Geodatabase for Archaeogenetics:

Ancient Peoples and Family Lines

by

Maria Leasure

A Thesis Presented to the
FACULTY OF THE USC DORNSIFE COLLEGE OF LETTERS, ARTS AND SCIENCES
University of Southern California
In Partial Fulfillment of the
Requirements for the Degree
MASTER OF SCIENCE
(GEOGRAPHIC INFORMATION SCIENCE AND TECHNOLOGY)

August 2021

To Jason, Qing'an, and Valentino for your faith, love, and support.

Acknowledgements

I would like to express my appreciation for conversations with Curtis Rogers concerning the relatively new field of genetic genealogy, and the volunteer work done by Open Street Map contributors in historical map projects.

I am indebted to Professor Laura Loyola for the insight and initial ideas to formulate this topic. I equally owe deep thanks to Professor Wu An-min, for inspiring an interest beyond "mapping" into the data and necessary structures and continuing to guide as a committee member, and to Professor Chiang Yao-yi, for generously sacrificing time to serve on the committee and share his expertise and important critiques. I am very grateful to Professor Jennifer Bernstein, my advisor, for so many improvements, motivation, and strong moral support.

The end of this project would not have come without a strong beginning, for which I thank Professors Darren Ruddell and Vanessa Osborne, who helped me overcome anxieties and direct my first efforts, and Robyn McNabb and Dr. Robert Vos for their belief and encouragement. A special thanks to Professor "Kirk" Oda, who taught me to love technical challenges and to learn from mistakes, and to Professor Elisabeth Sedano, who also helped me become more adaptable and appreciate the curveballs. All these individuals have improved not only a MS project but a person.

I am grateful to my colleagues, Samantha Bamberger and Timothy Williams, who helped accommodate my schedule to focus on this work at critical times, and from whom I have learned immeasurable amounts, and my classmates and friends at USC, who are unfailingly kind and ready to help. For a sounding board on all phases of this work and support in every possible way, I thank Jason Leasure.

Table of Contents

List of Tables

List of Figures

Abbreviations

BCE	Before the Common Era
CE	Common Era, equivalent to AD *anno domini*
DEM	Digital Elevation Model
DNA	Deoxyribonucleic acid
GIS	Geographic information system
GISci	Geographic information science
kya	Thousands of years ago
mtDNA	Non-recombining DNA from the mitochondria
SSI	Spatial Sciences Institute
SNP	Single Nucleotide Polymorphism
USC	University of Southern California
Y-DNA	Non-recombining DNA from the Y chromosome

Abstract

From its early beginnings from the parent genus Homo in Africa, the species *Homo sapiens* spread across the globe to every continent except Antarctica, long before the advent of large seafaring vessels or even the wheel. The dispersion of the first *Homo sapiens* occurred when other early human species, such as Neanderthals or Denisovans, were still in Europe and parts of Asia, and land features and climates were very different from northern and eastern Africa. As early modern humans encountered these new environments and possibly other, earlier peoples over centuries of migration, adaptations occurred, and new cultures arose. These migrations are of great interest to several disciplines, including physical anthropology, archaeology, and genetics. A global geodatabase as a repository of spatial and genetic data to facilitate Spatio-temporal models of models and various data visualizations would serve all these disciplines. Such a geodatabase also can incorporate other related data for investigation, such as global regions, early coastlines, glacier limits, or the overall continental geography of earlier ages for investigation of their possible effects on movement or settlement of ancient peoples. Additionally, a geodatabase offers many options to share or limit access to data. This project offers a comprehensive data source and tool for creating and sharing analyses with other research efforts.

Chapter 1 Introduction

Many species remain forever restricted to a certain habitat essential to their survival, only to be found in one region on the earth. Others, like our species *Homo sapiens*, succeed in adapting to diverse conditions and expand their range across multiple continents. Humans dispersed early in their history and developed many different material cultures in response to these varied environments, as seen in tools, weapons, and other items discovered today. They also developed great diversity in other aspects of culture, speaking numerous languages for communication, establishing a variety of religious belief systems, and creating unique forms of art. To understand the impact of these migrations and the rich diversity of our species, specialists in biology, chemistry, genetics, anthropology, archaeology, and other fields study not only the artefacts left by early peoples but their remains. Work specifically related to their genetic markers – sometimes called archaeogenetics – is adding a new dimension to these studies.

1.1.1. Outline

This chapter first introduces key problems in studies of ancient peoples, their cultures, and the impact of archaeogenetics on established ideas. In section 2, the chapter discusses the objective of fulfilling unmet needs of consolidated data for research, which will provide for essential spatial analysis (described in sub-section 1.2.1) and enabling wider scale studies (1.2.2). The advantage of the geodatabase format to these goals is addressed in sub-section 1.2.3. Section 3 shares the initial motivations for building a database as a tool inspired by the visualization of migrations (1.3.1) and by a successful application of a spatial database in genetics (1.3.2). Section 4 expands the objective and motivations by describing the geodatabase's applications and intended users in more specific scenarios, including unexpected global connections (1.4.1), increased multi-disciplinary work (1.4.2), and geographic distributions studies (1.4.3). Section 5

gives a basic methodology description and presents a more general outline and overview for the project in 1.5.1.

1.2. Objective

This geodatabase will bridge the efforts of researchers in different fields investigating early origins, migrations, and relationships of ancient peoples by offering a tool useful for multi-disciplinary and global spatial analysis. Many works have been published on single excavation sites or regional investigations, or geographically broader studies of specific time periods. However, a comprehensive, consolidated resource for reference and analysis is lacking in a format that can be readily used by software tools. The geodatabase is intended to be such a resource, by providing a unified spatial catalogue of a variety of data collected from the excavation and laboratory investigation of ancient human remains.

1.2.1. Geodatabase Design to Aid Spatial Analysis

A comprehensive spatial catalogue of data that can be queried will aid spatial analysis in migration and cultural studies. For example, to what extent past languages and cultural practices were spread through observations or exchanges between neighboring groups and how much was a product of actual migration is still researched and debated by Cavalli-Sforza (1994), Renfrew (2014), and others. Spatial analysis can determine which scenarios are more likely. However, without a unified resource for the cultural and genetic data necessary, this analysis and visualization is more difficult and time-consuming. A geodatabase is a valuable tool to provide structure for an analytic process. This geodatabase will include some data concerning the cultural identification of individuals whose DNA is catalogued here. The design of the geodatabase will allow for the future incorporation of cultural data sets, such as extents of spoken languages. Evidence of early genetics can support or cast doubt on theories in other studies of ancient

peoples, such as reconstructed migrations based upon linguistics. A genetic geodatabase provides storage and structure to use spatial analysis with this evidence.

1.2.2. Scope: Enabling Wider-Scale Studies

This project aims to include archaeological finds dated from the oldest Paleolithic discoveries to Iron Age peoples at the beginnings of recorded human history, from every continent, to provide a technology-friendly resource that is not limited in scope to a single place, time, or culture. The geodatabase also will include genetic marker data in addition to cultural assessments, estimated dates, and the spatial data to map the archaeological finds.

