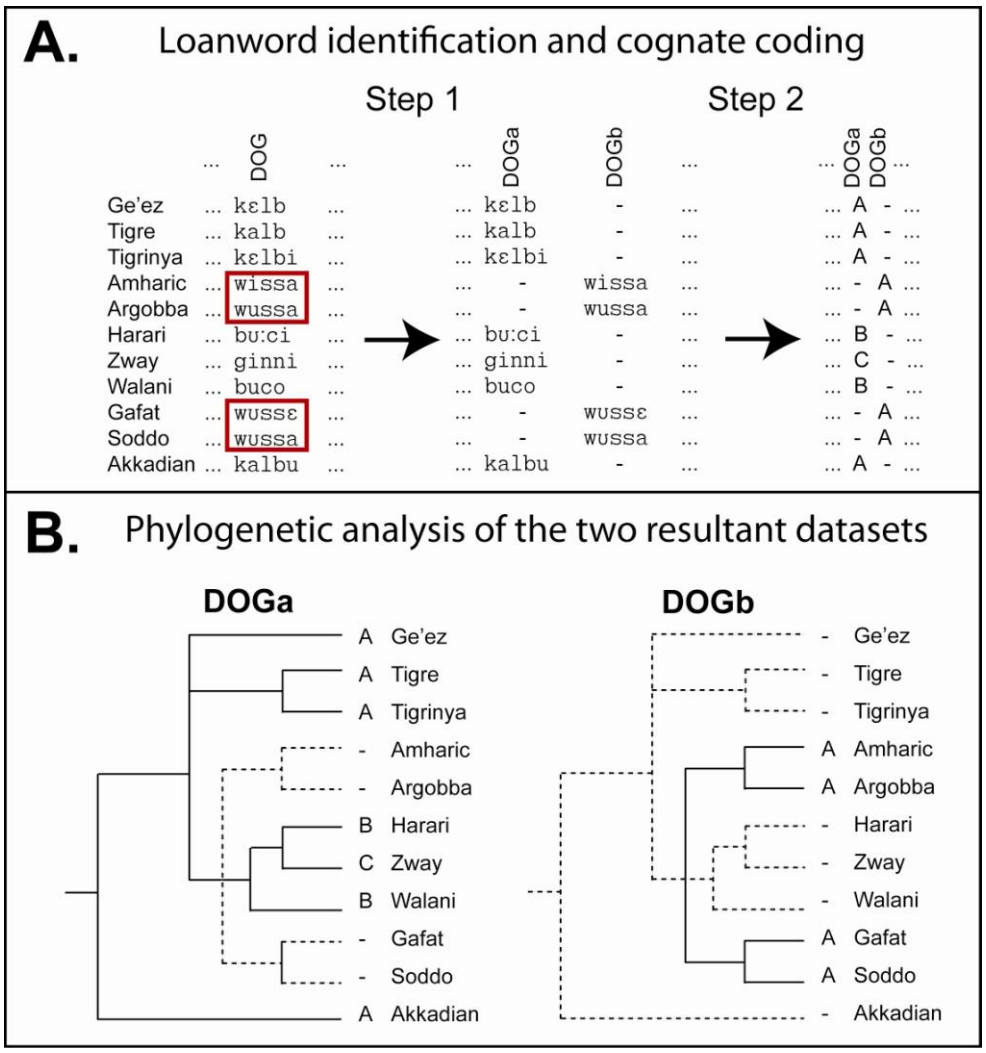

Figure 3-S2. Example of the cognate coding process as applied to the gloss of the word 'dog.' Cognate coding involved two steps: the identification of loanwords and the determination of cognate classes (depicted in panel A). Step 1: Loanwords shared by descent (identified by a red box placed around them) are placed in a new gloss category (e.g., DOGb) and coded as missing data (i.e., '-') with regard to the original gloss (e.g., DOGa). Step 2: Cognate classes are then identified from the words in each gloss category based on similar morphology (e.g., "k-lb" roots vs. "b-c" or "g-nn" roots) and given state codes (e.g., 'A,' 'B,' or 'C'). Panel B depicts the effect of this treatment on phylogenetic inference. When a language has missing data for a gloss category, the phylogenetic algorithm prunes the tree to not consider the position of that language when assessing the fit of the tree to the data for that particular gloss category. This pruning is depicted by dashed lines on the language trees of the DOGa (ancestral gloss for 'dog') and DOGb (the loanword gloss for 'dog') gloss categories. As you can see from the DOGb language tree, the data in this gloss category serves to hold together the four languages sharing the loanword by descent, but does not influence the relationships between any of the other languages.

	ALL	ASHES	BARK (TREE)	BELLY	BIG	BIRD	BITE (V.)	BLACK	BLOOD	BONE	BREAST	BURN (TR. V.)	CLOUDa	CLOUDb	COLD (OF AIR)	COME	DIE	DOGa	DOGb	DRINK (V.)	DRY (ADJ.)	EAR
Ge'ez	A	A	A	A	A	A	A	A	A	A	A	A	-	A	A	A	A	A	-	A	A	A
Tigre	A	B	B	B	A	B	A	A	A	A	A	B	-	A	A	A	A	A	-	A	A	A
Tigrinya	A	C	A	B	A	A	A	A	A	A	A	-	-	A	B	A	A	A	-	A	B	A
Amharic	A	A	A	-	B	A	A	B	A	A	B	C	-	A	A	A	A	-	B	B	C	B
Argobba	B	A	A	C	C	A	A	B	A	A	B	C	-	A	A	A	A	-	B	A	C	B
Harari	A	A	A	D	D	A	A	C	A	A	B	D	-	A	A	B	A	C	-	A	C	A
Zway	A	A	C	-	D	A	A	-	A	A	A	B	-	A	A	A	A	D	-	A	C	A
Walani	A	A	D	E	D	A	-	D	-	A	A	-	-	A	A	-	-	C	-	-	C	A
Gafat	C	-	-	F	E	A	A	B	A	A	A	C	-	A	C	C	B	-	B	B	C	A
Soddo	A	A	B	C	F	A	A	B	A	A	A	B	-	A	A	A	A	-	B	C	C	A
Mesqan	A	A	E	E	G	A	A	B	A	A	-	C	-	A	D	D	A	D	-	C	D	A
Innemor	A	A	G	C	H	A	A	E	A	A	A	D	-	A	E	A	A	D	-	C	C	A
Mesmes	D	A	F	C	I	A	A	-	A	B	A	E	-	A	E	A	A	D	-	C	C	A
Geto	A	A	G	E	J	A	A	B	A	A	A	D	-	A	D	D	A	D	-	C	D	A
Chaha	A	A	G	E	J	A	A	B	A	A	-	D	-	A	D	D	A	D	-	C	D	A
Hebrew	A	D	H	G	K	C	B	F	A	A	B	F	A	-	B	E	A	A	-	A	A	A
Ugaritic	A	E	H	B	K	D	C	-	A	A	-	F	B	-	-	E	A	A	-	A	F	A
Aramaic	A	F	B	C	L	C	A	G	A	C	B	G	A	-	B	A	A	A	-	A	A	A
Akkadian	A	F	H	C	L	D	B	A	A	A	C	F	B	-	F	F	A	A	-	A	F	A
Moroccan Arabic	A	G	I	C	M	E	D	H	A	A	D	H	C	-	A	G	A	A	-	D	A	C
Ogaden Arabic	A	G	K	G	M	E	D	L	A	A	F	H	C	-	A	G	A	A	-	D	A	C
Mehri	A	H	I	H	N	E	C	I	B	D	E	I	D	-	A	H	A	E	-	E	G	D
Jibbali	A	H	J	I	O	F	E	I	B	E	E	J	D	-	G	I	C	A	-	E	G	A
Harsusi	D	G	I	H	Q	G	G	K	C	E	E	H	D	-	F	H	A	E	-	E	G	D
Soqotri	D	G	B	J	P	F	F	J	B	F	E	J	E	-	G	K	D	A	-	F	G	A

Figure 3-S3. Cognate lists for 25 Semitic languages. This final Semitic cognate lists contained 126 words for 25 languages. The 126 words include the 97 meanings (see Figure 3-S1) and 29 loanwords (see Materials and Methods for our loanword identification and treatment methods), with gaps in the data coded by '-' characters.

	EAT	EGGa	EGGb	EGGc	EGGd	EYE	FATa	FATb	FIREa	FIREb	FISHa	FISHb	FLY (N)	FOOT	GIVE	GO	GOODa	GOODb	GRASS	HAIRa	HAIRb	HAIRc
Ge'ez	A	-	A	-	-	A	A	-	A	-	-	A	A	A	A	A	A	-	A	-	A	-
Tigre	A	-	A	-	-	A	A	-	A	-	-	A	B	A	B	B	B	-	B	-	A	-
Tigrinya	A	-	-	-	-	A	A	-	B	-	-	A	B	A	B	C	C	-	B	-	A	-
Amharic	A	-	-	B	-	A	-	A	A	-	-	A	A	A	C	C	D	-	B	-	A	-
Argobba	A	-	-	B	-	A	-	A	A	-	-	A	A	A	B	C	D	-	B	-	A	-
Harari	A	A	-	-	-	A	B	B	A	-	A	B	A	A	C	D	E	-	B	-	A	-
Zway	A	A	-	-	-	A	-	A	-	A	A	B	A	A	B	C	F	-	B	-	A	-
Walani	-	-	A	-	-	A	-	A	-	A	-	A	-	A	-	-	D	-	B	A	B	-
Gafat	A	-	A	-	-	A	C	C	A	-	-	A	A	A	B	A	-	-	B	-	A	-
Soddo	A	-	A	-	-	A	-	A	A	-	-	A	A	A	B	D	-	A	B	B	C	-
Mesqan	A	-	A	-	-	A	-	-	A	-	-	A	A	A	B	-	-	A	B	-	-	A
Innemor	A	-	-	-	B	A	D	D	A	-	-	A	A	A	B	E	G	-	B	B	C	-
Mesmes	A	-	-	-	B	A	-	A	A	-	-	-	C	A	B	A	H	-	B	-	-	A
Geto	A	-	-	-	B	A	E	E	A	-	-	A	A	B	B	-	G	-	B	B	C	-
Chaha	A	-	-	-	B	A	D	D	A	-	-	A	A	A	B	E	G	-	B	B	C	-
Hebrew	B	B	-	-	-	A	F	-	C	-	B	-	D	C	D	F	I	-	C/J	C	-	-
Ugaritic	B	-	-	-	-	A	-	-	A	-	B	-	D	D	E	-	I	-	-	C	-	-
Aramaic	B	B	-	-	-	A	G	-	D	-	C	-	E	C	F	F	I	-	C	C	-	-
Akkadian	B	C	-	-	-	A	H	-	A	-	C	-	F	E	D	G	J	-	C	C	-	-
Moroccan Arabic	B	B	-	-	-	A	I	-	E	-	D	-	G	C	G	H	K	-	D	C	-	-
Ogaden Arabic	B	B	-	-	-	A	I	-	D	-	F	-	G	J	B	I	P	-	E	C	-	-
Mehri	C	B	-	-	-	A	J	-	F	-	E	-	H	F	H	J	L	-	F	D	-	-
Jibbali	C	D	-	-	-	A	J	-	F	-	E	-	I	G	H	K	M	-	G	D	-	-
Harsusi	C	B	-	-	-	A	J	-	F	-	D	-	H	I	H	J	O	-	H	E	-	-
Soqotri	C	D	-	-	-	A	J	-	F	-	E	-	H	H	I	L	N	-	I	D	-	-

Figure 3-S3 continued.

	HAND	HEAD	HEAR	HEARTa	HEARTb	HORN	I	KILL	KNOW	KNEEa	KNEEb	LEAF	LIVER	LONG	LOUSE	MAN	MANY	MEATa	MEATb	MOON	MOUNTAIN	MOUTH
Ge'ez	A	A	A	A	-	A	A	A	A	A	-	A	A	A	A	A	A	-	A	A	A	A
Tigre	A	A	A	A	-	A	A	A	B	-	-	B	A	B	A	-	A	-	A	A	A	A
Tigrinya	A	A	A	A	-	A	A	A	C	A	-	A	B	A	A	A	A	-	A	A	B	A
Amharic	A	A	A	A	-	A	A	B	D	-	A	A	A	B	A	A	A	-	-	B	C	A
Argobba	A	B	A	A	-	A	B	B	D	-	A	A	C	B	A	A	A	A	-	B	-	A
Hararl	A	A	A	A	-	A	A	B	-	-	A	A	A	C	A	A	A	B	-	B	D	A
Zway	A	C	A	-	A	A	B	A	E	-	A	A	A	C	A	A	A	B	-	B	D	A
Walanl	A	B	-	-	A	A	B	-	-	A	-	A	A	C	A	A	A	B	-	A	-	A
Gafat	A	B	A	-	-	A	A	-	E	-	-	A	-	C	A	A	B	C	-	B	B	-
Soddo	A	D	A	A	-	A	C	B	E	-	A	A	A	D	A	A	C	B	-	B	-	A
Mesqan	A	D	A	A	-	-	B	A	F	-	-	A	A	D	A	A	A	B	-	-	E	A
Innemor	A	D	A	B	-	A	B	A	F	-	A	A	-	E	A	A	D	B	-	B	F	A
Mesmes	A	D	A	A	-	A	B	A	F	-	A	A	D	C	A	A	E	B	-	B	E	A
Geto	A	D	A	B	-	-	B	A	F	-	-	C	A	D	A	A	F	B	-	-	F	A
Chaha	A	D	A	B	-	A	B	A	F	-	-	A	A	D	A	A	A	B	-	B	F	A
Hebrew	B	A	A	A	-	A	A	C	G	A	-	D	A	F	B	B	G	B	-	C	G	B
Ugaritic	B	A	A	A	-	A	A	C	G	A	-	-	A	F	-	-	H	B	-	C	H	B
Aramaic	B	A	A	A	-	A	A	A	G	A	-	E	A	F	C	C	G	B	-	D	I	C
Akkadian	C	A	A	A	-	A	A	D	G	A	-	F	A	F	B	D	I	D	-	C	J	B
Moroccan Arabic	B	A	A	C	-	A	A	A	H	B	-	H	A	G	A	E	J	E	-	E	K	C
Ogaden Arabic	B	A	A	C	-	A	A	A	H	B	-	H	A	G	A	E	M	E	-	E	K	C
Mehri	D	E	A	-	-	A	C	E	I	A	-	-	A	-	B	F	K	F	-	A	-	D
Jibbali	E	A	A	A	-	A	C	E	I	A	-	I	A	H	B	F	K	F	-	D	K	D
Harsusi	F	E	A	-	-	A	C	F	J	A	-	-	A	-	B	F	K	F	-	D	M	D
Soqotri	E	A	B	A	-	A	C	E	I	A	-	J	A	H	B	G	L	F	-	D	L	D

Figure 3-S3 continued.