Existing publications generally are restricted to one location or subdivision of an epoch to allow the focus required for detailed study, such as that of a Later Stone Age burial site near Eulau, Germany (Haak et al. 2008). In Eulau, the in-depth investigation of several members of a family group revealed many insights into the lives of early peoples, including their social structures and family practices, such as a pattern of men seeking partners outside the local area. These localized, more granular studies are necessary to determine social patterns and interactions of ancient peoples. However, although they are important as evidence to form and evaluate theories on a wider scale, some effort to aggregate them must be made for this broader analysis. This geodatabase project gathers data from many separate studies of this type and unifies them for analysis on a larger scale, both in terms of time and place.

1.2.3. Advantages of Geodatabase Format

The geodatabase has several distinct advantages to researchers besides serving as a consolidated source of data. The geodatabase is superior to a simple folder collection of geojsons or kmls for example, because unlike files that use WGS84 (“Converting GIS Vector Data to KML | Keyhole Markup Language” n.d.; “RFC 7946 - The GeoJSON Format” n.d.), the feature

classes can be reprojected into a coordinate system more suitable for smaller regional or local analysis. The shapefile format shares this advantage, but again the geodatabase has advantages in organization. A shapefile is not only a single .shp but a collection of files, including .dbf and .shx, and if these components are missing, the shapefile is useless. The shapefile also could include other files, such as .cpg or .prj, and a loss of these files may cause issues with performance. When transferring a collection of shapefiles, care must be taken to not only move the primary .shps, but all their other essential components. A geodatabase provides one item to copy, move, or share over a network. Data stored and organized in a geodatabase potentially could simplify data retrieval and preparation for use in other software tools besides GIS desktop programs, thanks to its structure.

Many static or interactive maps with archaeogenetic themes are available, but unlike these, a geodatabase makes use of aspects of relational databases to allow for queries, kernel density estimation, and other operations beyond simple visual presentation. Because it can be accessed by the most used GIS desktop tools, the geodatabase can be used to test new models of migrations and spatial distributions of ancient groups. The aspects of relational databases can be exploited to simplify the analytic process; by setting up initial permanent relationship classes, repeated relate and join operations can be avoided.

These relationship classes can do more than temporary joins and relates. They also establish a referential integrity, such as when attributes are changed then their related objects may be automatically updated. The relationship class aids the edit process by granting quicker access to related objects. The advantage of editing one entry in a related table in a geodatabase over locating and updating multiple instances across several tables and joined shapefiles is obvious, both for convenience and maintenance overhead as well as data integrity.

1.3. Motivations

Clues to the daily lives of ancient peoples are present not only through the evidence of the objects they left behind but also their remains. The development of genetic analysis has added another component to the investigation and analysis of early human remains. There are several theories concerning the migrations of our earliest ancestors in Africa and their first journeys to other continents (Beyin 2011). Despite new techniques in analysis and a growing body of literature, no established global database of early human finds yet exists with the required data for this multi-disciplinary spatial analysis.

1.3.1. Spatial Analysis Essential to Migration Studies

A study of any species must consider its environment; such a study without a sense of *place* is incomplete. It naturally follows that these studies of early peoples would make extensive use of spatial analysis and visualization to reconstruct the patterns of migration and settlement. Beyin's (2011) mapping of some of the earliest human migrations out of Africa is one such reconstruction. Some research teams have used genetic markers of human remains and comparisons with modern-day haplogroup distributions to find possible paths of migration not previously considered (Reich et al. 2012; Achilli et al. 2013). A geodatabase of the data used by these researchers would be useful to replicate their efforts and expand upon these ideas.

1.3.2. Spatial Database Precedent in Related Fields

The investigation of archaeogenetics, or the DNA characteristics of early peoples, is important to many investigations of ancient groups and their migrations. One aspect of archaeogenetics is the distribution of various haplogroups. Several efforts to map the occurrence of these haplogroups geographically in modern populations have been successful. In 2005, McDonald of the University of Illinois published maps showing the percentage of various DNA

haplogroups present today in different regions of the world. More recently, a database for the present-day distribution of mtDNA, the non-recombinant DNA passed exclusively through the maternal line, has been created (Rasheed et al. 2017) called "mtDNAmap". It is a spatial database implemented using MySQL. Access to the database is not limited to academics, as Rasheed's team has enabled access to the public via the Internet at URL http://www.dnageography.com/mtDNAmap.php.

1.4. Intended Users and Applications

New research by specialists in genetics, anthropology, and anthropology challenges long-standing theories on the arrival of modern humans and their arrival in certain geographic areas. The dominant theory of human origins is that of an African origin, often assumed to be one location from which early humans spread to the rest of the continent, but work in north Africa suggests that from earliest times the species was found in more than one region (Hublin et al. 2017). Genetics work also has caused a re-evaluation of the traditional theory on human settlement of the Americas by one wave of hunter-gatherers from Siberia by way of Beringia, a former landmass now covered by the waters of the Bering Strait (Achilli et al. 2013; Reich et al. 2012; Wei et al. 2018; Yang et al. 2010). Many opportunities for continued inquiry in these and other regions of the world remain and would benefit from a geodatabase to categorize, analyze, and visualize data.

1.4.1. Exploring Global Connections

For many years, the early 1900s view of Aleš Hrdlička that early peoples from Siberia came across a land bridge into the Americas in one large wave of migration (Hrdlička 1907; Hrdlicka 1936) remained pre-eminent. Later researchers have found evidence to challenge this, some even finding a three-wave view too simplistic (Arias et al. 2018; Achilli et al. 2013; Reich

et al. 2012). Teams led by Yang (2010) and Skoglund (2015) also found evidence for migrations over water, rather than a slow overland trek, and some evidence of a link between South America and Polynesia emerged in other genetic work by Gonçalves et al. (2013). A global geodatabase of human samples and their genetic attributes could assist in the exploration of links between widely separated incidences of genetic markers.

1.4.2. Multi-Disciplinary Work

Researchers have explored the movements of peoples in Europe and Asia using in-depth analysis of genetics, while also considering archaeologists' research of cultural artifacts and linguists' constructions of language trees. Prominent scholar (Gimbutas 1963) theorized that a linguistic wave originating from a unified group (i.e. the Kurgan culture) crossed from the Pontic steppes to southern and western regions of Europe. Cavalli-Sforza also created reconstructions of the migrations of early peoples throughout Asia and Europe using the study of languages alongside archaeogenetics (Cavalli-Sforza 1994). During the early years of archaeogenetics, researchers Sokal, Oden, and Thomson (1992) countered that neither large scale Pontic steppes migration nor other competing theories completely explained language distribution in Europe. When geography was held constant, Sokal, Oden and Thomson (1992) found a "markedly lower" correlation between language and genetics, which they admitted was still statistically significant. However, later researchers with data from newer genetic extraction and analysis techniques found evidence supporting the earlier ideas of a group of related, migrating people carrying their language from the steppes into new territories (Krzewińska et al. 2018). Investigation including geospatial analysis using new genetic data can either contradict or support earlier theories derived from work in physical anthropology, linguistics, or other fields.

1.4.3. Geography and Spatial Distributions

Even though genetic studies can reconstruct trees of haplogroups and clades, the role of geography in understanding pathways and motivations for migrations cannot be overstressed. Yang et al. (2010) found that dispersal of genetic markers from the extreme northwestern parts of the Americas correlated to a least-cost path using coasts as facilitators and looked forward to future studies on the impact of other major geographic features such as mountains and river valleys. Baumann (2017) used agent-based models in similar explorations of the impact of coastline navigation in the Mediterranean. Baumann (2017) constructed these models to assess the feasibility of using early seacraft to fish, hunt, or settle in islands during the earliest ages of humanity – including times when earlier, related species as *Homo neanderthalensis* existed in large numbers. Previously, scientists believed that such early peoples could only journey over land given the level of technology and seafaring knowledge supposedly held at that time (Baumann 2017). This project will include basic geographic data, but its geodatabase structure can easily accommodate raster or vector files delineating ancient coastlines, glacial expanses, and other potential factors that facilitate or impede migrations.

1.5. Overview of Methodology

The organization of data from these finds into a geodatabase could aid anthropologists and genetic researchers in several ways. Users can identify each archaeological find uniquely and within a spatial context. The geodatabase provides an extensible set of attributes, easily updated to accommodate newly published GIS files and tables. This is a more uniform and simple method to update data quickly and accurately than multiple separated GIS files and tables, since options like domains to limit input to acceptable values or ranges exist. An additional advantage to a geodatabase over a loose collection of separated tables, shapefiles

(.shp) or keyhole markup language (.kml) layers, or images is that use of a geodatabase can facilitate SQL queries by providing a unified set of tables and establishing their relationships. Finally, our solution can scale from a file geodatabase to an enterprise geodatabase if needed in the future to allow for multiple users with versioning and access or editing controls.

1.5.1. Overview

This study will continue with the review of work of researchers from multiple disciplines in chapter 2, to provide understanding for the types of analysis and data used to investigate the key problems in human origins and migrations described in this chapter. The second chapter also describes literature used to explore design and use of geodatabases in similar applications from various related fields. Chapter 3 provides details of the methodology, including design and the data sets used in the project. Chapter 4 shares the results of the geodatabase project, and ideas for future work and desired improvements are outlined in chapter 5.

Chapter 2 Related Disciplines and Literature Review

The exploration of human dispersion and migration involves the disciplines of genetics, biology, anthropology, archaeology, and history. Literature concerning the creation and application of geodatabases in the service of research questions in these related fields was essential. To better inform the process of building a geodatabase for studies of early human populations, literature involving ancient peoples, their migrations, and genetic markers also was reviewed.

Section 2.1 explores the questions and controversies that remain in the study of early human origins and the use of genetics as one component of related studies. Sections 2.1.1 and 2.1.2 establish the use of Y-DNA and mtDNA markers as useful tools for current analysis of the movements of ancient peoples and cement their inclusion in the data sets for this project. Section 2.2 outlines the use of geodatabases in a variety of scientific fields, providing models for this project. Sections 2.2.1 and 2.2.2 specifically apply the use of databases to the question of distributions in biology and to cataloguing archaeological samples, which have close parallels with aspects of this work. Section 2.2.3 highlights the benefits of a digital approach in general, and in particular one that enables the use of the Internet, to data sharing for historical research. Section 3.1 describes the structure and beginnings of the Esri geodatabase format, and Sections 3.2 and 3.3 offer some insight into the benefit and practicality of selecting this format for this project.

2.1. Genetics in Population Studies

Although the origins of our genus lie in Africa, the first development of the modern form of humans, *Homo sapiens*, and its arrival on different continents continue to spark many questions. Geneticists and anthropologists have looked at different types of evidence to find

answers to these questions and support various positions. For decades there has been some support for a multi-regional hypothesis of humans arising from earlier related species, but others opt for a somewhat similar "candelabra" hypothesis. Many others support the idea of a single origin of modern humans differentiating themselves from earlier, related humans to become a distinct new species, *Homo sapiens* (Hublin et al. 2017; C. B. Stringer and Andrews 1988; C. Stringer and Andrews 2005; Wilson and Cann 1992).

Genetics work contributes strongly to the growth of the single origin theory (Gibbons 1997), as researchers can trace the origin of the so-called 'Adam' and 'Eve' of Y-DNA and mitochondrial lines of all living humans and locate them in the African continent. However, genetics studies reveal a combined ancestry of *Homo sapiens* and *Homo neanderthalensis* in some regions (Pääbo 2014). The intermingling of closely related yet technically separate species is part of the foundation of the multi-regional and candelabra hypotheses. Furthermore, work involving agent-based modeling and migrations admits the possibility that human beings prior to *Homo sapiens* could have crossed bodies of water to find food sources or settle (Baumann, n.d.; Yang et al. 2010), not necessarily restricting the travels of earlier forms of our genus to slow, overland treks. Although genetics has become an important part of investigation into early human history, genetic analysis has not yet conclusively provided the answers to questions of the first origins and expansions, and other disciplines and their methodologies will continue to play an important role.

The single-origin or "out of Africa" theory usually ascribes an East African origin to earliest *Homo sapiens*, then dispersion into other regions and continents, a journey as summarized by Beyin (2011) as an origin in East Africa, then migration to Arabia, SE Asia, and

the Levant sometime after 150 kya. Eventually Europe and NW Asia would become inhabited. Whether from one common origin, as Beyin (2011) describes, or from neighboring regions already populated in part by our earlier ancestors and relatives, modern humans soon dispersed across vastly different environments. Studies of these migrations frequently use genetic markers in addition to other types of data to trace their paths out of regions and environments, such as the use of hundreds of thousands of SNPS, rather than only mtDNA and Y-DNA haplogroups, to reveal possibly three streams of Asian descent into the Americas at least 15 kya (Reich et al. 2012).

2.1.1. The Primacy of Nonrecombinant Markers

The uninterrupted "father line" and "mother line" genetic markers inherited, Y-DNA and mtDNA respectively, are frequently the genetic markers used for tracing early human journeys (Grugni et al. 2019). Although nuclear DNA is used to study the genetic evidence of the early *Homo sapiens,* as in the Reich (2012) study of the Americas, and also is useful for tracking hybrid population (e.g. offspring of *Homo sapiens* and *Homo denisova*) movements, mitochondrial DNA (mtDNA) and Y chromosome DNA (Y-DNA) continue to figure prominently in these kinds of geographic dispersal studies. These markers, particularly examination of mtDNA haplogroup M dispersals, were used in Beyin's (2010) examination of Upper Pleistocene movements out of Africa. As Underhill and Kivisild (2007) remark, neither mtDNA nor Y-DNA can be considered informative concerning the first speciation event of modern humans from earlier types, but they remain "the most well resolved genetic loci for the study of population histories since the out-of-Africa migration." These types of markers have been useful in following biological evidence of demic diffusion compared to the diffusion of

cultures, evidenced by characteristics of artifacts and languages, such as the question of Indo-European (Gimbutas 1963).

2.1.2. Objections to Reliance on Non-Recombinant Markers

The ancient DNA of a group may be persistent, even if a direct father-to-son (Y-DNA) or mother-to-daughter (mtDNA) lineage becomes extinct. These ancestors continue to pass down other forms of DNA to descendants of the opposite sex for generations, allowing for earlier ancestors to potentially be studied using nuclear DNA techniques (Reich et al. 2012). Furthermore, Y-DNA and mtDNA account for only a small percentage of each human's individual genome. There are some limitations and valid concerns relating to exclusive reliance on these types of genetic markers (Arias et al. 2018; Reich et al. 2012), but mtDNA and Y-DNA remain important, established markers for analysis and will form the basis of genetic data for inclusion in the geodatabase.