	NAIL	NAME	NECK	NEW	NIGHT	NOSE	ONE	RAIN (N)	REDa		REDb	ROADa		ROADb	ROOT	SANDa		SANDb	SAY	SEE	SEED	SIT	SKIN (N)a		SKIN (N)b		SKIN (N)c
Ge'ez	A	A	A	A	A	A	A	A	A	-	A	-		A	A	-		A	A	A	A	A	A	-		-	
Tigre	A	A	B	A	A	A	B	A	A	-	B	-		B	A	-		A	A	A	A	B	-	B		-	
Tigrinya	A	A	A	A	A	A	A	B	A	-	C	-		A	A	-		A	A	A	A	B	-	B		-	
Amharic	A	A	C	A	A	B	A	A	A	-	C	-		A	-	A		B	B	A	C	-	-		-	
Argobba	A	A	C	A	B	B	A	A	A	-	D	-		A	-	A		B	C	A	D	-	-		-	
Harari	A	A	C	A	C	B	A	A	A	-	-	A		A	B	B		A	A	B	E	-	-		A	
Zway	A	A	C	B	C	A	A	A	B	-	-	A		A	B	B		A	D	B	E	-	-		A	
Walani	A	A	C	A	C	C	A	A	-	-	-	A		A	-	A		-	-	A	-	-	-		A	
Gafat	A	A	C	A	A	A	C	A	A	-	E	-		C	-	-		A	B	A	F	-	-		A	
Soddo	A	A	C	A	D	A	A	A	-	A	F	-		A	-	A		A	C	A	F	-	-		A	
Mesqan	-	A	C	C	D	A	A	A	-	A	G	-		A	-	A		C	C	A	G	-	-		A	
Innemor	B	A	C	C	D	A	A	C	-	A	G	-		A	-	A		A	C	A	G	-	-		A	
Mesmes	B	A	C	B	-	A	A	C	-	A	G	-		D	C	C		A	D	A	G	-	-		A	
Geto	A	A	C	C	D	A	A	A	-	A	G	-		A	-	A		A	C	A	G	-	-		A	
Chaha	-	A	C	C	D	A	A	A	-	A	G	-		A	-	A		A	C	A	G	-	-		A	
Hebrew	A	A	D	A	A	A	A	D/I	C	-	H	-		E	D	-		D	A	A	H	B	-	-			
Ugaritic	-	A	E	-	A	A	A	D/I	C	-	I	-		E	-	-		E	E	A	H	-	-	-			
Aramaic	A	A	D	A	A	D	A	H	D	-	J	-		E	D	-		D	C	A	H	C	-	-			
Akkadian	A	A	A	A	D	A	D	A	E	-	J	-		E	E	-		F	F	A	H	C	-	-			
Moroccan Arabic	A	A	F	D	A	A	A	E	F	-	K	-		F	F	-		G	G	A	I	D	-	-			
Ogaden Arabic	A	A	G	D	A	A	A	I	F	-	K	-		F	J	-		G	G	E	I	G	-	-			
Mehri	A	A	-	E	A	D	E	F	G	-	L	-		G	G	-		D	H	D	J	E	-	-			
Jibbali	A	A	H	E	E	D	E	G	G	-	M	-		G	H	-		H	I	D	L	F	-	-			
Harsusi	A	A	H	E	G	D	A	D	G	-	N	-		G	-	-		D	K	D	J	F	-	-			
Soqotri	A	A	I	D	F	D	E	G	G	-	L	-		H	I	-		D	J	-	M	F	-	-			

Figure 3-S3 continued.

	SKIN (N)d	SLEEP (V.)	SMALL	SMOKE (N.)a	SMOKE (N.)b	SNAKE	STAND	STAR	STONE	SUNa	SUNb	SWIMa	SWIMb	TAILa	TAILb	TAILc	THIN	THIS	THREE	THOU	TONGUEa	TONGUEb
Ge'ez	-	A	A	-	A	A	A	A	A	A	-	A	-	A	A	-	A	A	A	A	A	-
Tigre	-	B	A	A	B	A	B	A	A	A	-	-	A	A	A	-	A	B	A	A	A	-
Tigrinya	-	C	A	B	C	B	C	A	A	A	-	-	A	-	B	-	A	A	A	B	-	-
Amharic	A	D	B	-	A	C	A	A	B	A	-	A	-	-	-	A	A	A	A	A	B	A
Argobba	A	D	C	-	A	C	A	A	B	A	-	A	-	-	-	A	A	C	B	C	B	A
Harari	-	E	D	A	B	C	D	B	A	-	A	B	-	B	C	-	A	A	B	D	-	B
Zway	-	D	D	A	B	C	D	A	A	-	A	C	-	C	D	-	A	A	A	A	-	B
Walani	-	-	E	A	B	-	-	A	A	-	A	-	-	C	D	-	-	-	B	A	-	B
Gafat	-	F	B	A	B	C	A	A	B	B	-	-	-	-	B	-	A	D	B	A	B	A
Soddo	-	E	F	A	B	C	A	A	A	C	-	A	-	-	B	-	A	A	B	E	B	A
Mesqan	-	F	G	A	B	D	A	C	A	A	-	D	-	D	E	-	A	A	B	C	-	B
Innemor	-	E	G	A	B	E	E	A	A	D	-	E	-	E	F	-	A	E	B	C	-	B
Mesmes	-	G	B	A	B	C	E	A	A	C	-	F	-	E	F	-	A	E	B	C	-	B
Geto	-	E	B	A	B	E	A	A	A	A	-	E	-	E	F	-	A	A	B	C	-	B
Chaha	-	E	B	A	B	E	A	A	A	D	-	E	-	E	F	-	B	A	B	C	-	B
Hebrew	-	H	H	C	-	F	F	A	A	E	-	F	-	A	-	-	C	E	A	A	A	-
Ugaritic	-	H	I	D	-	-	-	A	A	E	-	-	-	A	-	-	-	F	A	A	A	-
Aramaic	-	I	J	A	-	G	G	A	C	E	-	F	-	A	-	-	D	A	A	A	A	-
Akkadian	-	J	K	E	-	H	H	A	A	E	-	G	-	E	-	-	A	D	A	A	A	-
Moroccan Arabic	-	K	I	F	-	F	I	D	D	E	-	H	-	A	-	-	F	E	A	A	A	-
Ogaden Arabic	-	N	I	F	-	K	K	A	D	E	-	I	-	A	-	-	E	E	A	A	A	-
Mehri	-	L	L	G	-	G	J	D	E	F	-	I	-	A	-	-	A	A	A	F	A	-
Jibbali	-	M	M	G	-	H	J	A	F	G	-	J	-	A	-	-	A	A	A	F	A	-
Harsusi	-	L	L	G	-	J	J	A	G	F	-	I	-	A	-	-	A	A	A	A	A	-
Soqotri	-	M	L	H	-	I	L	A	A	G	-	I	-	A	-	-	A	A	A	F	A	-

Figure 3-S3 continued.

	TOOTH	TREEa	TREEb	TWO	WARM	WATERa	WATERb	WATERc	WE	WET	WHAT?	WHITEa	WHITEb	WHO?	WOMAN	YE
Ge'ez	A	A	-	A	A	A	-	-	A	A	A	A	-	A	A	A
Tigre	B	A	-	A	B	A	-	-	A	B	A	A	-	A	B	A
Tigrinya	A	B	-	A	A	A	-	-	A	C	A	A	-	A	-	B
Amharic	C	-	A	A	A	-	A	-	A	A	A	B	-	A	C	C
Argobba	A	-	A	A	A	-	A	-	A	A	A	C	-	A	D	D
Harari	A	D	-	A	C	A	-	-	A	A	A	B	-	A	E	E
Zway	A	A	-	A	A	A	-	-	A	A	A	-	A	A	C	A
Walani	A	A	-	A	A	A	-	-	A	-	A	-	A	A	F	A
Gafat	A	-	A	A	A	-	-	A	A	C	-	B	-	A	F	C
Soddo	A	C	-	A	A	-	-	A	A	D	B	B	-	B	F	-
Mesqan	D	C	-	A	D	-	-	A	A	E	A	B	-	-	F	E
Innemor	A	E	-	A	A	-	-	A	A	E	A	D	-	A	F	E
Mesmes	A	F	-	A	A	-	-	A	A	E	A	E	-	C	F	E
Geto	A	G	-	A	A	-	-	A	A	E	A	B	-	A	F	E
Chaha	A	A	-	A	A	-	-	A	A	F	A	B	-	A	F	E
Hebrew	A	A	-	B	E	A	-	-	A	A	B	F	-	D	G	A
Ugaritic	A	A	-	B	E	A	-	-	-	-	B	F	-	D	H	-
Aramaic	A	H	-	C	E	A	-	-	A	-	B	G	-	A	H	A
Akkadian	A	A	-	B	E	A	-	-	A	-	A	H	-	A	I	A
Moroccan Arabic	A	I	-	D	F	A	-	-	A	H	C	I	-	E	J	A
Ogaden Arabic	A	I	-	B	E	A	-	-	A	A	C	I	-	A	K	A
Mehri	E	-	-	C	H	B	-	-	A	G	D	F	-	A	-	A
Jibbali	A	J	-	C	-	A	-	-	A	G	E	F	-	A	L	A
Harsusi	G	J	-	C	I	B	-	-	A	G	G	F	-	A	H	A
Soqotri	F	J	-	C	-	C	-	-	A	G	F	J	-	A	M	A

Figure 3-S3 continued.

CHAPTER 4
UTILITY OF DNA VIRUSES FOR STUDYING HUMAN HOST HISTORY: CASE STUDY OF JC VIRUS[4]

Introduction

Recent research has focused on the use of microbial pathogens and commensals as complements to traditional genetic markers to investigate the population histories and demographies of their hosts (Ashford 2000; Holmes 2004). Microbial pathogens and commensals have generally faster mutation rates (μ) and shorter generation times than their hosts, which thereby often produce significant population differentiation faster than that observed in any host genetic system. These attributes also offer greater resolution for the estimation of μ and therefore other population genetic parameters such as effective population size (N_e) and coalescent time.

In turn, the utility of microbial pathogens and commensals to study host history and demography also depends on their mode of transmission [i.e.,, vertical, horizontal, or some mixture of both (Ashford 2000; Holmes 2004)]. Vertically transmitted pathogens and commensals are passed from parents to offspring within their host populations, thereby closely tying them to their host genealogies and generation times (with the latter often measured in years to decades). Thus, such microbes are expected to show older coalescent times and slower population dynamics that reflect the more ancient historical events within their hosts. In contrast, horizontally transmitted pathogens and commensals are not so constrained as they can also be passed among unrelated individuals within their host populations by direct or indirect contact with infected non-relatives. Thus, these microbes are expected to exhibit more recent coalescent

[4] Reproduced with permission from: Kitchen A, Miyamoto M, Mulligan C. 2008b. Utility of DNA viruses for studying human host history: Case study of JC virus. Mol Phylogenet Evol. 46:673-682.

times and faster population dynamics that represent changes in their infection rates due to younger historical events within their hosts.

In particular, viruses have shown great utility as markers that both corroborate and extend the population histories inferred from human DNA (Holmes 2004). Given their fast rates of μ = 0.01 to 3.4 x 10^{-3} substitutions per site per year (Jenkins et al. 2002), RNA viruses have proven their utility for studying recent human events due to societal and epidemiological changes as well as population genetic ones (e.g.,, migration). For example, Pybus et al. (2003) and Drummond et al. (2005) documented a link between virus population expansion and increased infection rates in the rapidly evolving hepatitis C virus (μ = 7.9 x 10^{-4}) due to changing public health policies during the 1920s-1950s in Egypt.

In contrast to RNA viruses, DNA viruses have presumably more variable μ, thereby making their applications to studies of human population events less clear but potentially broader. Slowly evolving DNA viruses, such as human papillomavirus (HPV-18, μ = 2 x 10^{-7}), have been used as markers of ancient human population events, including the establishment of host phylogeographic substructure (Ong et al. 1993; Bernard 1994). Conversely, hepatitis B virus (HBV) is a fast evolving virus (μ = 4.2 x 10^{-5}) whose population dynamics have been shown to reflect recent human events such as the spread of HBV infections in Japan due to postwar societal changes after World War II (Michitaka et al. 2006). Collectively, DNA viruses have the potential to track major events in human history ranging from recent societal and epidemiological changes to older phylogenetic co-divergences.

JC virus (JCV) is a double-stranded, circular, DNA virus of humans that belongs to the polyomavirus family. Although ~ 70 to 90% of all adults are seroprevalent for JCV (Padgett and Walker 1973), this virus is not normally associated with any disease, except in

immunocompromised patients (Weber and Major 1997). The normal target organ for JCV is the kidneys and the virus is thought to be passed among both relatives and non-relatives through their urine. JCV replicates in the host cell nucleus, but does not integrate into the latter's chromosomes (Khalili et al. 2007). Evolutionary studies indicate that the JCV genome has not evolved under widespread positive selection (Pavesi 2005).

JCV has been widely assumed to be a slowly evolving virus, which has co-evolved for at least 100,000 years with its human host (Sugimoto et al. 1997). For these reasons, the virus has been used to infer ancient global (Sugimoto et al. 1997, 2002; Wooding 2001) and regional (Agostini et al. 1997; Ikegaya et al. 2005) human population history. This assumption of slow evolution and ancient co-divergence is largely based on medical, epidemiological, and comparative information for individual patients, ethnic groups, JCV strains, and other polyomaviruses (Agostini et al. 1997; Sugimoto et al. 1997). These results have proven most consistent with the hypothesis of an effectively vertically transmitted and slowly changing virus. However, other such studies have instead argued that horizontal transmission from extra-familial sources occurs in > 50% of infections [Kitamura et al. 1994; Kunitake et al. 1995; see also Chen et al. (2004) for evidence of a rapid polyomavirus rate]. Significant horizontal transmission and rapid change are more characteristic of a fast evolving virus that could instead be tracking recent events in human history.