Regardless of the necessarily limited picture drawn from these lines, until techniques in sequencing and analyzing ancient DNA improve these nonrecombinant markers likely will remain prominent in research. Concerning analysis and genetic trees constructed of the other types of nuclear DNA, Underhill and Kivisild (2007) point out that such trees may be robust under some conditions. However, they also note that they are of low molecular resolution, which provides a lesser understanding of our genetic history during "the pivotal past 100,000 years, which is the time window of interest for most of the human migrations." (Underhill and Kivisild 2007)

2.2. Geodatabases as Aids to Analysis

Many researchers have explored geodatabases within their respective areas to aid analysis and visualization, in addition to providing a basic digital chronicle with a spatial component. They have proven useful both in physical sciences, social sciences, and history. For example, an Italian research team led by Viciani (2018) built a geodatabase for a botanical study, and researchers Szabo et al. (2018) created a geodatabase for use in examining the history of a forest. Archaeologists established numerous archaeological online databases - including some spatial databases - for studies of ancient cultures in the Middle East. Rodriguez (2019) created a geodatabase based on historical records to successfully explore trends in Mexican migration into the United States. The study of ancient human migrations involves techniques from biological sciences, archaeology, and social sciences, so successful use of a geodatabase in related fields of study support its use in this project and provide models for design and implementation.

2.2.1. Geodatabases in Biology

Biodiversity or species distribution studies benefit from a geodatabase approach. The idea of building a geodatabase to aid analysis of biodiversity in vegetation was proposed by (Viciani et al. 2018) who expected to gain "…identification of spatial patterns of species richness and of sampling effort..." among its benefits, as well as the examination of possible relationships between species distribution and topographic factors. This approach was designed for the Parco Nazionale delle Foreste Casentinesi, near Campigna, Italy, to support conservation studies, but its methodology and research benefits could likewise be applied to studies of virtually any species, either plant or animal, on land. For this vegetation study, (Viciani et al. 2018) the team used roughly 680 reports from field studies using a hand-held GPS to collect spatial data on

vegetation species in the park in addition to literature data, much of which was lacking in specific geographical data.

Viciani's team obtained a digital elevation model (DEM) to investigate the topography of the park alongside the various specimen data. Topography can be important to study of plant distributions, since elevation, aspect, and slope can influence the growth of various species. It is possible to imagine that the propagation of a plant species could be influenced by topological conditions allowing or hindering seeds or roots to spread and increase the range of that variety. The ability to incorporate DEMS or other raster files as part of analysis is useful not only in botany, but other life sciences like anthropology, as demonstrated by Baumann's work (2017) in agent-based modeling of human migrations.

2.2.2. The "Internet Museum" and Digital Collections

The use of digital inventories with spatial data capabilities are not new to management of archaeological finds. Many museums make use of EMu software by Axiell to keep inventories of these artifacts and artworks. This allows institutions to share details of their collections with others around the world and to maintain accountability for the locations of each item. IMu Maps and IMu Object Locator allow for the mapping of objects in collections using floor plans ("EMu – Collections Trust" n.d.).

Users can also use EMu for other diverse functions, such as tracking requirements for repatriation of human remains samples, as required by some nations, or monitoring problems areas that threaten the safe storage of objects in the collections (e.g. insect pests or mildew). However, EMu users chiefly seem to use the system for collections management, and do not appear to be using its developing spatial capabilities to carry out spatial analysis using the object

attributes filed in the database (Axiell 2018). This system, as currently implemented, cannot be used in the place of a geodatabase for study of early genetics and migrations.

The use of geodatabases can aid researchers in locating considerable amounts of data and easily visualizing it spatially, compared to other forms of web publication. Morrish and Laefer as early as 2010 advocated for the use of web-enabled architectural heritage inventories. Morrish and Laefer (2010) noted the considerable resources to compile inventories, resulting in many "tomes" yet they were not fully utilized because of accessibility and ease of use.

In 1999, the historical preservation organization English Heritage began a project to digitize and publish on the web. The system followed its original survey structure, with independent entries lacking integration into a mapping interface or advanced inquiry tools (Morrish and Laefer 2010). Monuments or sculptures commissioned during certain periods could be specified with a search of the system, for example, but Morrish and Laefer (2010) pointed out that its structure did not allow for a geographic visualization of their distribution. The geodatabase format, like the one designed for this project, easily would allow for both these queries and visualizations with proprietary and open source GIS tools.

2.2.3. Databases and Geodatabases in Archaeology

The approach of a database for archaeological research was investigated by Drzewiecki and Arinat (2017). Drzewiecki and Arinat (2017) surveyed archaeological professionals in Jordan regarding their use of databases for archaeology, and their concerns about utility, access, and reliability. The survey by Drzewiecki and Arinat (2017) included all types of online databases used in Jordanian archaeology studies, but at least one (MEGA-J) had full spatial capabilities to display polygons. The use of databases had become very important to the

archaeologists, according to the researchers' findings, notwithstanding some reservations regarding reliability of the data. Most archaeologists would also conduct a literature review to verify what was retrieved from the databases.

One important consideration illuminated by Drzewiecki and Arinat (2017) was that improved locational accuracy could be a negative, in that the databases could serve as a "catalogue of 'looting-to-order'". They added, however, that limiting access would only slow down professional looters in their efforts to steal artifacts, not eliminate the problem. Drzewiecki and Arinat (2017) summarized respondents in stating that the benefits of wide-spread and open access to the databases outweighed the looting risk, an ever-present threat to sites. These concerns account for the lower precision of coordinates used in this project, as journals typically publish these because they would not increase this threat.

A site in Bethsaida, Israel, was the subject of efforts by Burrows in 2016 to create an archaeology-specific geodatabase, designed to improve accuracy and utility when used by personnel of various levels of skill or experience. Fieldwork at the site created an "unwieldy" amount of data to record, according to Burrows, but modern digital methods were surprisingly not deployed frequently to handle the problem. Burrows began constructing a geodatabase to log the location of the artifacts discovered and accurately record characteristics in a systematic way that could be easily reviewed using SQL queries. Both legacy data and new field study data went into populating the database. The design and adoption of this geodatabase allowed use of handheld GPS devices ((Burrows, n.d.) to be used to record more precise and accurate spatial data and immediately describe basic attributes for the finds in the field.

This archaeogenetic geodatabase project necessarily will use less precise data, as it currently relies upon the general locations of excavation sites in published papers. These seldom offer the actual coordinates of each find, since as Drzewiecki and Arinat (2017) pointed out, there is a real need to protect research locations, at least until work has been totally completed in the field. However, the lack of precision for entries in this geodatabase is not an obstacle for the project because unlike Burrows' project (n.d.), it is not for the purpose of locating the sites or additional specimens, but for spatially visualizing the patterns in related data on wider, regional levels.

2.3. Developing the Geodatabase

2.3.1. Beginnings - Esri

The geodatabase (.gdb) is the native data format used by Esri, and is the usual data format for editing and data management in Esri tools ArcMap, ArcGIS Pro, etc. Esri developed the geodatabase (geographic database) to store both spatial and attribute data together simultaneously in a single database management system. This integration of the relational and object-oriented concepts in geodatabases allows its single management system storage for both spatial data and attributes, according to (Twumasi 2002). Before this, the data model for databases isolated geographic data from attributes of entities, for example ArcInfo used separate files for attributes, called INFO tables, and spatial data was kept within indexed binary files (Twumasi 2002).