Recently, Shackelton et al. (2006) used a very different approach to estimate μ for JCV, a method that did not rely on the a priori assumption of an ancient basal co-divergence between virus and host. Instead, these authors used the different sampling dates for JCV samples to provide a new estimate of its μ, one which was independent of this assumption. Specifically, they used Bayesian, maximum likelihood (ML) and distance phylogenetic methods to show that

there was substantial geographic subdivision among global JCV populations. However, above the subtype level, they then statistically documented that the virus and human phylogenies showed no significant co-divergence, which thereby led to their rejection of the old external calibration for estimating the JCV rate. Rather, using the viral sampling dates, Shackelton et al. produced an independent estimate of μ for global JCV that was two orders of magnitude faster ($\mu = 1.7 \times 10^{-5}$) than those based on the assumption of an old basal co-divergence between virus and host ($\mu = 1$ to 4×10^{-7}; Hatwell and Sharp 2000). Correspondingly, the historical population dynamics of global JCV was now scaled in centuries rather than in tens of thousands of years.

In this study, we evaluate the utility of DNA viruses, as represented by JCV, to investigate recent versus ancient events in the history and demography of their human host. Specifically, we assess the support for a fast versus slow μ for JCV and we evaluate the ramifications for each μ when investigating human population dynamics. Our approach relies on a combination of phylogenetic and coalescent methods, which include the first use of Bayesian skyline plots to trace regional changes in N_e over time in both JCV and its human host. Coupled with different lines of evidence, comparisons of these skyline plots for JCV and humans indicate that the virus is evolving on a timescale similar to that for other fast evolving DNA viruses and RNA viruses. Thus, like these other fast evolving viruses, we conclude that JCV can be used to track recent human events, including those that are due to societal and epidemiological changes.

Materials and Methods

JC Virus and Human Mitochondrial DNA Sequences

A dataset of 407 genome sequences for JCV was collected from GenBank (Table 4-S1[5]). Two hundred and eighty contiguous nucleotides, representing the 5' hypervariable non-coding intergenic region, were removed from each genome prior to their multiple alignment with ClustalX (Thompson et al. 1994). The resultant 4,850 base pair alignment for the full coding genomes was edited by hand to minimize the number of unique gaps and to ensure the integrity of the reading frame (Figure 4-S1[6]). A subset of 92 of the 231 dated sequences from Shackelton et al. (2006) was assembled for estimation of population genetic parameters, consisting of 11 African, 21 Native American, 25 European and 35 Japanese sequences. These four regions were chosen because each was well represented by dated JCV sequences that were sampled from the major lineages of the longer-term diversity in that region (Figure 4-1). Furthermore, each region was also well represented by complete coding mitochondrial DNA (mtDNA) genomes for its human host population (see below), as well as by associated archaeological, paleoclimatological, and paleogeographic information. The sampling dates for this subset of 92 dated JCV ranged from 1970 to 2003.

A total of 196 genome sequences for human mtDNA was collected from the online Human Mitochondrial Genome Database (mtDB: www.genpat.uu.se/mtDB/; Ingman and Gyllensten

[5] Table 4-S1 (available online at *Molecular Phylogenetics and Evolution* journal website): Directory for the 407 JCV coding genomes used in this study. "Identifier" refers to the unique name adopted for each sequence (see Figure 4-S1). The fourth, fifth, and sixth columns identify the subset of 337 JCV that were considered by Shackelton et al. (2006); the subset of 231 JCV for which known sampling dates were provided by this reference; and the subset of 92 JCV (in boldface) used in our internal and external rate estimations.

[6] Figure 4-S1 (available online at *Molecular Phylogenetics and Evolution* journal website): Multiple sequence alignment for the 407 JCV coding genomes. Each sequence is represented by its unique identifier that is keyed to its GenBank accession number in Table 4-S1. Gaps are represented by "-." Position 1 of this alignment corresponds to the first site of the VP2 gene. The "v" at position 4,597 marks the location of the contiguous non-coding IR region that was excluded from this study.

2006). As reviewed in Pakendorf and Stoneking (2005), the non-coding control region sequences of each genome were then removed to limit subsequent analyses to the more conserved full coding regions. These full coding genomes were then aligned as for JCV, resulting in a multiple sequence alignment of 15,465 base pairs (Figure 4-S2[7]). In all, 56 African, 20 Native American, 60 European and 60 Japanese genomes were collected (the 60 European and 60 Japanese genomes were randomly sampled from larger datasets available in mtDB).

JC Virus Phylogenetic Analysis

A ML phylogenetic analysis was performed on the 407 JCV sequence alignment to complement the distance, ML and Bayesian analyses of 337 sequences completed by Shackelton et al. (2006). Modeltest 3.7 (Posada and Crandall 1998) was used to determine that the GTR + Γ+ I model of nucleotide substitution was most appropriate for this dataset (using the Akaike Information Criterion). The GTR + Γ+ I model allows for unequal rates among the six pairs of reciprocal substitutions (e.g.,, the probability of change from A to C equals that from C to A, but can vary from that for A to G), while accounting for rate heterogeneity among sites by a gamma distribution and proportion of invariable positions. This ML analysis was performed using PhyML (Guindon and Gascuel 2003) with branch swapping by tree bisection-and-reconnection. ML bootstrap analysis was based on 1000 pseudo-replicates to determine group support.

The available genome sequences for three other members of the polyomavirus family [simian virus 40 (SV40), BK virus (BKV), and simian agent 12 (SA12)] were initially

[7] Figure 4-S2 (available online at *Molecular Phylogenetics and Evolution* journal website): Multiple sequence alignment for the 196 human, coding, mtDNA genomes used in this study. Here, "coding" refers to both protein and structural RNA genes following Pakendorf and Stoneking (2005). Gaps are represented by "-." Position 1 of this alignment corresponds to site 577 of the Anderson Reference Sequence (ARS), which represents the first position of the tRNA-PHE gene. The final position of this alignment (15,465) corresponds to site 16,023 of the ARS, which represents the last position of tRNA-PRO. Sequences starting with "mtDNA" and "AmtDNA" follow the naming conventions of Herrnstadt et al. (2002) for the European and American sequences in mtDB (Ingman and Gyllensten 2006), respectively. All other sequences are identified by their GenBank accession numbers.

considered, but later rejected, as outgroups to root the JCV phylogenies. Each of these three other polyomavirus genomes was aligned to the set of 407 JCV sequences and then subjected to a neighbor-joining bootstrap analysis (with 1000 pseudo-replicates) using ML distances under the same GTR + Γ + I model as accepted above (Saitou and Nei 1987). However, none of these three provided a well-supported root for JCV, as judged by their bootstrap scores of <50% (results not shown). These failures can be attributed to the extensive sequence divergence between JCV and SV40, SA12 and BKV, which all have proportional sequence differences from the former of 20% to 30%. In addition to this concern with ambiguous rooting due to substitution saturation, the two related polyomaviruses, BKV and SA12, may also have undergone horizontal transfers between their human and baboon hosts, thereby questioning their use as outgroups (Cantalupo et al. 2005). For these reasons, we instead followed convention and rooted the JCV tree at its midpoint, which thereby allowed for direct comparison of this ML phylogeny to those of Shackelton et al. (2006) and others.

Estimation of μ for JC Virus

Two Bayesian skyline analyses (Drummond et al. 2005) were performed on the dataset of 92 dated JCV sequences using the program BEAST v1.3 (http://evolve.zoo.ox.ac.uk) to estimate μ. The first analysis relied on the sampling dates for each dated JCV, whereas the second assumed that these viral sequences were all sampled contemporaneously. Following the terminology introduced by Rambaut (2000) in his ML study, these two approaches for estimating μ are herein referred to as the single rate dated tips (SRDT) and single rate (SR) molecular clock models.

The SRDT analysis relied on an uninformative flat prior of μ=0 to 100, whereas the SR approach used a strong uniform prior where the total tree depth was set to 90-110 thousand years

ago (kya). The latter prior was based on the assumption of a basal co-divergence between JCV and humans calibrated to 100 kya (see Figure 4-1). These two approaches are herein referred to as "internal" and "external." In both analyses, the number of stepwise changes in N_e was fixed at 10. Thus, these two Bayesian analyses were analogous to the generalized ML skyline approach, where the number of stepwise changes in N_e can be less than their maximum (i.e.,, the number of sequences (n) minus one) (Strimmer and Pybus 2001).

All Bayesian skyline analyses were performed with the GTR + Γ+ I substitution model and a relaxed molecular clock with an uncorrelated log-normal rate distribution (Drummond et al. 2006). Markov chains were run for 40,000,000 generations and sampled every 1,000 generations with the first 4,000 samples discarded as burn-in. Unless otherwise specified (e.g.,, as above for μ), default priors were used for all parameters. The program Tracer v1.3 (http://evolve.zoo.ox.ac.uk) was used to visually inspect sampled posterior probabilities for Markov chain stationarity and to calculate summary statistics for the population genetic parameters. At least two independent runs were completed for each analysis to corroborate these final results.

JC Virus Population Dynamics

Bayesian skyline plots were generated for the four regional sets of dated JCV using the calculated mean internal and external rates to estimate historical changes in N_e. The SR rate model was used for the regional skyline plots using the external rate of 1.356×10^{-7} (Figure 4-2). The SRDT rate model was used to analyze the Native American, European and Japanese datasets with the internal rate of 3.642×10^{-5} (Figure 4-3). The original African dataset, with only 11 dated sequences, was supplemented with 55 undated African JCV. Skyline plots for Africa were then generated using the SR model with both the internal and external rates. In these various

analyses, the number of stepwise changes in N_e was set to their maximum of (n – 1) as done in the classic ML skyline approach (Pybus et al. 2000). All other conditions of these BEAST runs were as above.

Human mtDNA Population Dynamics

Bayesian skyline plots were generated for the mtDNA coding genomic sequences from the same four regions. The SR model was used on the assumption that the sampling interval of the mtDNA sequences was insignificant relative to μ. The Bayesian counterpart to the classic ML skyline model (see above) and the widely accepted μ of 1.7×10^{-8} for coding mtDNA (Ingman et al. 2000) were used while all other conditions were as above.

Bayes Factor Model Comparison

Two Bayesian skyline analyses of the 92 dated JCV dataset were performed using the SRDT and SR models with the mean rate fixed at 3.642×10^{-5} and 10 piecewise estimates of N_e. The marginal likelihood under each model was approximated by the harmonic means of the log likelihoods sampled. Both analyses were performed using identical priors and protocols for their proposals, moves and sampling to minimize known limitations in using the harmonic mean to estimate marginal likelihoods (Lartillot and Philippe 2006). Log Bayes Factor (BF) was calculated according to the equation: log BF = log [(marginal likelihood for SRDT model) / (marginal likelihood for SR)] (Raftery, 1996). In this way, we tested the significance of including the sampling dates for the JCV sequences in their SRDT analysis against the exclusion of this information in their SR treatment.

Results

ML Phylogeny for JC Virus

Our ML phylogeny of 407 JCV coding genomes (Figure 4-1) was consistent with the distance, ML and Bayesian phylogenies of Shackelton et al. (2006) for 337 JCV. All 21 previously identified JCV subtypes clustered together and ML bootstrap analysis demonstrated substantial (83% to 100%) support for 18 of these 21 groups (Sugimoto et al. 2002). The JCV sequences within each subtype cluster corresponded to host populations from distinct geographic regions and historically related ethnicities. For example, subtypes 7B1 and 7B2, which were supported by bootstrap scores of 100%, included only East Asian sequences. At the inter-subtype level, our ML phylogeny was congruent with the two of Shackelton et al. (2006) for the 13 highlighted subtypes that were used in their six jungle analyses of virus/host co-divergence (Figure 4-1). Correspondingly, our JCV phylogeny also showed no significant association above the subtype level with the three human host phylogenies as used in their jungle analyses. In short, there was substantial phylogeographic structuring of JCV diversity by subtypes in regional human populations, but no significant co-divergence at the higher levels between virus and host phylogenies.

JC Virus Rates and Skyline Plots

The mean estimate of μ for JCV, when calculated under the SR model that assumes a basal co-divergence between virus and human host of 100 kya (i.e.,, the external rate), was 1.356×10^{-7}, with a 95% credible interval of $1.089 - 1.563 \times 10^{-7}$. We used this external rate estimate to produce four regional skyline plots (Africa, Americas, Europe and Japan) that traced the changes in N_e for JCV over time on a scale of thousands to tens of thousands of years (Figure 4-2). We defined a significant change in population size as the occurrence of non-overlapping 95% credible intervals at the beginning and end of an increase or decrease. Using this criterion, we

interpreted the ~40-fold and ~25-fold increases of JCV population size beginning ~10-12 kya in Africa and Europe, respectively, to be significant, while the largest increases seen in the Americas and Japan to be suggestive but not significant.

The mean estimate of μ for JCV, when calculated using virus sampling dates under the SRDT model (i.e.,, the internal rate), was 3.642×10^{-5}, with a 95% credible interval of 1.227 to 6.149×10^{-5}. This internal rate estimate, which is close to that of Shackelton et al. (2006) using a different set of dated JCV sequences, was thus more than 100-fold faster than the external rate estimate. Skyline plots generated with this internal rate had timescales on the order of decades and centuries (Figure 4-3). As μ is a scalar of N_e and time (Pybus et al. 2000), the relative magnitudes and statistical significance of the changes in these regional plots were therefore similar to those calculated with the external rate, i.e., only Africa and Europe showed significant increases in population size.

Human mtDNA Skyline Plots

The human skyline analyses, estimated with the widely accepted μ of 1.7×10^{-8} for its coding mtDNA, produced skyline plots that were scaled in thousands to tens of thousands of years, similar to the JCV external rate plots (Figure 4-4). Using our definition of significance, we interpreted the ~30-fold population increase starting at ~14 kya in Europe and the ~65-fold increase ~55 to 25 kya in Japan to be significant, while the largest changes in Africa and the Americas and the recent increase in Japan to be suggestive of population growth.

Log Bayes Factor Model Comparison

Using the full set of dated sequences, the marginal log likelihood of the SRDT model with the internal rate was -12817.73, whereas the marginal log likelihood of the SR model with the

internal rate was -12829.22. This produced a log BF of 11.49 for the comparison of the two models.

Discussion

Bayesian Skyline Plots, Human mtDNA, and Historical Demography

Bayesian skyline plots provide a dynamic representation of changes in N_e over time. As such, they provide a moving picture of population change and capture far more information than summary statistics or simpler coalescent models that offer only snapshots of N_e. Our Bayesian skyline plots of mtDNA sequences represent the first application of this powerful method to the extensive comparative database of human mtDNA (Figure 4-4). Our plots reveal coalescences that range from 50 kya in Europe to 140 kya in Africa and, thus, capture the complete time period during which anatomically modern humans lived in these regions. Interestingly, the significant 30-fold increase noted in the European population is consistent with paleoclimatological and archaeological evidence for the retreat of the European ice sheet 13-15 kya and the subsequent spread of agriculture in Neolithic Europe (Gamble et al. 2004; Pinhasi et al. 2005). The significant 65-fold increase between 55 and 25 kya in the Japan skyline plot is consistent with a mid-Pleistocene population expansion in Asia and reflects the ancestral diversity sampled by the relatively large group of immigrants from East Asia that founded the modern Japanese population (D'Andrea et al. 1995). The other suggestive population increases are also associated with major shifts in climate, subsistence practices, or migrations, which are events thought to have the most extreme effects on regional human demography (Figure 4-4).

The mtDNA skyline plots also allow for absolute estimates of the effective number of females (N_{ef}) over time, which are of great interest as alternatives to point estimates of N_{ef}. The τ values from the skyline plots can be converted to N_{ef} estimates by dividing by the widely accepted value of 25 years-per-generation for humans (Fenner 2004). In this way, we calculate

the founder N_{ef} of the Americas to be ~2,000 females, which is consistent with much of the literature (Bonatto and Salzano 1997b) and, therefore, disagrees with the extremely low estimate of ~70 males + females by Hey (2005). Furthermore, we calculate the median estimate of African N_{ef} at the coalescent as ~1,300, which rises to ~6,000 by ~75 kya and then increases even further during the Pleistocene expansion ~50 kya. These skyline estimates of African N_{ef} add substantial detail to previously published estimates that indicate a N_{ef} of ~5,000 averaged over the ~150 kya since the origin of modern humans (Jorde et al. 1998).

JC Virus Rates and Historical Demography

The slow external JCV rate estimate produces skyline plots that are scaled in thousands to tens of thousands of years and are thus comparable to the skyline plots of its human host. We find that there are apparent associations between the JCV skyline plots using the external rate and the major human expansions detailed in the mtDNA plots (Figures 4-2 and 4-4). In particular, the external rate plots show increases in JCV population sizes for the five expansion events described in the mtDNA skyline plots (Figure 4-4).

In contrast, JCV dynamics estimated under the fast internal rate occur on the order of decades to centuries (Figure 4-3), which is a product of the internal rate being more than 100 times faster than the external rate. This extreme difference in timescale precludes a comparison of JCV population changes estimated with the fast rate to ancient population dynamics of its human host. The implication of accepting a fast internal rate is that JCV is not a marker of ancient human population dynamics. Rather, the geographic patterning of JCV diversity may reflect recent human events that are behavioral, sociological or technological in nature.

Current Support for the Fast Internal Rate of JC Virus

Which JCV evolutionary rate, fast internal or slow external, is correct? In support of the slow rate, there is the apparent similarity between the skyline plots using the external JCV rate

and those for human mtDNA (Figures 4-2 and 4-4). While this convergence supports the regional co-divergence of JCV and its human host, it is the only current result supportive of the slow external rate estimate.

In contrast, there are three lines of current evidence that favor a substantially faster rate of JCV evolution. First, our phylogenetic analysis provides no significant overall support for ancient co-divergence between JCV and humans, thereby contradicting a parallel slow rate in the virus as well as in its host. Specifically, our ML analysis of a JCV dataset that is 22% larger than that of Shackelton et al. (2006) produced the same inter-subtype topology as used in their six jungle tests (Figure 4-1), thus corroborating the lack of significant support for a long-term association between virus and host and a slow rate in the former (see these authors for details, particularly about the grouping of subtypes 2A1 and 2A2 from East Asia and the New World versus the non-basal position of African JCV). Correspondingly, this phylogeny opposes the unique hypothesis of Pavesi (2003) for an ancient African origin of JCV and thereby a similar slow rate for both virus and humans. In contrast to his study of only the slowly evolving and invariant sites for 18 JCV genomes and the highly divergent SV40 as their outgroup, the current phylogenetic conclusions are based on six different jungle analyses of 13 JCV versus human nodes; from ~200-400 viral genomes; and from all available coding positions (including the rapidly evolving ones) with their rate heterogeneity accommodated by a gamma distribution and proportion of invariable sites.

Second, our log BF test now provides a new line of evidence in support of a faster rate for JCV. In the statistics literature, log BF values greater than 5 are generally considered “decisive” in support of one model over another (Kass and Raftery 1995). In our log BF test, an indecisive result would indicate that the known sampling dates for the 92 JCV offer no more information

for their skyline analyses than does a comparable set of undated sequences. That is, if JCV is slowly evolving, the yearly differences in sampling dates would have little impact on the study of its population dynamics, which would have occurred on a timescale of thousands to tens of thousands of years. However, our log BF value of 11.49 indicates instead that these sampling dates for JCV do indeed provide critical information for its SRDT rate estimation and skyline analyses. The reason is that JCV is rapidly evolving on a timescale of decades to centuries, where yearly differences in its sampling dates can have a large effect.

Third, acceptance of the fast internal rate of $\mu = 3.642 \times 10^{-5}$ for JCV leads to a novel explanation for virus/host interactions, one that is based on recent societal and epidemiological changes in humans ~50 years ago. The fast internal rate for JCV means that its population dynamics are occurring on a timescale that is obviously too recent to track ancient human evolution. Instead, this fast internal rate is more consistent with the phylogeography of JCV reflecting recent changes in its infection rates due to societal and epidemiological shifts in human behavior or technology. In short, if the fast internal rate for JCV is accurate, we would expect to find associations between its skyline plots and known recent events in modern human history, as were initially noted for its slow external versus host mtDNA graphs (see above).

Thus, the third line of current support for the fast internal rate comes from the associations of the major recent expansions for JCV in Africa, Europe, and Japan (but not in the Americas) with known postwar societal changes at the end of World War II in 1945 (Michitaka et al. 2006). The significant 40- and 25-fold increases of the JCV populations in Africa and Europe, respectively, start ~50 years ago (Figure 4-3). The JCV skyline plot for Japan also shows a suggestive increase that begins at about the same time. In contrast, the largest increase in the Americas begins only ~15 years ago. Taken together, these results point to the fact that postwar

societal changes in the human host of JCV were much more extensive in regions near the centers of fighting during World War II than elsewhere (i.e.,, Europe and Japan versus the Americas) (Weinberg 1995). In turn, while major fighting and destruction did not occur across Africa, postwar political, economic, and technological changes, such as the replacement of colonial rule and the subsequent spread of a "green revolution," resulted in sub-Saharan Africa having the largest population growth and urbanization rates in the world over the last ~50 years (United Nations 2003, 2004). These obvious ties (or lack thereof) to known postwar changes in its human host offer a new third line of corroboration in support of the fast internal rate for JCV. In the process, they also reinforce the potential utility of JCV to address other, less, well documented events in human host history (see below).

In light of these three lines of current support, we accept the fast internal rate for JCV and its associated skyline plots as more accurate reflections of its population dynamics. Correspondingly, we conclude that the apparent similarities between the JCV external rate and human mtDNA skyline plots are coincidental.

Utility of JC Virus and Other Fast Evolving DNA Viruses for Studying the Human Host

Multiple tests are needed to document the fast versus slow evolutionary rates of DNA viruses. Based on such tests, this study shows that the evolutionary rate and population dynamics of JCV are most similar to those of other fast evolving DNA viruses (e.g., Michitaka et al. 2006). Thus, like them, the historical population dynamics of JCV are to a large extent the consequence of rapid horizontal transmissions due to recent societal and epidemiological changes in its human host. Conversely, they are not primarily the result of slow vertical transmissions across human generations, which span ~25 years. Slowly evolving DNA viruses that exhibit vertical transmission patterns (e.g.,, HPV) are more appropriate for the study of older evolutionary events within their hosts (Bernard 1994; Holmes 2004).

In light of this need for multiple testing, further follow-up studies are now encouraged to assess critically our arguments for a fast JCV rate. As used in this study, BEAST v1.3 does not account for migration. Thus, the upcoming availability of a newer version that implements a structured coalescent model is particularly welcomed, since it will allow for Bayesian rate and skyline plot estimates for subdivided JCV populations that have experienced/are experiencing migration as well as growth (http://evolve.zoo.ox.ac.uk/beast/manual.html). Along these lines, the SR and SRDT models with their relaxed, uncorrelated, log normal, molecular clocks were chosen for this study on the basis of their nested relationship that allows for straightforward comparisons and their documented successes in the treatment of both real and simulated datasets (Rambaut 2000; Drummond et al. 2005). The coefficients of variation for these two models, given their estimated mean rates of 1.356×10^{-7} and 3.642×10^{-5}, were 0.127 and 0.146, respectively. An obvious next step relative to the present study is to extend its Bayesian comparisons of these relaxed molecular clock models to include a non-clock model where rates are entirely free to vary from branch to branch (i.e.,, the unrooted model of phylogeny). Such follow-up tests are crucial to improving our understanding of the history, demography, and epidemiology of JCV both regionally and worldwide (Shackelton et al. 2006).

In conclusion, fast evolving DNA viruses like JCV can complement RNA viruses to study human events that have occurred too recently to be detected by any host genetic system currently in use (Holmes 2004). Such studies of recent human history can include the tracking of political and economic changes, new vaccination programs, and dispersal events to name just a few of the societal, epidemiological, and population genetic areas that can be addressed with fast evolving viruses such as JCV.

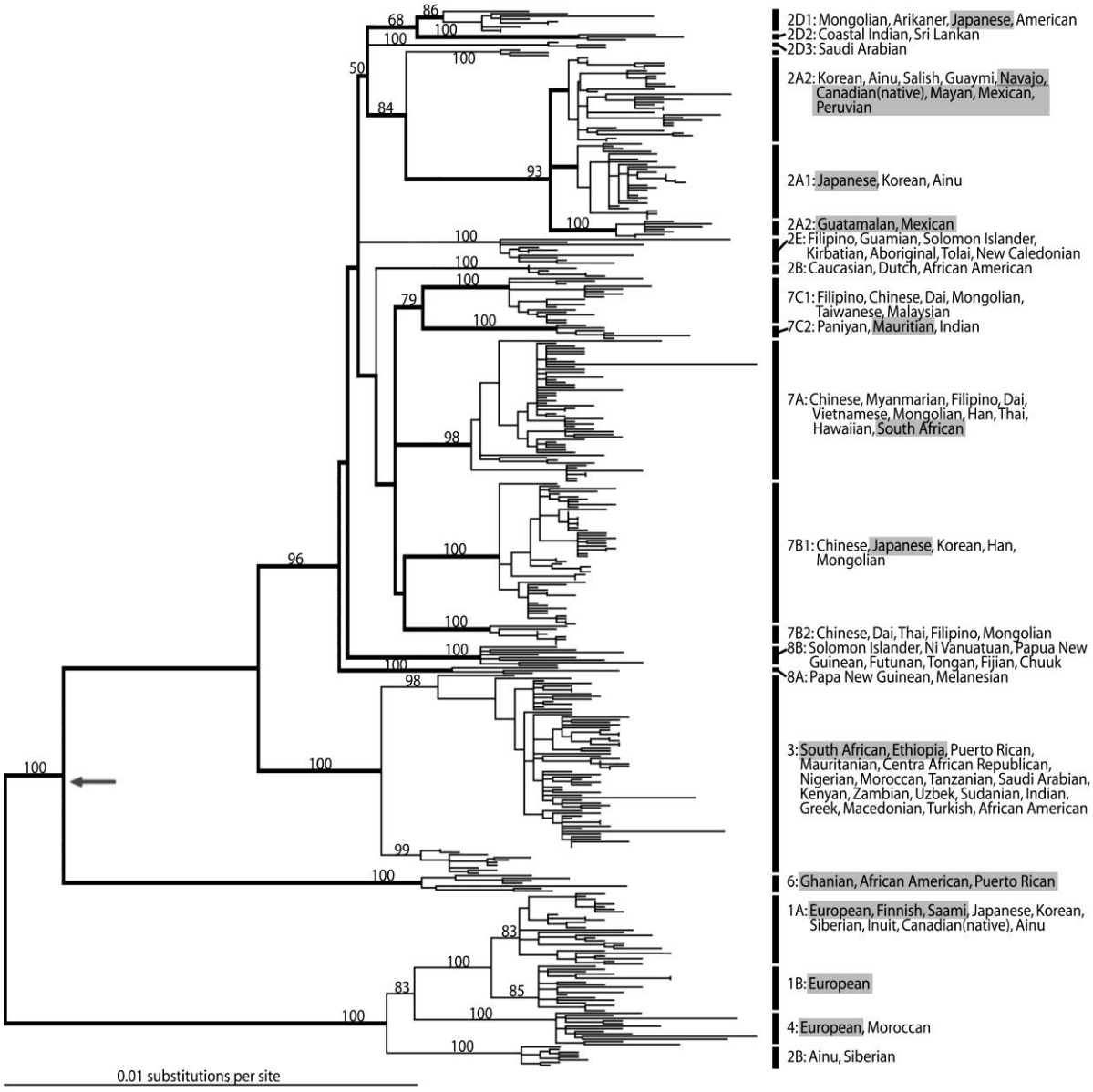

Figure 4-1. Optimal ML phylogeny for 407 JCV coding genomes. Major JCV groups are classified according to their recognized subtypes (1A to 8B; Sugimoto et al. 1997) and the ethnic origins of their human hosts are specified as well. Bootstrap proportions are given for those subtypes and other higher-order clusters with scores >50%. Boxed labels identify the 92 JCV from the four regional groups studied here (Africa, Europe, Japan, and the Americas). The bold branches and nodes trace the relationships of the 13 highlighted subtypes used in the jungle analyses by Shackelton et al. (2006). The arrow points to the node that has been widely used to reinterpret the midpoint root of other JCV phylogenies as an unresolved trichotomy of its populations from Europe, Africa, and Africa plus other regions. Correspondingly, this basal trichotomy has been widely related to the initial split within the human phylogeny and has thereby been dated at ~100 kya (Sugimoto et al. 1997; Hatwell and Sharp 2000).