2.3.2. Spatial Data Model

The geodatabase makes use of a spatial data model, which is divided first into either raster/field-based or vector/object-based approaches. In turn, the raster approach may follow a

grid or regular tessellation scheme, or it may opt for an irregular scheme using variant sizes in partitions. The vector approach may opt for an unstructured or "spaghetti" model or a structured topologic model, the latter of which is more popular as it can offer a rich topology (Worboys 1995, Twumasi 2002) although it is less simple than an unstructured model.

2.3.3. Not Limited to Proprietary Format

The primacy of Esri in the GIS world is hard to deny, but many organizations have begun to embrace other tools to support of open source, to have an alternative to expensive licensing, and the ability to work in a more diverse set of formats, among other reasons. Some believe that the geodatabase cannot be used with non-Esri tools, but this is not the case. Recent versions of QGIS, past version 3.0, can read a geodatabase created by current versions of ArcGIS ("Opening File Based Geodatabases in QGIS 2.4 • North River Geographic Systems Inc" 2014). The Open Source Data Manager can read the directory using "OpenFileGDB" as type. As early as 2015, GDAL versions were able to read and extract information from .gdb files ("Working with File Geodatabases (.Gdb) Using QGIS and GDAL | Geospatial @ UCLA" n.d.), so QGIS and other tools using GDAL can work with a geodatabase system. The ability to use the geodatabase with other platforms makes it a suitable choice for a spatial database, since it can be accessed by other popular applications if needed.

Chapter 3 Data and Research Methods

This chapter describes data and methods for design and construction of the geodatabase of ancient peoples and genetic markers. This part of the project is divided into roughly four stages, described in sections 3.1 through 3.4. The first stage described in 3.1 is the initial determination of data needs and plans for data acquisition. Section 3.2 describes decisions made between the various database options with spatial capabilities, and for software in general. Section 3.3 outlines the main design and concept of the database, including table relationships. The fourth stage described in section 3.4 involves exploration of the data within the database structure and testing the proof of concept. It outlines plans for SQL queries and basic visualizations to evaluate its performance and fitness to its purpose. These stages initially were carried out in sequential order but are repeated in a "feedback loop" as shortcomings or possible improvements to the geodatabase were identified.

3.1. Data Needs and Acquisition

The data for the project consists of three main kinds: the archaeological find data (e.g. the specimens themselves and their characteristics), male and female line genetic data, and geographical data. The ideal was to have a wide scope and include global data for all three main divisions. The human remains data were obtained from published papers by anthropologists or archaeologists that included results of genetic analysis of mtDNA or Y-DNA haplogroups. Reference tables for all the branches of the human genetic tree would need to be located or created to include the genetic data in a normalized form suitable for an expandable database project. Several public references from the International Society of Genetic Genealogists and companies offering family DNA products (i.e. FamilyTree DNA) aided in the construction of these genetic data sets when none could be located in a format suitable for use in a geodatabase.

The geography data was the easiest of the three to obtain, as multiple open source and Esri data sets for countries, regions, and even past geological ages of the earth were available.

3.1.1. Spatial Context

A data set of world regions was included to provide a geographic context to the other data sets. This 'region' data set came from Esri's data portal. It is a shapefile of polygons with the typical fields of OID, name, shape length, shape area, square kilometers, and square miles. Data types include object ID, geometry, text, double, double, double, and double, in order. This will be converted into a feature class for the geodatabase.

3.1.2. Finds – the Central Data Set of Ancient Human Specimens

The central data set, the finds, was anticipated to be the most difficult to obtain. Early research showed that many papers described the needed data in narrative form, and did not include tables, charts, JSONs, or formats that could be easily input into a database by automated means. Furthermore, most papers described a single excavation site. These seldom contained remains of more than five individuals, usually members of a family group. One exception was the work by a team in 2017, Matheisen et al., who published a study on a whole collection of hundreds of European finds and shared an Excel table with the designators, genetic data, locational data, and many other important attributes to examine the temporal element, such as identified culture or carbon dated estimates of age. Wherever possible, resources like this are used to import data with less possibility of error than manual entry.

Since smaller and less easily formatted sources of data were the rule, work began immediately on the data acquisition stage of the project. A timetable was set for 18 months to locate and complete at least a basic compilation of the data, with a target of 300 entries in a form for preliminary testing. Collection later exceeded this goal, with a total of over 2,700 entries for

the finds data set, more than 300 different Y-DNA clades (younger and usually smaller haplogroups "downstream" from older, usually larger, "downstream" related haplogroups) and over 340 distinct mtDNA clades or haplogroups. It is anticipated that some continued acquisition and import into the database will be ongoing until the end stages of the project. A sample of the initial spreadsheet used to compile data from published papers is provided in the table below. The spreadsheet also used fields to record latitude and longitude coordinates for the find locations ("findloc") and notes on the culture period to which the find is attributed, but these are omitted here for space considerations.

Table 1. Original data spreadsheet

Country	findloc	Desig	Se	newY	frmrnon	simpleY	simplmt	mtDNA
Austria	Tyrol-Bolzano	Oetzi	M	K1	K1	K1	G	GL91
Bulgaria	Vratitsa	V2	M				U	U2e1'2'3
Bulgaria	Krushare	K8	M				R	R
Bulgaria	Svilengrad	P192-1	M				U	U3b
Bulgaria	Stambolovo	T2G2	M				H	H1c9a
Czech Republic	Brandysek	RISE568	F				H	H
Czech Republic	Knezeves	RISE566	M	R-P310	R1b1a2a1a	R	H	H2a
Czech Republic	Brandysek	RISE569	F				H	H1af2
Czech Republic	Velke Prilepy	RISE577	F				T	T2b

Many later additions to this central dataset of human remains were possible due to discovery of a specialized OpenStreetMap. As of early 2019, OpenStreetMap (OSM) contributors had created and published a global map of early human remains studied by scientists. This map apparently incorporates several .geojson or .kml layers for different time periods and regions, displayed in a map view with clickable pop-ups for related information, and as such there is no mechanism for queries or filtered views. Each layer displayed presents not only some facts about each find in the pop-ups but citations of the papers describing the related scientific work. The inclusion of source citations in the OSM map allowed for quick location of additional material for incorporation into the geodatabase. This was especially helpful to locate

data sets from areas where climate apparently had hindered early efforts at DNA extraction, and so there was a much smaller body of published work with relevant data to locate. This resource helped to expand the geodatabase to its global goal.

Details about the format of the map and files used in the OSM project could be seen using the developer tools option in Google Chrome browser. Had there been many entries included in the OSM project not already present in the geodatabase, it would have been possible to acquire some of the .geojsons used in the OpenStreetMap project and to convert them into Esri shapefile format, rather than performing a completely manual input from the cited papers.

Once the initial data phase concluded and the geodatabase was assembled, any similar new finds could be incorporated. For any ongoing data collection, new material could be merged with the existing 'archaeological_find' feature class once it was converted into new shapefiles from tables or other common GIS files types. Using ArcGIS Pro, it is possible for field data to be amended or added to fit the schema of the geodatabase and updated with the data management tools.