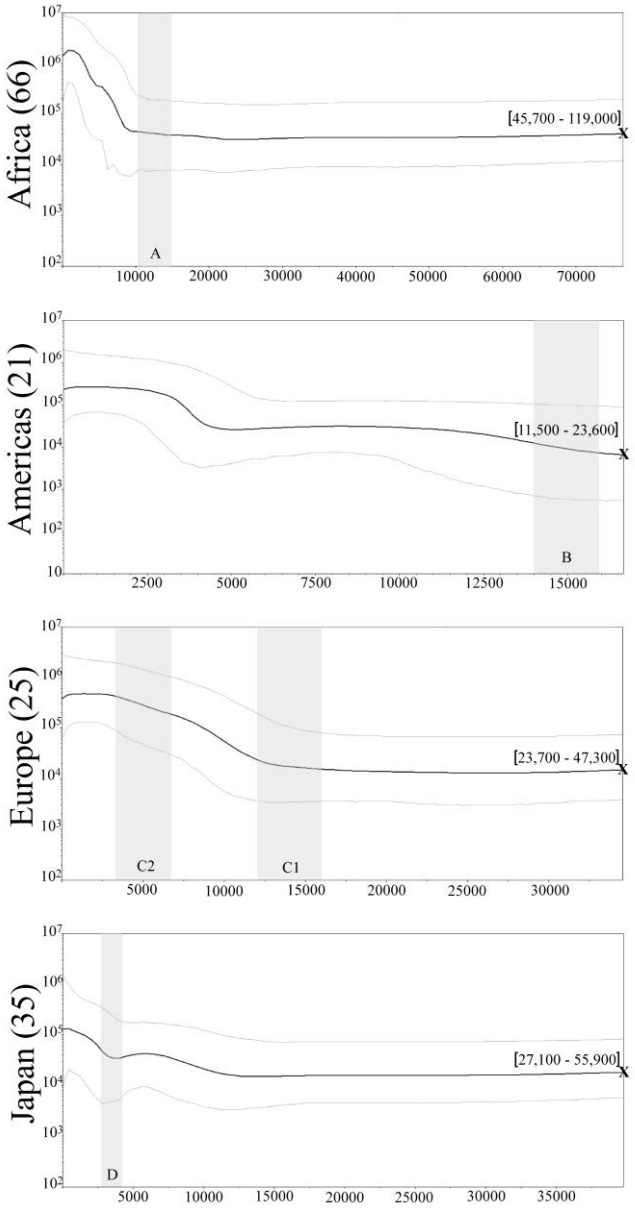

Figure 4-2. Bayesian skyline plots for the four regional groups of JCV generated with the slow external rate of 1.356×10^{-7}. The number of sequences for each region is noted in parentheses. The x-axis is time as measured in years before present and the y-axis is the scaled population size (τ, which is the product of N_e multiplied by generation time). Each curve is a plot of median τ, with its 95% credible interval indicated by the light gray lines. "**X**" marks the median coalescent time for each regional sample, with its 95% credible interval given in brackets. Note the different scales for both axes across plots. **A** to **D** highlight five increases in median τ for JCV that correspond to major events and episodes of growth in human history (see Figure 4-4).

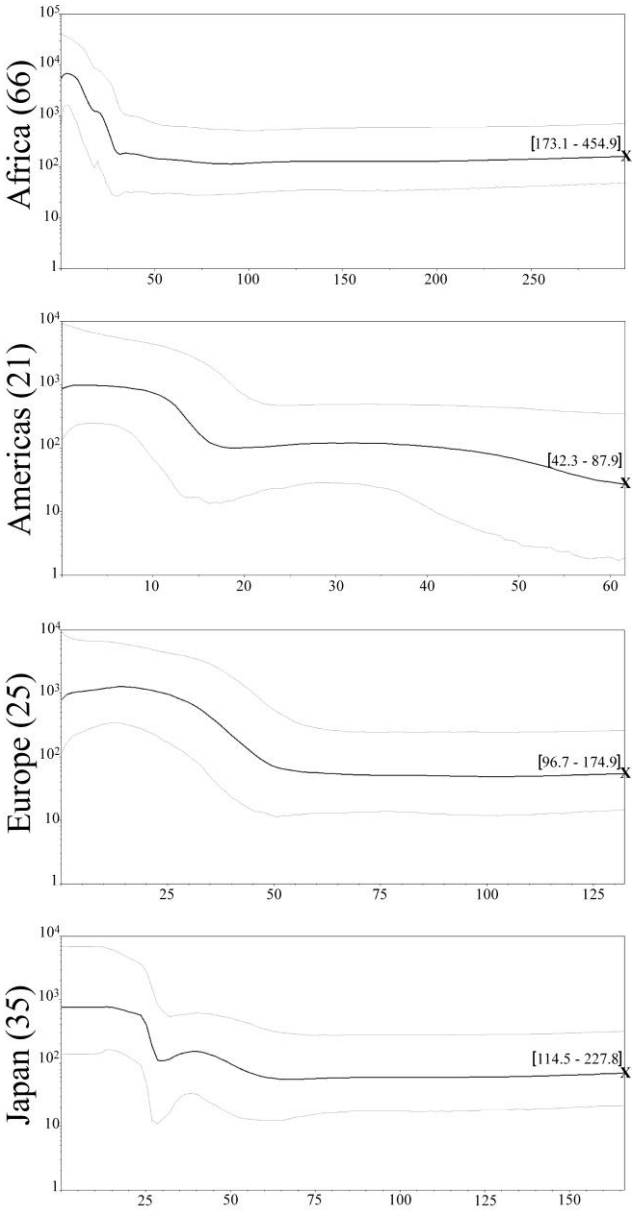

Figure 4-3. Bayesian skyline plots for the four regional groups of JCV generated with the fast internal rate of 3.642 x 10^{-5}. The x- and y-axes are in the same units as in Figure 4-2, although both are two to three orders of magnitude smaller. This difference in the dimensions of both axes reflects the fact that the genetic diversity of each population is determined by its product of N_e times μ (Tajima 1983). Correspondingly, given the recent timescales of these plots, their x-axes are best interpreted as years before the date of the most recently sampled JCV sequence(s) for that region (e.g., 2003 for Japan; see Table S1). Otherwise, these plots follow the conventions of Figure 4-2.

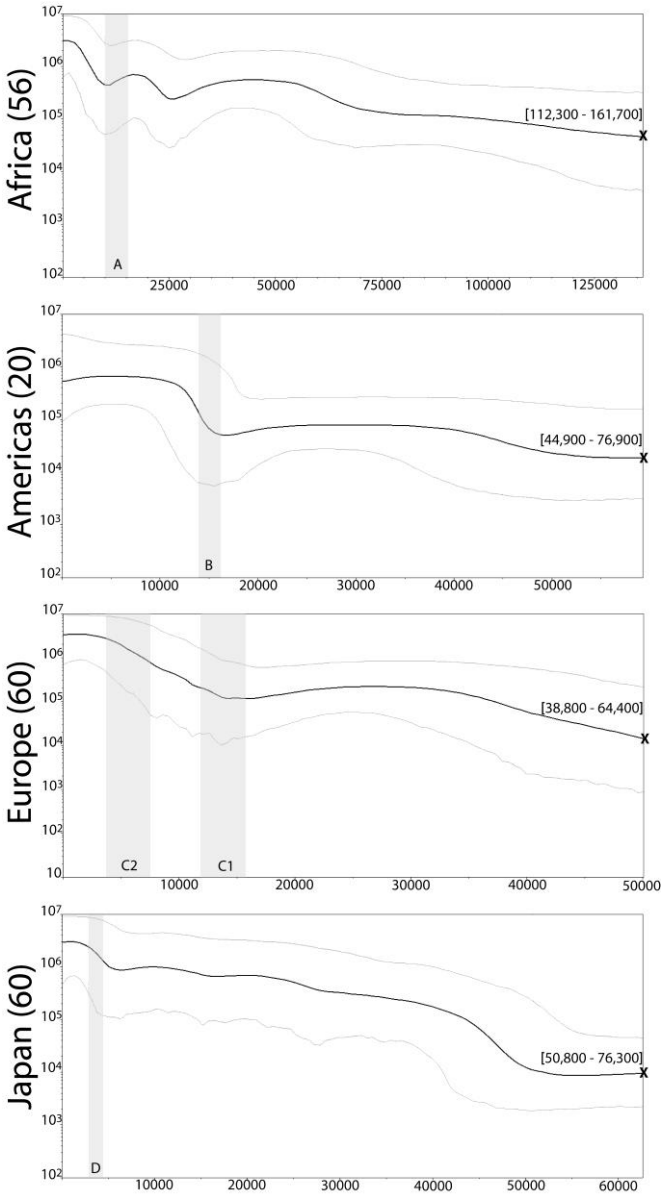

Figure 4-4. Bayesian skyline plots for the four regional groups of humans as estimated with complete mtDNA coding genomes. In these plots, τ is the product of N_{ef} multiplied by generation time. Otherwise, these plots follow the conventions of Figure 4-2. **A** to **D** highlight five episodes of population size change that correspond to known major events in human history. These events include: (**A**) increasing aridity in Africa coinciding with a reduction in observed archaeological sites (Mitchell 2002; Kuper and Kropelin 2006); (**B**) the peopling of the Americas after migration across the Bering landbridge (Greenberg et al. 1986); (**C1**) the retreat of the glaciers in Europe following the last glacial maximum (Gamble et al. 2004); (**C2**) the rise and spread of agriculture in Neolithic Europe (Pinhasi et al. 2005); and (**D**) the introduction of rice agriculture and subsequent migration of people from the Korean peninsula to Japan (D'Andrea et al. 1995).

CHAPTER 5
CONCLUSION

The demographic history of human populations has been one of shifting population sizes and complex migrations since the expansion of modern humans out of Africa at least ~50,000 years ago (e.g., Tamm et al. 2007; Atkinson, Gray, and Drummond 2008; Kayser et al. 2008). Changes in human demography have left patterns in human genetic diversity, via the evolutionary processes of genetic drift and gene flow, which can be used to infer the timing and magnitude of events in the population history of humans. However, natural selection and mutation have also shaped variation in the human genome (e.g., Reich et al. 2002; Smith, Webster, and Ellegren 2002; Webster, Smith, and Ellegren 2002; Voight et al. 2006; Nielsen et al. 2007; Sabeti et al. 2007) and limit our ability to infer past demographic events from the genetic diversity of modern human populations. While the effects of natural selection can be minimized by analyzing neutrally evolving DNA sequences, the relatively slow mutation rate of human DNA [from ~10^{-7} substitutions per site for mitochondrial DNA (Hasegawa et al. 1993; Ingman et al. 2000; Howell et al. 2003) to ~10^{-9} for nuclear DNA (Kaessmann et al. 1999)] makes it extremely difficult to infer events in human demographic history that occurred more recently than thousands of years ago from human DNA alone and also increases the variance of estimates of demographic parameters.

The limited utility of human DNA alone for investigating only ancient population events necessitates the use of complementary approaches that utilize alternative sources of data to investigate more recent events or to increase the precision of demographic parameter estimates. My dissertation demonstrates how genetic anthropology may combine complementary forms of data evolving at different rates with existing lines of evidence to investigate human demographic events from decades and centuries to tens of thousands of years in the past. I believe that genetic

anthropologists, who must consider cultural and historical as well as biological processes in human evolution, are uniquely positioned to use complementary forms of data and multiple lines of evidence from diverse disciplines. This inter-disciplinary perspective can, at best, instill in genetic anthropologists an appreciation of alternative forms of evidence and an inherent disregard for disciplinary boundaries that aid in the pursuit of understanding human history. My study demonstrates that such an approach can provide significant new insights, such as a surprisingly fast evolutionary rate for a double-stranded DNA virus or a long occupation of Beringia by the Amerind, and new strategies for addressing human demographic history.

Each of the three projects included in my dissertation presented approaches to investigating human demography at different timeframes. First, with regard to the colonization of the New World, I was able to use the mitochondrial genetic diversity of Amerind populations, along with archaeological, climatological, and paleoecological data, to propose a three-stage model for the peopling of the Americas. In this model, the proto-Amerind population first diverged from the Asian gene pool ~43,000 years ago, then underwent a long period (~20,000 years) of stable population size, and ultimately expanded into the Americas ~16,000 years ago. Second, I showed that including independent evidence into an analysis of molecular data dramatically improves inferences about the Amerind expansion into the Americas. By including archaeological and historical evidence in a reanalysis of a dataset previously used to produce an unrealistic model for the migration, I produced estimates of demographic parameters similar to those made from mitochondrial data and consistent with the archaeological record. Specifically, Hey (2005) estimated that ~70 people colonized the entire New World ~7,000 years ago, whereas I used archaeological evidence and contemporary migration rates to estimate a more plausible migration of ~1,000 Amerind ~15,000 years ago from the same dataset as Hey. Third, I

was able to estimate substantially narrower ranges for the time of the Amerind expansion (~14,000 to 16,000 years ago) into the Americas as well as the effective population size of the Amerind founder population (~1,000 to ~5000 individuals). These dates and population sizes are consistent with the physical (i.e., archaeological, climatological, and paleoecological) and molecular (maternal, paternal, and autosomal markers) evidence. This study was able to produce a model for the peopling of the Americas that reconciles the existing molecular, archaeological, and climatological evidence, dramatically narrow estimates of important demographic parameters, and is amenable to the inclusion of new data and future hypothesis testing.

In the second project, I investigated the evolution of the Semitic language family to make inferences about the history of human populations in the Middle East and Horn of Africa over the most recent centuries and millennia. I used phylogenetic techniques borrowed from evolutionary biology to estimate a language tree with divergence dates for the Semitic languages, which I then correlated with the archaeological evidence of Semitic populations. This analysis indicated an Early Bronze Age origin of Semitic ~5900 years ago between Mesopotamia and the Levant, while rejecting an African origin for the Semitic family. It also supported a single Iron Age migration of Semitic to Ethiopia ~3000 years ago from across the Red Sea, which is consistent with cultural myths linking Ethiopia to Semitic populations in the Levant. These results provide a statistically robust model of Semitic population history that is consistent with linguistic and archaeological data and will provide a hypothesis that will be tested in the future by the inclusion of genetic data.

In the final project, I studied the recent demographic history of JC virus and correlated it to events in human history within the very recent past. First, I estimated a mutation rate for JC virus that was much faster than previously assumed. This rate was ~300 times faster than

previous estimates, which indicates that JC virus is evolving on the order of decades and centuries and is a poor marker of ancient human history. Second, I used the demographic history of JC virus to infer events in recent human history. By applying the newly estimated fast rate to study JC virus's demography, I saw evidence for rapid population growth that correlates with significant events in human history, such as increased urbanization rates and population growth following World War II. Finally, I demonstrated that DNA viruses, even double-stranded DNA viruses, can be used as fast evolving markers of recent human history. The fast rate estimated for JC virus was robustly supported by multiple tests, and the inferences about human history made from JC virus demography are consistent with the known historical record. This invites the possibility that fast evolving DNA viruses can complement slow-evolving human DNA to investigate recent human history.

This study has broader implications for genetic anthropology as a whole. First, I implement a strategy for investigating human history across multiple timeframes using data evolving at vastly different rates. This increases the temporal reach of genetic anthropologists, who have traditionally addressed only ancient events that occurred tens to hundreds of thousands of years ago, to include events in the very recent past. This provides unique opportunities for combining historical and archaeological records with genetic or linguistic data into comprehensive analyses of recent population events. Indeed, this strategy is already generating much excitement, and at least one multi-million dollar Mellon grant has been awarded (to UCLA) to facilitate collaborations between geneticists, historians, and linguists to study historical events as recent as the Middle Ages. Second, I demonstrate the importance of including lines of evidence from other disciplines into the analysis of molecular data. Genetic analyses must account for and often benefit from the inclusion of existing non-genetic evidence;

for example, the inclusion of archaeological evidence for the occupation of North America by 14,000 years ago provided a more realistic estimate of the founder population size in my analysis of the peopling of the Americas. Incorporating prior information guides genetic analyses to consider only realistic scenarios and often increases the precision of inferences made from such analyses. Lastly, I show how data that are coevolving with human populations can be used to infer events in human history when the analysis of human DNA alone is not sufficient to do so. Languages and human viruses, though not vertically transmitted like human DNA and studied by fundamentally different disciplines, are intrinsically linked to human populations and can be exploited to investigate aspects of human demography. In sum, these insights demonstrate that genetic anthropologists who are interested in human demography and human history would benefit from an approach that includes biological, cultural, and historical perspectives.

The pursuit of a truly multi-disciplinary approach to genetic anthropology fits within a larger perspective that not only encompasses all of biological anthropology, but extends throughout anthropology as a whole. This perspective advocates not only the use of alternative forms of data, but also emphasizes the adoption of epistemological frames and analytical methods across disciplinary boundaries. I call this approach to anthropology "integrative anthropology" to highlight the value of integrating both data and methods across the sub-disciplinary lines (archaeological, biological, cultural, and linguistic) of anthropology to perform question-based research about human history.

I believe the time is ripe for research projects with this outlook to flourish. Previously, such integrative research has been accomplished in stops and starts, yielding interesting results but lacking sustained output, possibly due to historical reasons. For example, a first attempt to unify archaeological, biological, and linguistic evidence for the peopling of the New World

(Greenberg, Turner, and Zegura 1986) resulted in much further research on the subject, but did not produce a sustained inter-disciplinary approach to studying the Amerind colonization of the Americas, possibly because the attempt occurred before its time. In a sign that the outlook for such research is improving, my multi-disciplinary analysis of the peopling of the New World has been met with much interest, as evidenced by the numerous general science articles, requests to use my map, and citation in a *Science* review article (Goebel, Waters, and O'Rourke 2008). Another initiative toward an integrative approach is the analysis of individual samples from multiple perspectives, such as the generation of archaeological, stable isotope (also good for tracking human migrations), and molecular genetic data from the same set of samples. Technological change has also increased the viability of such research, especially with regard to genetics. The ever-decreasing cost of obtaining genetic data, in conjunction with the increasing amount of such data available in public databases, has begun to shift the focus of genetic anthropologists away from simple data collection and toward the development or application of innovative analytical methods. This increased emphasis on method application and development directly led to the incorporation of non-genetic data into genetic analyses as it became clear that some questions could not be answered with sufficient accuracy or precision by human genetic data alone, regardless of sample size. For example, Wang et al. (2007) analyzed a massive microsatellite dataset (>800 chromosomes typed for 678 loci) to investigate the peopling of the Americas from the genetic structure of Amerind populations. Though they found intriguing patterns of population structure, they were unable to obtain estimates for important demographic parameters (such as the time and size of the migration) that I was able to obtain from the analysis of smaller datasets in my study of the same event. A side-effect of the interest in new data has been the emergence of research using human pathogens as markers of human population history.

This research is maturing quickly as large datasets have led to novel insights about human disease dynamics and corrected false assumptions about the utility of specific pathogens as markers (e.g., my research demonstrating JC virus is a marker of recent instead of ancient human history).

The most radical direction of such research is the investigation and identification of patterns that extend from human biological evolution to changes in human culture. This endeavor has an inglorious history rooted in social Darwinism and the mis-application of natural selection to social and cultural outcomes. However, current examples of this research have been more circumspect and careful in their use of evolutionary theory to answer questions about human cultural and linguistic history. This research has had the most impact on the study of language history (e.g., my phylogenetic study of Semitic languages), but has recently spread to include the use of 'neutral' and 'adaptive' models of cultural or linguistic change, patterned off of 'neutral theory' and natural selection in evolutionary biology, to predict expected distributions of languages or cultural outcomes under different scenarios. Two examples of this are the melding of game theory and population genetics to understand the evolution of human social behavior (see Boyd 2006 for a review of recent work) and the use of selection and drift to investigate links between linguistic change and human demography (see Nettle 2007 for a discussion of new studies). While such research will certainly experience growing pains characterized by the over-interpretation of cultural change from an evolutionary perspective, it will at the very least provoke discussion about such broad approaches to studying human history.

The 'integrative anthropology' approach to studying human history takes an explicitly broad view of what constitutes appropriate data and the application of analytical techniques. The integration of multiple sub-disciplines of anthropology, as well as data from other disciplines

(e.g., virology or climatology), into a single, broad-based analysis requires a willingness to look for general patterns and focus on question-based research rather than expertise- or discipline-based research. Research along these lines holds great promise to produce novel insights about human history as well as define the extent to which data and analytical techniques can be combined into a single anthropological research program.

LIST OF REFERENCES

Agostini H, Yanagihara R, Davis V, Ryschkewitsch C, Stoner G. 1997. Asian genotypes of JC virus in Native Americans and in a Pacific island population: markers of viral evolution and human migration. Proc Natl Acad Sci USA. 94:14542-14546.

Anderson S, Bankier A, Barrell B, de Bruijn M, Coulson A, et al. (14 co-authors). 1981. Sequence and organization of the human mitochondrial genome. Nature. 290:457-465.

Ashford R. 2000. Parasites as indicators of human biology and evolution. J Med Microbiol. 49:771-772.

Athanasiadis G, Esteban E, Via M, Dugoujon J, Moschonas N, Chaabani H, Moral P. 2007. The X chromosome Alu insertions as a tool for human population genetics: data from European and African human groups. Eur J Hum Genet. 15:578-583.

Atkinson Q, Gray R. 2005. Curious parallels and curious connections--phylogenetic thinking in biology and historical linguistics. Syst Biol. 54:513-526.

Atkinson Q, Gray R, Drummond A. 2008. mtDNA variation predicts population size in humans and reveals a major Southern Asian chapter in human prehistory. Mol Biol Evol. 25:468-474.

Atkinson Q, Meade A, Venditti C, Greenhill S, Pagel M. 2008. Languages evolve in punctuational bursts. Science. 319:588.

Barbujani G, Bertorelle G, Chikhi L. 1998. Evidence for paleolithic and neolithic gene flow in Europe.Am J Hum Genet. 62:488-491.

Beerli P. 2006. Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. Bioinformatics 22:341-345.

Bender ML. 1971. Languages of Ethiopia - New Lexicostatistic Classification and Some Problems of Diffusion. Anthropol Linguist. 13:165-288.

Bernard H. 1994. Coevolution of papillomaviruses with human populations. Trends Microbiol. 2:140-143.

Blau J. 1978. Hebrew and Northwest Semitic: reflections on the classification of the Semitic languages. Hebrew Annual Review. 2:21-44.

Blust R. 2000. Why lexicostatistics doesn't work: the 'universal constant' hypothesis and the Austronesian languages. In: Renfrew C, McMahon A, Trask L, editors. Time depth in historical linguistics. Cambridge: The McDonald Institute for Archaeological Research. p. 311-331.

Bonatto S, Salzano F. 1997a. A single and early migration for the peopling of the Americas supported by mitochondrial DNA sequence data. Proc Natl Acad Sci USA. 94:1866-1871.

Bonatto S., Salzano F, 1997b. Diversity and age of four major mtDNA haplogroups, and their implications for the peopling of the New World. Am J Hum Genet. 61:1413-1423.

Boyd R. 2006. Evolution. The puzzle of human sociality. Science. 314:1555-1556.

Buccellati G. 1997. Akkadian. In: Hetzron R, editor. The Semitic Languages. London: Routledge. p. 69-99.

Campbell L. 2000. Time perspective in linguistics. In: Renfrew C, editor. Time depth in historical linguistics. Cambridge: The McDonald Institute for Archaeological Research. p. 3-31.

Cann R, Stoneking M, Wilson A. 1987. Mitochondrial-DNA and Human-Evolution. Nature. 325:31-36.

Cantalupo P, Doering A., Sullivan C, Pal A., Peden K, Lewis A, Pipas J. 2005. Complete nucleotide sequence of polyomavirus SA12. J Virol. 79:13094-13104.

Capelli C, Redhead N, Romano V, et al. (18 co-authors). 2006. Population structure in the Mediterranean basin: A Y chromosome perspective. Ann Hum Genet. 70:207-225.

Cavalli-Sforza L, Edwards A. 1964. Analysis of human evolution. Proceedings of the 11th International Congress of Genetics. 2:923-933.

Cavalli-Sforza LL, Menozzi P, Piazza A. 1994. The history and geography of human genes. Princeton, New Jersey: Princeton University Press.

Chen Y, Sharp P, Fowkes M, Kocher O, Joseph J, Koralnik I. 2004. Analysis of 15 novel full-length BK virus sequences from three individuals: evidence of a high intra-strain genetic diversity. J Gen Virol. 85:2651-2663.

Chikhi L, Destro-Bisol G, Bertorelle G, Pascali V, Barbujani G. 1998. Clines of nuclear DNA markers suggest a largely neolithic ancestry of the European gene pool. Proc Natl Acad Sci USA. 95:9053-9058.

Connah G. 2001. African Civilizations: An Archaeological Perspective. Cambridge: Cambridge University Press.

D'Andrea A, Crawford G, Yoshizaki M, Kudo T. 1995. Late Jomon cultigens in northeastern Japan. Antiquity 69:146-152.

Daniels PT. 1997. Scripts of Semitic languages. In: Hetzron R, editor. The Semitic Languages. London: Routledge. p. 16-45.

Derenko M, Malyarchuk B, Grzybowski T, et al. (12 co-authors). 2007. Phylogeographic analysis of mitochondrial DNA in northern Asian Populations. Am J Hum Genet. 81:1025-1041.

Diem W. 1980. Die genealogische Stellung des Arabischen in den semitischen Sprachen: ein eingelostes Problem der Semitistik. In: Diem W, Wild S, editors. Studien aus Arabistik und Semitistik, A. Spitaler zum 70. Wiesbaden: Harrassowitz. p. 65-85.

Dillehay TD, editor. 1997. The archaeological context and interpretation. Washington, DC: Smithsonian Institution Press.

Drummond A, Ho S, Phillips M, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biology. 4:699-710.

Drummond A, Rambaut A, Shapiro B, Pybus O. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol. 22:1185-1192.

Drummond A, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 7:214.

Duffy S, Shackelton L, Holmes E. 2008. Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet. 9:267-276.

Ehret C. 1995. Reconstructing proto-Afroasiatic (proto-Afasian): Vowels, tone, consonants, and vocabulary. Berkeley: University of California Press.

Ehret C. 1988. Social transformations in the early history of the Horn of Africa: linguistic clues to developments of the period 500 BC to AD 500. In: Bayene T, editor. Proceedings of the Eigth International Conference of Ethiopian Studies. Addis Ababa: Institute of Ethiopian Studies. p. 639-651.

Ehret C. 2000. Testing the expectations of glottochronology against the correlations of language and archaeology in Africa. In: Renfrew C, McMahon A, Trask L, editors. Time depth in historical linguistics. Cambridge: The McDonald Institute for Archaeological Research. p. 373-399.

Ehret C, Keita S, Newman P. 2004. The origins of Afroasiatic. Science. 306:1680.

Ehrich RH, editor. 1992. Chronologies in Old World archaeology. Chicago: University of Chicago Press.

Elias S, Short S, Nelson C, Birks H. 1996 Life and times of the Bering land bridge. Nature. 382:60-63.

Faber A. 1980. Genetic subgroupings of the Semitic languages. Austin, Texas: Linguistics Department, University of Texas at Austin.

Faber A. 1997. Genetic subgrouping of the Semitic languages. In: Hetzron R, editor. The Semitic Languages. London: Routledge. p. 3-15.

Fagundes N, Ray N, Beaumont M, Neuenschwander S, Salzano F, Bonatto S, Excoffier L. 2007. Statistical evaluation of alternative models of human evolution. Proc Natl Acad Sci USA. 104:17614-17619.