The initial data acquisition phase resulted in a point shapefile called 'archaeological_find', now converted into a feature class for inclusion in the geodatabase. Figure 1 given below shows the table view in Arc Catalog immediately after its initial conversion into a feature class, prior to eliminating unneeded fields and renaming others for consistency in the database design. More details about this feature class and others will be provided in section 3.3 concerning the methods of the geodatabase.

OBJECTID	Shape	X	Y	Age	EarliestDateEstimated	LatestDateEstimated	Culture_Period	Designator	mtDNA	Ydna	Citation
1	Point Z	14.326193	50.162701	2300 - 1900 BCE	-2300	-1900	Bell Beaker	RISE577	T2b	<Null>	http://www.nature.c
2	Point Z	71.17527	57.701396	43 000 BC	-42000	-41000	no context	Ust'-Ishim	R	K(xLT)	http://www.nature.c
3	Point Z	21.829963	45.016303	40 000-35 000 BC	-40000	-35000	no context	Oase1	N	F	http://www.nature.c
4	Point Z	115.893338	39.661387	38 000 BC	-38000	-38000	Tianyuan	Tianyuan	B	<Null>	http://www.pnas.org
5	Point Z	39.033566	51.389044	35 000 BC	-35000	-35000	Aurignacian culture	Kostenki 14	U2	C-M130, C1b	http://www.sciencem
6	Point Z	40.498626	56.175389	33 000 - 30 000 BC	-33000	-30000	Upper Palaeolithic	Sunghir 1	U8c	C1a2	http://science.scienc
7	Point Z	40.498626	56.175389	33 000 - 30 000 BC	-33000	-30000	Upper Palaeolithic	Sunghir 2	U2	C1a2	http://science.scienc
8	Point Z	40.498626	56.175389	33 000 - 30 000 BC	-33000	-30000	Upper Palaeolithic	Sunghir 3	U2	C1a2	http://science.scienc
9	Point Z	40.498626	56.175389	33 000 - 30 000 BC	-33000	-30000	Upper Palaeolithic	Sunghir 4	U2	C1a2	http://science.scienc
10	Point Z	136.189377	70.746489	30100-29300 BC	-30100	-29300	Upper Paleolithic	Yana1	U2'	P1	https://www.nature.c
11	Point Z	136.189377	70.746489	30100-29300 BC	-30100	-29300	Upper Paleolithic	Yana2	U2'	P1	https://www.nature.c
12	Point Z	15.616894	48.403565	29 250-28 690 BC	-29250	-28690	Gravettian	Krems WA3	U5	<Null>	http://www.nature.c
13	Point Z	16.644931	48.888395	28 700-27 300 BC	-28700	-27300	Gravettian	Vestonice16	U5	IJK	http://www.nature.c
14	Point Z	17.577868	40.726348	25 800-25 500 BC	-25800	-25500	Gravettian	Ostuni1	M	<Null>	http://www.nature.c
15	Point Z	103.520565	52.838136	22 000 BC	-22000	-22000	Mal'ta-Buret' culture	MA-1	U*	R*	http://www.nature.c
16	Point Z	-3.46653	43.260581	16 800-16 600 BC	-16800	-16600	Magdalenian	El Miron	U5b	<Null>	http://www.nature.c
17	Point Z	92.853441	56.029903	15 000-14 500 BC	-15000	-14500	AfontovaGora3	AfontovaGora3	R1b	<Null>	http://www.nature.c
18	Point Z	-2.410142	34.809765	13000-12000 BC	-13000	-12000	Iberomaurusian	TAF009	U6a	E1b1b1a1b1	http://science.scienc
19	Point Z	-2.410142	34.809765	13000-12000 BC	-13000	-12000	Iberomaurusian	TAF010	U6a	E1b1b1a1	http://science.scienc
20	Point Z	-2.410142	34.809765	13000-12000 BC	-13000	-12000	Iberomaurusian	TAF011	U6a	E1b1b1a1	http://science.scienc
21	Point Z	-2.410142	34.809765	13000-12000 BC	-13000	-12000	Iberomaurusian	TAF012	U6a	<Null>	http://science.scienc
22	Point Z	-2.410142	34.809765	13000-12000 BC	-13000	-12000	Iberomaurusian	TAF013	U6a	E1b1b1a1	http://science.scienc
23	Point Z	-2.410142	34.809765	13000-12000 BC	-13000	-12000	Iberomaurusian	TAF014	M1b	E1b1b1a1	http://science.scienc
24	Point Z	12.318517	37.934996	12 250-11 800 BC	-12250	-11800	Late Epigravettian	Oriente C	U2'	<Null>	https://www.science

Figure 1. Table view in ArcCatalog after feature class conversion

3.1.3. Genetic Data – the Family Lines of mtDNA and Y-DNA

The genetic data in the finds database was used to create the two data tables for the mtDNA and Y-DNA genetic haplogroup. Unique instances of the 'archaeological_find' subclade, the "mtDNA" or "Ydna" fields shown above, were found using Excel to create the basis of the two tables. The most common clades also were added. The resulting columns were copied and used as the first column field of two new tables. Although the tables will expand to include many clades not yet present in the finds dataset, this step ensures the presence of the most relevant groups for analysis and queries in the finds for testing from the beginning.

After the initial list of Y-DNA subclades or clades was made, an additional field was created for the SNP nomenclature of those groups, if determined. The haplogroup to which the clades belong is another field, and the haplogroup's direct ancestor is the last field included in the table. The mtDNA table follows an identical structure, except that a separate SNP nomenclature is not used for mtDNA groups. Both tables were created as .csv in Excel, then

imported into the geodatabase. Except for the object ID, all fields are 'text' data types for this schema. The details of these fields in the design will be discussed further in section 3.3.

The consistency in mtDNA naming conventions makes working with the data much simpler, especially when designing the tables for queries. Accommodating a changing and inconsistent system of nomenclature for Y-DNA data has posed several challenges to the correct import of the data as well as construction of the tables and relationships. The solution at this point is to use the International Society of Genetic Genealogists (ISOGG) naming conventions for most purposes, but to keep an SNP field in the table for cross-referencing. They may possibly be fully incorporated later when these SNP designators are universally accepted and standardized.

3.2. Software and Tools

Software offers a variety of potential solutions for this spatial database project. Several requirements are present for a geodatabase of archaeogenetic data from around the world. A chief concern is the ability of the software to support complex analysis and to offer visualizations that can differentiate between many objects or characteristics. The data for the specimens consists of numerous material cultures identified by researchers, a wide range of dates, and hundreds of clades or subclades, among many other attributes included for visualization and analysis. The data for the specimens alone could grow to number several thousand entries, but storage needs are still minimal compared to many commercial database applications. Nevertheless, the database must handle potentially complex queries well. It is also extremely important to be able to update and edit the data easily, since new techniques are making more samples available with DNA analysis and the field of genetics is changing rapidly.

3.2.1. Prototype and Final Software Solution

The initial trial of this database was created in Microsoft SQL Server Management Studio and was a fully functional prototype. A test was also done of the database using Access. SQL Server Management Studio makes database management relatively simple and can accommodate spatial data. However, the capabilities of visualization and analysis offered by other tools are not available in SQL Server Management Studio, so it was not considered a viable option for this project. Other options, including several open source software solutions, were explored until the choice was made to construct this spatial database in a geodatabase (.gdb) format using Esri.