Falush D, Wirth T, Linz B, et al. (18 co-authors). 2003. Traces of human migrations in Helicobacter pylori populations. Science 299:1582-1585.

Fattovich R. 1990. Remarks on the pre-Aksumite period in northern Ethiopia. Journal of Ethiopian Studies. 23:1-33.

Fenner J. 2004. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am J Phys Anthropol. 128:415-423.

Fisher RA. 1930. Genetical theory of natural selection. Oxford: Clarendon Press.

Fix A. 2005. Rapid deployment of the five founding Amerind mtDNA haplogroups via coastal and riverine colonization. Am J Phys Anthropol. 128:430-436.

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

Fu Y. 2001. Estimating mutation rate and generation time from longitudinal samples of DNA sequences. Mol Biol Evol. 18:620-626.

Gamble C, Davies W, Pettitt P, Richards M. 2004. Climate change and evolving human diversity in Europe during the last glacial. Phil Trans R Soc Lond B. 359:243-254.

Garrigan D, Kingan S, Pilkington M, et al. (12 co-authors). 2007. Inferring human population sizes, divergence times and rates of gene flow from mitochondrial, X and Y chromosome resequencing data. Genetics 177:2195-2207.

Gilbert M, Rambaut A, Wlasiuk G, Spira T, Pitchenik A, Worobey M. 2007. The emergence of HIV/AIDS in the Americas and beyond. Proc Natl Acad Sci USA. 104:18566-18570.

Goebel T. 2007. The missing years for modern humans. Science. 315:194-196.

Goebel T, Waters M, O'Rourke D. 2008. The late Pleistocene dispersal of modern humans in the Americas. Science. 319:1497-1502.

Gordon CH. 1997. Amorite and Eblaite. In: p. Hetzron R, editor. The Semitic Languages. London: Routledge. p. 100-113.

Gordon RG, editor. 2005. Ethnologue: languages of the world. Dallas: SIL International.

Gray R, Atkinson Q. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. Nature. 426:435-439.

Gray R, Jordan F. 2000. Language trees support the express-train sequence of Austronesian expansion. Nature. 405:1052-1055.

Greenberg J, Turner C, Zegura S. 1986. The settlement of the Americas: a comparison of the linguistic, dental and genetic evidence. Curr Anthropol. 77:477-497.

Guindon S, Gascuel O. 2003. Simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52:696–704.

Guthrie RD. 1990. Frozen fauna of the mammoth steppe. Chicago: University of Chicago Press.

Hamilton M, Buchanan B. 2007. Spatial gradients in Clovis-age radiocarbon dates across North America suggest rapid colonization from the north. Proc Natl Acad Sci USA. 104:15625-15630.

Hammer M. 1994. A recent insertion of an alu element on the Y chromosome is a useful marker for human population studies. Mol Biol Evol. 11:749-761.

Hammer M, Redd A, Wood E, et al. (12 co-authors). 2000. Jewish and Middle Eastern non-Jewish populations share a common pool of Y-chromosome biallelic haplotypes. Proc Natl Acad Sci USA. 97:6769-6774.

Hammer M, Spurdle A, Karafet T, et al. (12 co-authors). 1997. The geographic distribution of human Y chromosome variation. Genetics 145:787-805.

Handt O, Meyer S, von Haeseler A. 1998. Compilation of human mtDNA control region sequences. Nucleic Acids Res. 26:126-129.

Handt O, Richards M, Trommsdorff M, et al. (13 co-authors). 1994. Molecular-Genetic Analyses of the Tyrolean Ice Man. Science 264:1775-1778.

Hasegawa M, Dirienzo A, Kocher T, Wilson A. 1993. Toward a More Accurate Time-Scale for the Human Mitochondrial-DNA Tree. J Mol Evol. 37:347-354.

Hatwell J, Sharp P. 2000. Evolution of human polyomavirus JC. J Gen Virol. 81:1191-1200.

Hayward RJ. 2000. Afroasiatic. In: Heine B, Nurse D, editors. African languages. Cambridge: Cambridge University Press. p. 74-98.

Herrnstadt C, Elson J, Fahy E, et al. (11 co-authors). 2002. Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences for the major African, Asian, and European haplogroups. Am J Hum Genet. 70:1152-1171.

Hetzron R. 1976. Two principles of genetic reconstruction. Lingua. 38:89-104.

Hetzron R, editor. 1997. The Semitic languages. London: Routledge.

Hey J. 2005. On the number of New World founders: a population genetic portrait of the peopling of the Americas. PLoS Biology. 3:e193.

Hoffecker J, Elias S. 2003. Environment and archaeology in Beringia. Evol Anthropol. 12:34-49.

Hoffecker J, Powers W, Goebel T. 1993 The colonization of Beringia and the peopling of the New World. Science. 259:46-53.

Holden C. 2002. Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony analysis. P Roy Soc B-Biol Sci. 269:793-799.

Holmes E. 2004. The phylogeography of human viruses. Mol Ecol. 13:745-756.

Hopkins DM. 1982. Aspects of the paleogeography of Beringia during the Late Pleistocene. In: Hopkins DM, Matthews JV, Schweger CE, Young SB, editors. Paleoecology of Beringia. New York: Academic Press. p. 3-28.

Howell N, Smejkal C, Mackey D, Chinnery P, Turnbull D, Herrnstadt C. 2003. The pedigree rate of sequence divergence in the human mitochondrial genome: There is a difference between phylogenetic and pedigree rates. Am J Hum Genet. 72:659-670.

Huehnergard J. 1992. Languages of the ancient Near East. In: The Anchor Bible Dictionary. p. 155-170.

Huehnergard J. 1990. Remarks on the classification of the Northwest Semitic Languages. In: Hoftizjer J, editor. Deir 'Alla Symposium. Leiden: Brill. p. 282-293.

Ikegaya H, Zheng H, Saukko P, et al. (12 co-authors). 2005. Genetic diversity of JC virus in the Saami and the Finns: implications for their population history. Am J Phys Anthropol. 128:185-193.

Ilan D. 2003. The Middle Bronze Age (circa 2000 - 1500 B.C.E.). In: Richard S, editor. Near Eastern archaeology: a reader. Winona Lakes, Indiana: Eisenbrauns. p. 331-342.

Ingman M, Gyllensten U. 2006. Human Mitochondrial Genome Database, a resource for population genetics and medical sciences. Nucleic Acids Res. 34:D749-D751.

Ingman M, Kaessmann H, Paabo S, Gyllensten U. 2000. Mitochondrial genome variation and the origin of modern humans. Nature. 408:708-713.

Jenkins G, Rambaut A, Pybus O, Holmes E. 2002. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. J Mol Evol. 54:156-165.

Jobling MA, Hurles ME, Tyler-Smith C. 2004. Into new found lands. In: Human Evolutionary Genetics. New York: Garland Science. pp. 339-372.

Jorde L, Bamshad M, Rogers A. 1998. Using mitochondrial DNA and nuclear DNA markers to reconstruct human evolution. Bioessays. 20:126-136.

Joyce D. 2006. Chronology and new research on the Schaefer mammoth (*Mammuthus primigenius*) site, Kenosha County, Wisconsin, USA. Quat Intl. 142:44-57.

Kaessmann H, Heissig F, von Haeseler A, Paabo S. 1999. DNA sequence variation in a non-coding region of low recombination on the human X chromosome. Nat Genet. 22:78-81.

Kass R, Raftery A. 1995. Bayes factors. J Am Stat Assoc. 90:773-795.

Kaye AS, Rosenhouse J. 1997. Arabic dialects and Maltese. In: Hetzron R, editor. The Semitic Languages. London: Routledge. p. 263-311.

Kayser M, Choi Y, van Oven M, Mona S, Brauer S, Trent R, Suarkia D, Schiefenhovel W, Stoneking M. 2008. The impact of the Austronesian expansion: evidence from mtDNA and Y chromosome diversity in the Admiralty Islands of Melanesia. Mol Biol Evol. 25:1362-1374.

Khalili K, White M, Lublin F, Ferrante P, Berger J. 2007. Reactivation of JC virus and development of PML in patients with multiple sclerosis. Neurology 68:985-990.

Kimura M. 1968. Evolutionary rate at the molecular level. Nature. 217:614-626.

Kitamura T, Kunitake T, Guo J, Tominaga T, Kawabe K, Yogo Y. 1994. Transmission of the human polyomavirus JC virus occurs with the family and outside the family. J Clin Microbiol. 32:2359-2363.

Kitchen A, Miyamoto M, Mulligan C. 2008a. Three-stage colonization model for the peopling of the Americas. PLoS ONE. 3:e1596.

Kitchen A, Miyamoto M, Mulligan C. 2008b. Utility of DNA viruses for studying human host history: Case study of JC virus. Mol Phylogenet Evol. 46:673-682.

Kivisild T, Shen P, Wall D, et al. (17 co-authors). 2006. The role of selection in the evolution of human mitochondrial genomes. Genetics. 172:373-387.

Knight A, Batzer M, Stoneking M, Tiwari H, Scheer W, Herrera R, Deininger P. 1996. DNA sequences of Alu elements indicate a recent replacement of the human autosomal genetic complement. Proc Natl Acad Sci USA. 93:4360-4364.

Kogan LE, Korotayev AV. 1997. Sayhadic (Epigraphic South Arabian). In: Hetzron R, editor. Semitic Languages. London: Routledge. p. 220-241

Kolman C, Bermingham E, Cooke R, Ward R, Arias T, Guionneau-Sinclear F. 1995. Reduced mtDNA diversity in the Ngobe Amerinds of Panama. Genetics. 140:273-283.

Kolman C, Sambuughin N, Bermingham E. 1996. Mitochondrial DNA analysis of mongolian populations and implications for the origin of New World founders. Genetics 142:1321-1334.

Kunitake T, Kitamura T, Guo J, Taguchi F, Kawabe K, Yogo Y. 1995. Parent-to- child transmission is relatively common in the spread of human polyomavirus JC virus. J Clin Microbiol. 33:1448-1451.

Kuper R, Kropelin S. 2006. Climate-controlled Holocene occupation in the Sahara: motor of Africa's evolution. Science. 313:803-807.

Kuzmin Y, Keates S. 2005. Dates are not just data: paleolithic settlement patterns in Siberia derived from radiocarbon records. Am Antiquity. 70:773-789.

Landsteiner K. 1901. Uber Agglutinationserscheinungen normalen menschlichen. Weiner Klin. Wochenschr. 14.

Lartillot N, Philippe H. 2006. Computing Bayes factors using thermodynamic integration. Syst Biol. 55:195-207.

Lloyd S. 1984. The archaeology of Mesopotamia. New York: Thames and Hudson.

Lovell A, Moreau C, Yotova V, Xiao F, Bourgeois S, Gehl D, Bertranpetit J, Schurr E, Labuda D. 2005. Ethiopia: between Sub-Saharan Africa and Western Eurasia. Ann Hum Genet. 69:275-287.

Macaulay V, Hill C, Achilli A, et al. (21 co-authors). 2005. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science. 308:1034-1036.

Malhi R, Eshleman J, Greenberg J, Weiss D, Schultz Shook B, Kaestle F, Lorenz J, Kemp B, Johnson J, Smith D. 2002. The structure of diversity within new world mitochondrial DNA haplogroups: Implications for the prehistory of North America. Am J Hum Genet. 70:905-919.

Mandryk C, Josenhans H, Fedje D, Mathewes R. 2001. Late Quaternary paleoenvironments of Northwestern North America: implications for inland versus coastal migration routes. Quat Sci Rev. 20:301-314.

Merriwether D, Rothhamer F, Ferrell R. 1995. Distribution of the 4 founding lineage haplotypes in Native Americans suggests a single wave of migration for the New World. Am J Phys Anthropol. 98:411-430.

Michitaka K, Tanaka Y, Horiike N, Dhoung T, Chen Y, Matsuura K, Hiasa Y, Mizokami M, Onji M. 2006. Tracing the history of hepatitis B virus genotype D in western Japan. J Med Virol. 78:44-52.

Mishmar D, Ruiz-Pesini E, Golik P, et al. (13 co-authors). 2003. Natural selection shaped regional mtDNA variation in humans. Proc Natl Acad Sci USA. 100:171-176.

Mitchell P. 2002. The Archaeology of Southern Africa. Cambridge: Cambridge University Press.

Moorey PRS. 1994. Ancient Mesopotamian materials and industries: the archaeological evidence. Oxford: Clarendon Press.

Mulligan C, Hunley K, Cole S, Long J. 2004. Population genetics, history, and health patterns in native Americans. Annu Rev Genom Hum Genet. 5:295-315.

Mulligan C, Kitchen A, Miyamoto M. 2006. Comment on "Population size does not influence mitochondrial genetic diversity in animals." Science. 314:1390.

Nardo D. 2007. Ancient Mesopotamia. Detroit: Greenhaven Press.

Nebel A, Filon D, Brinkmann B, Majumder P, Faerman M, Oppenheim A. 2001. The y-chromosome pool of Jews as part of the genetic landscape of the Middle East. Am J Hum Genet. 69:1095-1112.

Nebel A, Landau-Tasseron E, Filon D, Oppenheim A, Faerman M. 2002. Genetic evidence for the expansion of Arabian tribes into the Southern Levant and North Africa. Am J Hum Genet. 70:1594-1596.

Nettle D. 1999. Linguistic diversity of the Americas can be reconciled with a recent colonization. Proc Natl Acad Sci USA. 96:3325-3329.

Nettle D. 2007. Language and genes: a new perspective on the origins of human cultural diversity. Proc Natl Acad Sci USA. 104:10755-10756.

Newton M, Raftery A, Davison A, et al. (33 co-authors). 1994. Approximate Bayesian-Inference with the Weighted Likelihood Bootstrap. J Roy Stat Soc B Met. 56:3-48.

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark A. 2007. Recent and ongoing selection in the human genome. Nat Rev Genet. 8:857-868.

Non A, Kitchen A, Mulligan C. 2007. Identification of the most informative regions of the mitochondrial genome for phylogenetic and coalescent analyses. Mol Phylogenet Evol. 44:1164-1171.

O'Rourke D, Hayes M, Carlyle S. 2000. Spatial and temporal stability of mtDNA haplogroup frequencies in native North America. Hum Biol. 72:15-34.