3.2.2. Software Solutions for Complete Spatial Database

ArcGIS Pro by Esri was selected as the software of choice after reviewing some comparisons of other database management software with spatial capabilities, such as the comparison of MySQL and PostGIS capabilities by Piórkowski (2011), as well as considering personal experiences with the performance of the various tools available for database management (DBM) and GIS. A key factor in this decision is the simplicity of a unified solution for all DBM, GIS, data sharing, and cartographic needs. As Rodriguez (2019) points outs in his work creating a geodatabase for historical migration data, using a comprehensive software program that can store, handle, and manage both spatial and nonspatial data is essential.

ESRI's GIS software and geodatabase format not only provide these requirements, but also offer the benefit of a widely used platform with rich visual displays, user-friendly cartographic layouts, and many advanced analytic tools at the user's disposal. This toolkit uses a largely intuitive GUI that allows for quick setting of parameters and can reduce typos or coding language errors. Esri has the additional advantage of providing a data portal to make ongoing data acquisition and sharing between researchers easier.

Besides Microsoft Excel for compiling notes and initial data preparation, ArcGIS Pro version 2.3 is the main tool for this project. Tables and shapefiles can be imported into the geodatabase and converted into feature classes or geodatabase tables. Relationship classes may also be created with ArcGIS Pro. If domains or field changes are needed during the project, it can be used to make them. The multifunctionality and highly satisfactory performance of both tools have fully met the software needs of the project.

3.3. Database Design

To move beyond a simple collection of shapefiles and into a geodatabase, it was necessary first to construct a relationship diagram. The design took a few different shapes before the ERD was finally constructed after some trials and feedback. Following the design with normalized tables, some initial fields from 'archaeological_find' were removed after it had been converted into a feature class in the geodatabase. The initial ERD is provided in figure 2 below. It is anticipated that additional adjustments will be made after the database is constructed and further tested.

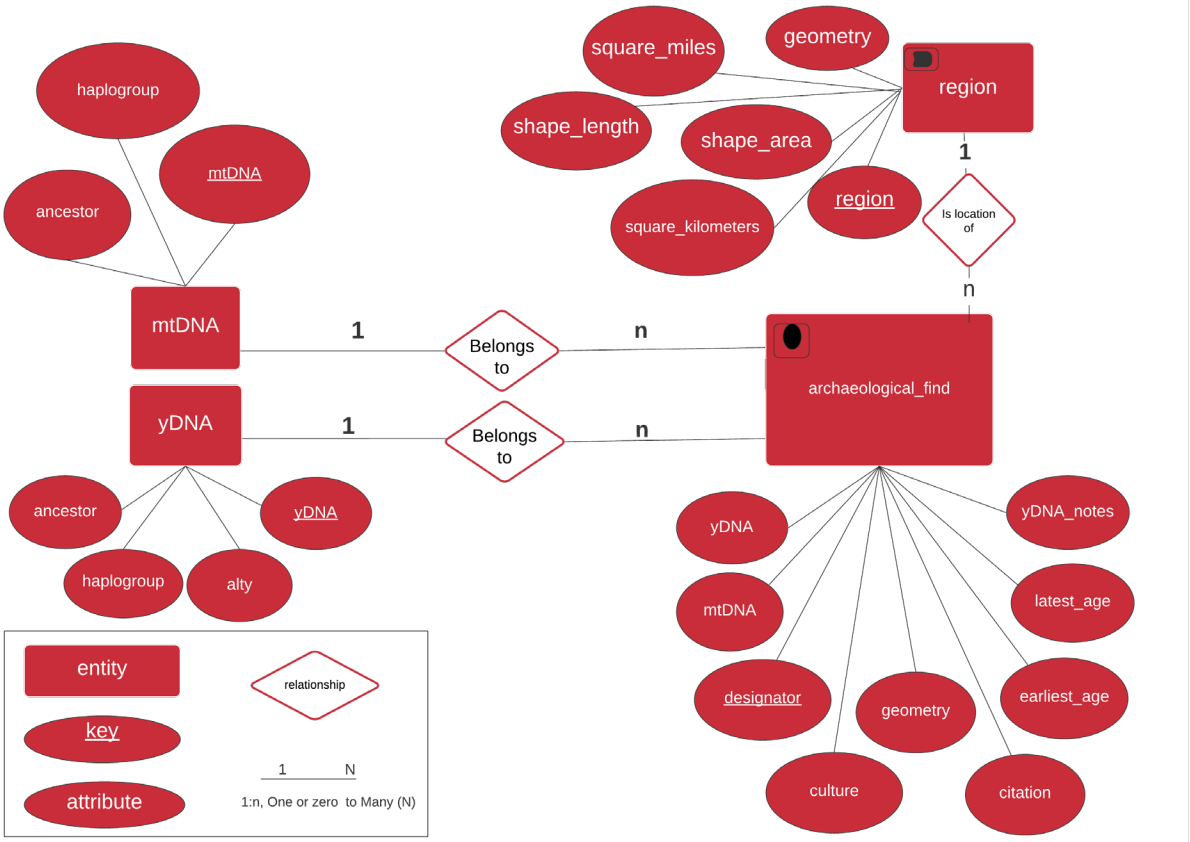

Figure 2. Conceptual ERD

3.3.1. Overview of Tables

The current design includes four parts, two feature classes containing spatial data types, 'region' and 'archaeological_find', and two genetic data tables, 'yDNA' and 'mtDNA', that contain only non-spatial data. Ideally, all table names would use either singular or plural forms and letter case to avoid needless confusion during queries or any scripting involving the tables. Field names also were somewhat constrained by Esri requirements, such as avoiding spaces. To aid clarity, the names for DNA-related tables retained the upper-case form, and these were the only exceptions to the lower case and singular forms used in the geodatabase. Again, for clarity

the name of the central feature class of samples from human remains is called ‘archaeological_find’ to provide more immediate understanding of the contents.

In the geodatabase design, the central table is the feature class for the archaeological finds themselves. It contains entries for the human specimens and their attributes, including the mtDNA and Y-DNA clade or subclade identified, and spatial data expressed as geometry. For this data set, the geometry is point data derived from medium precision and accuracy latitude and longitude coordinates, which was originally converted using XY Table to Point tool. Two other tables contain mtDNA data and Y-DNA, respectively, and are named ‘mtDNA’ and ‘yDNA’. These tables both contain fields for names of each clade or subclade, simply called either “mtDNA” or “yDNA”, the upstream “haplogroup” to which it belongs, and the haplogroup’s direct ancestral haplogroup, called “ancestor”. The ‘yDNA’ contains an additional ‘alty’ attribute of alternative nomenclature. The ‘region’ table has the “region” or name, its polygon “geometry”, and the standard Esri fields for “shape_length”, “shape_area”, “square_kilometers”, and “square_miles”.