Ohta T. 1992. The Nearly Neutral Theory of Molecular Evolution. Annu Rev Ecol Syst. 23:263-286.

Ong C, Chan S, Campo M, et al. (11 co-authors). 1993. Evolution of human papillomavirus type 18: an ancient phylogenetic root in Africa and intratype diversity reflect coevolution with human ethnic groups. J Virol. 67:6424-6431.

Overstreet DF. 2005. Late-glacial ice-marginal adaptation in southeastern Wisconsin. In: Bonnichsen R, Lepper BT, Stanford D, Waters MR, editors. Paleoamerican origins: beyond Clovis. College Station, TX: Center for the Study of the First Americans. p. 183-195.

Paabo S. 1985. Molecular-Cloning of Ancient Egyptian Mummy DNA. Nature. 314:644-645.

Padgett B, Walker D. 1973. Prevalence of antibodies in human sera against JC virus, an isolate from a case of progressive multifocal leukoencephalopathy. J Infect Dis. 127:467-470.

Pagel M, Atkinson Q, Meade A. 2007. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. Nature. 449:717-720.

Pakendorf B, Stoneking M. 2005. Mitochondrial DNA and human evolution. Annu Rev Genomics Hum Genet. 6:165-183.

Passarino G, Semino O, Quintana-Murci L, Excoffier L, Hammer M, Santachiara-Benerecetti A. 1998. Different genetic components in the Ethiopian population, identified by mtDNA and Y-chromosome polymorphisms. Am J Hum Genet. 62:420-434.

Pavesi A. 2003. African origin of polyomavirus JC and implications for prehistoric human migrations. J Mol Evol. 56:564-572.

Pavesi A. 2005. Utility of JC polyomavirus in tracing the pattern of human migrations dating to prehistoric times. J Gen Virol. 86:1315-1326.

Pauling L, Itano H, Singer S, Wells I. 1949. Sickle-cell anemia, a molecular disease. Science. 110:543-548.

Payseur B, Cutter A, Nachman M. 2002. Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. Mol Biol Evol. 19:1143-1153.

Pinhasi R, Fort J, Ammerman A. 2005. Tracing the origin and spread of agriculture in Europe. PLoS Biology. 12:2220-2228.

Pitulko V, Nikolsky P, Girya E, Basilyan A, Tumskoy V, Koulakov S, Astakhov S, Pavlova E, Anisimov M. 2004. The Yana RHS site: humans in the Arctic before the last glacial maximum. Science. 303:52-56.

Polanski A, Kimmel M, Chakraborty R. 1998. Application of a time-dependent coalescence process for inferring the history of population size changes from DNA sequence data. Proc Natl Acad Sci USA. 95:5456-5461.

Posada D, Crandall K. 1998. Modeltest: testing the model of DNA substitution. Bioinformatics. 14:817-818.

Pybus O, Drummond A, Nakano T, Robertson B, Rambaut A. 2003. The epidemiology and iatrogenic transmission of hepatitis C virus in Egypt: a Bayesian coalescent approach. Mol Biol Evol. 20:381-387.

Pybus O, Rambaut A, Harvey P. 2000. An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics. 155:1429-1437.

Quintana-Murci L, Krausz C, Zerjal T, et al. (13 co-authors). 2001. Y-chromosome lineages trace diffusion of people and languages in southwestern Asia. Am J Hum Genet. 68:537-542.

Quintana-Murci L, Semino O, Bandelt H, Passarino G, McElreavey K, Santachiara-Benerecetti A. 1999. Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. Nat Genet. 23:437-441.

Rabin C. 1975. Lexicostatistics and the internal divisions of Semitic. In: Bynon T, Bynon J, editors. Hamito-Semitica. The Hague: Mouton. p. 85-102

Raftery AE. 1996. Hypothesis testing and model selection. In: Gilks WR, Richardson S, Spiegelhalter DJ, editors. Markov chain Monte Carlo in practice. London: Chapman & Hall. p. 163-188.

Rambaut A. 2000. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. Bioinformatics. 16:395-399.

Rambaut A, Drummond A. 2007. Tracer v1.4.

Redelings B, Suchard M. 2005. Joint Bayesian estimation of alignment and phylogeny. Syst Biol. 54:401-418.

Reich D, Schaffner S, Daly M, McVean G, Mullikin J, Higgins J, Richter D, Lander E, Altshuler D. 2002. Human genome sequence variation and the influence of gene history, mutation and recombination. Nat Genet. 32:135-142.

Rendsburg GA. 2003. Semitic languages (with special reference to the Levant). In: Richard S, editor. Near Eastern archaeology: a reader. Winona Lakes, Indiana: Eisenbrauns. p. 71-73.

Renfrew C, McMahon A, Trask L, editors. 2000. Time depth in historical linguistics. Cambridge: The McDonald Institute for Archaeological Research.

Richard S, editor. 2003a. Near Eastern Archaeology: a reader. Winona Lake, Indiana: Eisenbrauns.

Richard S. 2003b. The Early Bronze Age in the southern Levant. In: Richard S, editor. Near Eastern archaeology: a reader. Winona Lake, Indiana: Eisenbrauns. p. 286-302.

Richards M, Macaulay V, Bandelt H, Sykes B. 1998. Phylogeography of mitochondrial DNA in western Europe. Ann Hum Genet 62:241-260.

Rodgers J. 1992. The subgrouping of the South Semitic languages. In: Kaye AS, editor. Semitic studies in honor of Wolf Leslau. Wiesbaden: Harrassowitz. p. 1323-1336.

Rosenberg N, Mahajan S, Ramachandran S, Zhao C, Pritchard J, Feldman M. 2005. Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet. 1:e70.

Rosenberg, N, Pritchard J, Weber J, Cann H, Kidd K, Zhivotovsky L, Feldman M. 2002. Genetic structure of human populations. Science. 298:2381-2385.

Rosenberg N, Woolf E, Pritchard J, Schaap T, Gefel D, Shpirer I, Lavi U, Bonne-Tamir B, Hillel J, Feldman M. 2001. Distinctive genetic signatures in the Libyan Jews. Proc Natl Acad Sci USA. 98:858-863.

Sabeti P, Varilly P, Fry B, et al. (263 co-authors). 2007. Genome-wide detection and characterization of positive selection in human populations. Nature. 449:913-918.

Saitou N, Nei M. 1987. The neighbor-joining method for reconstructing evolutionary trees. Mol Biol Evol. 4:406-425.

Santos F, Pandya A, Tyler-Smith C, Pena S, Schanfield M, Leonard W, Osipova L, Crawford M, Mitchell R. 1999. The central Siberian origin for Native American Y chromosomes. Am J Hum Genet. 64:619-628.

Schurr T, Sherry S. 2004. Mitochondrial DNA and Y chromosome diversity and the peopling of the Americas: evolutionary and demographic evidence. Am J Hum Biol. 16:420-439.

Schurr T. 2004. The peopling of the New World: perspectives from molecular anthropology. Annu Rev Anthropol. 33:551-583.

Shackelton L, Rambaut A, Pybus O, Holmes E. 2006. JC virus evolution and its association with human populations. J Virol. 80:9928-9933.

Shen P, Wang F, Underhill P, et al. (13 co-authors). 2000. Population genetic implications from sequence variation in four Y chromosome genes. Proc Natl Acad Sci USA. 97:7354-7359.

Sherry S, Rogers A, Harpending H, Soodyall H, Jenkins T, Stoneking M. 1994. Mismatch distributions of mtDNA reveal recent human population expansions. Hum Biol. 66:761-775.

Shields G, Schmiechen A, Frazier B, Redd A, Voevoda M, Reed J, Ward R. 1993. mtDNA sequences suggest a recent evolutionary divergence for Beringian and northern North American populations. Am J Hum Genet. 53:549-562.

Silva W, Bonatto S, Holanda A, et al. (14 co-authors). 2002. Mitochondrial genome diversity of Native Americans supports a single early entry of founder populations into America. Am J Hum Genet. 71:187-192.

Smith N, Webster M, Ellegren H. 2002. Deterministic mutation rate variation in the human genome. Genome Res. 12:1350-1356.

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

Strimmer K, Pybus O. 2001. Exploring the demographic history of DNA sequences using the generalized skyline plot. Mol Biol Evol. 18:2298-2305.

Suchard M, Weiss R, Sinsheimer J. 2001. Bayesian selection of continuous-time Markov chain evolutionary models. Mol Biol Evol. 18:1001-1013.

Sugimoto C, Hasegawa M, Kato A, Zheng H, Ebihara H, Taguchi F, Kitamura T, Yogo Y. 2002. Evolution of human polyomavirus JC: implications for the population history of humans. J Mol Evol. 54:285-297.

Sugimoto C, Hasegawa M, Zheng H, et al. (17 co-authors). 2002. JC virus strains indigenous to northeastern Siberians and Canadian Inuits are unique but evolutionally related to those distributed throughout Europe and Mediterranean areas. J Mol Evol. 55:322-335.

Sugimoto C, Kitamura T, Guo J, et al. (19 co-authors). 1997. Typing of urinary JC virus DNA offers a novel means of tracing human migrations. Proc Natl Acad Sci USA. 94:9191-9196.

Swadesh M. 1955. Towards greater accuracy in lexicostatistic dating. Int J Am Linguist. 21:121-137.

Szathmary E. 1993. Genetics of aboriginal North Americans. Evol Anthropol. 1:202-220.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics. 105:437-460.

Tamm E, Kivisild T, Reidla M, et al. (21 co-authors). 2007. Beringian standstill and spread of Native American founders. PLoS ONE. 2:e829.

Templeton A. 1998. Human races: a genetic and evolutionary perspective. Am Anthropol. 100:632-650.

Thompson J, Higgins D, Gibson T. 1994. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673-4680.

Tishkoff S, Varkonyi R, Cahinhinan N, et al. (17 co-authors). 2001. Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science. 293:455-462.

Torroni A, Schurr T, Cabell M, Brown M, Neel J, Larsen M, Smith D, Vullo C, Wallace D. 1993a. Asian affinities and continental radiation of the four founding Native American mtDNAs. Am J Hum Genet. 53:563-590.

Torroni A, Sukernik R, Schurr T, Starikorskaya Y, Cabell M, Crawford M, Comuzzie A, Wallace D. 1993b. mtDNA variation of aboriginal Siberians reveals distinct genetic affinities with Native Americans. Am J Hum Genet. 53:591-608.

Twiddy S, Holmes E, Rambaut A. 2003. Inferring the rate and time-scale of dengue virus evolution. Mol Biol Evol. 20:122-129.

Underhill P, Shen P, Lin A, et al. (21co-authors). 2000. Y chromosome sequence variation and the history of human populations. Nature Genetics. 26:358-361.

United Nations, Department of Economic and Social Affairs, Population Division, 2003. World Urbanization Prospects: The 2003 Revision. www.un.org/esa/population/publications/wup2003/2003wup.htm

United Nations, Department of Economic and Social Affairs, Population Division, 2004. World Population Prospects: The 2004 Revision Analytical Report. www.un.org/esa/population/publications/WPP2004/WPP2004_Volume3.htm

Venter J, Adams M, Myers E, et al. (275 co-authors). 2001. The sequence of the human genome. Science. 291:1304-1351.

Voight B, Kudaravalli S, Wen X, Pritchard J. 2006. A map of recent positive selection in the human genome. PLoS Biology. 4:e72.

von Wissman H. 1975. Uber die fruhe Geschichte Arabiens und das Enstehen des Sabaerreiches. Wien: Austria Academy of Sciences Press.

Wang S, Lewis C, Jakobsson M, et al. (27 co-authors). 2007. Genetic variation and population structure in native Americans. PLoS Genet. 3:e185.

Waters M, Stafford T. 2007. Redefining the age of Clovis: implications for the peopling of the Americas. Science. 315:1122-1126.

Weale M, Weiss D, Jager R, Bradman N, Thomas M. 2002. Y chromosome evidence for Anglo-Saxon mass migration. Mol Biol Evol. 19:1008-1021.

Weber T, Major E. 1997. Progressive multifocal leukoencephalopathy: molecular biology, pathogenesis and clinical impact. Intervirology. 40:98-111.

Webster M, Smith N, Ellegren H. 2002. Microsatellite evolution inferred from human-chimpanzee genomic sequence alignments. Proc Natl Acad Sci USA. 99:8748-8753.

Weinberg GL. 1995. A World at Arms: A Global History of World War II. Cambridge: Cambridge University Press.

Wilson A, Cann R, Carr S, et al. (11 co-authors). 1985. Mitochondrial DNA and two perspectives on evolutionary genetics. Biol J Linn Soc. 26:375-400.

Wooding S. 2001. Do human and JC virus genes show evidence of host-parasite codemography? Infect Genet Evol. 1:3-12.

Wright S. 1931. Evolution in Mendelian populations. Genetics. 16:0097-0159.

Xiao F, Yotova V, Zietkiewicz E, et al. (11 co-authors). 2004. Human X-chromosomal lineages in Europe reveal Middle Eastern and Asiatic contacts. Eur J Hum Genet. 12:301-311.

Zazula G, Froese D, Schweger C, Mathewes R, Beaudoin A, Telka A, Harington C, Westgate J. 2003. Ice-age steppe vegetation in east Beringia. Nature. 426:603.

Zhang J. 2007. The drifting human genome. Proc Natl Acad Sci USA. 104:20147-20148.

Zheng H, Zhao P, Suganami H, Ohasi Y, Ikegaya H, Kim J, Sugimoto C, Takasaka T, Kitamura T, Yogo Y. 2004. Regional distribution of two related Northeast Asian genotypes of JC virus, CY-a and -b: implications for the dispersal of Northeast Asians. Microbes Infect. 6:596-603.

BIOGRAPHICAL SKETCH

I graduated from Homestead High School in Fort Wayne, Indiana, in spring 1997. I then attended the The Johns Hopkins University from fall 1997 to spring 2001, and graduated with a B.S. in biomedical engineering and a concentration in computer science. I began my graduate career in September 2001 when I matriculated at the University of Oxford (Hertford College) and took my M.Sc. in biology in 2002. I then enrolled in the anthropology graduate program at University of Florida in fall 2002, received my M.A. in May 2004 and my Ph.D. in August 2008.