3.3.2. Relationships and Cardinalities

The relationships for these tables are relatively simple and can be assigned using ArcGIS Pro. Each entry in the ‘archaeological_find’ table must be linked by its ‘mtDNA’ field to exactly one entry in the ‘mtDNA’ field of the ‘mtDNA’ table, since specimens without DNA analysis were not included in the project. However, each entry in the ‘mtDNA’ field of the ‘mtDNA’ table may have several ‘archaeological_find’ entries that belong to it. Therefore, the cardinality of the relationship of ‘mtDNA’ to ‘archaeological_find’ is “one-to-many”, for each ‘archaeological_find’ specimen belongs to exactly one ‘mtDNA’ entry but a single ‘mtDNA’ entry could have many specimens that belong to it. Likewise, entries in ‘archaeological_find’

may only have one single corresponding ‘yDNA’ value, but again each ‘yDNA’ table entry may also have many entries in the ‘archaeological_find’ table that belong to it. The cardinality therefore is “one to many” from the ‘yDNA’ table to the ‘archaeological_find’ table.

3.3.3. New vs. Original Design for Tables

The early prototype structure had been overly complex, so it was simplified in the new design for Esri’s geodatabase format. Keeping only a ‘region’ data set and not including a ‘country’ data set, which would not be relevant for the temporal span of the geodatabase, reduced the number of tables. There is no effort to connect DNA groups to regions where they are commonly found today. This is because the aim of the geodatabase is to allow for analysis of these relationships as they existed in the past, and so it was beyond the scope of this project. Instead, a ‘region’ feature class serves as an example of using geographic references or other spatial data in queries on the central feature class ‘archaeological_find’ and the DNA tables. The relationship of ‘region’ to ‘archaeological_find’ will utilize spatial operations rather than explicit keys in a departure from relational database design in the geodatabase design.

3.3.4. Data Types

Operations may fail if the proper data types and lengths are not specified. For example, dates in some data sets may be formatted as text, or a field that may require a decimal to the hundreds place for some records may be formatted as an integer; these would require changes to the data type to be used as intended. For this geodatabase, the estimated earliest and latest ages of the archaeological finds were included to provide further context and allow for easier queries involving multiple cultural periods or groups. However, they were left as integers, because of difficulties in working with such early and imprecise dates. Positive and negative integers were

used for CE dates and BCE dates, respectively, to not only facilitate queries on the finds from different eras but to allow the user to perform age calculations.

Aside from object IDs and the integers for estimated ages, most other data types in the schema are text, or varchar (255), except for the two feature classes that contain geometry. The 'region' entity additionally retains the four length and area related fields from Esri, which are double. The details of data types in the schema of the current geodatabase plan are illustrated by figure 3.

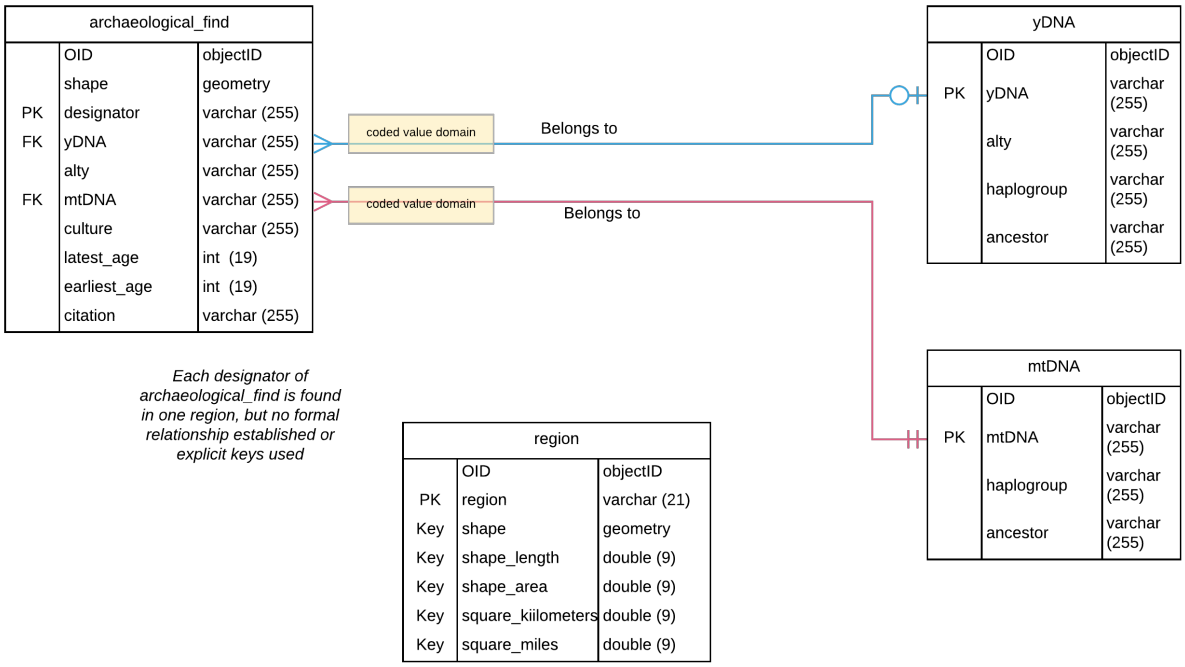

Figure 3. Schema drawing with data types of early geodatabase plan

3.3.5. Region Feature Class

The 'region' feature class was made from a data set created and provided by Esri, consisting of polygons representing common regional divisions of the world's land masses. A relationship class using key fields is not strictly necessary with ArcGIS Pro or other GIS because of spatial join capabilities. One region may be the location of many archaeological finds, and

each find will only belong to one region, but the cardinality is not shown here by the crow's foot notation for figure 3 because the tables 'region' and 'archaeological_find' were not formally linked. Although this data set was included as a permanent feature class in the geodatabase, keys were not used to create a relationship class in order to demonstrate this capability. The use case scenario includes analysis of the archaeological finds alongside historical geographic data, e.g. ancient coastlines or glacial extents. Many such data sets are available but are localized, and so were not incorporated into the geodatabase at this time. The 'region' feature class will show how this analysis of finds with respect to geographic areas can be accomplished in a GIS with a geodatabase, using spatial join, without the need for creating special key fields as would be required in a relational database.

3.3.6. Archaeological_find Feature Class

'Archaeological_find' is the central feature class of the geodatabase. As described in the data section 3.1, the 'archaeological_find' feature class was created using Arc Catalog from a point shapefile made in ArcGIS Pro. Its attribute table includes the geometry, along with other non-spatial attributes including the designator, "culture" indicating the material culture or period (e.g. Neolithic, Hittite, Magdalenian, etc.), "earliest_age" giving the specimen's earliest estimated age, "latest_age" giving its latest estimated age, mtDNA clade or "mtDNA", the Y-DNA clade "yDNA" in ISOGG form if specimen is male and this data could be obtained, alternative nomenclature "SNP" for the Y-DNA clade if available, and "citation" for a doi – useful to quickly verify the data or obtain more information concerning the specimen in the original paper. Each entry in 'archaeological_find' belongs to exactly one 'mtDNA' entry as shown in the crow's foot in figure 3, but belongs to one or zero in 'yDNA' since many samples are genetically female. The 'mtDNA' and 'yDNA' entries may both have many

‘arcaheological_find’ entries that belong to them. Details of the fields and data types as shown in ArcGIS Pro’s Catalog can be seen in Table 2 outlining the ‘archaeological_find’ feature class alone. Its geography view provided immediately following in figure 4.

Table 2. Data types for ‘archaeological_find’

Field	Data type
ObjectID	object ID
Shape	geometry
earliest_age	long
latest_age	long
culture	text
designator	text
mtDNA	text
yDNA	text
citation	text

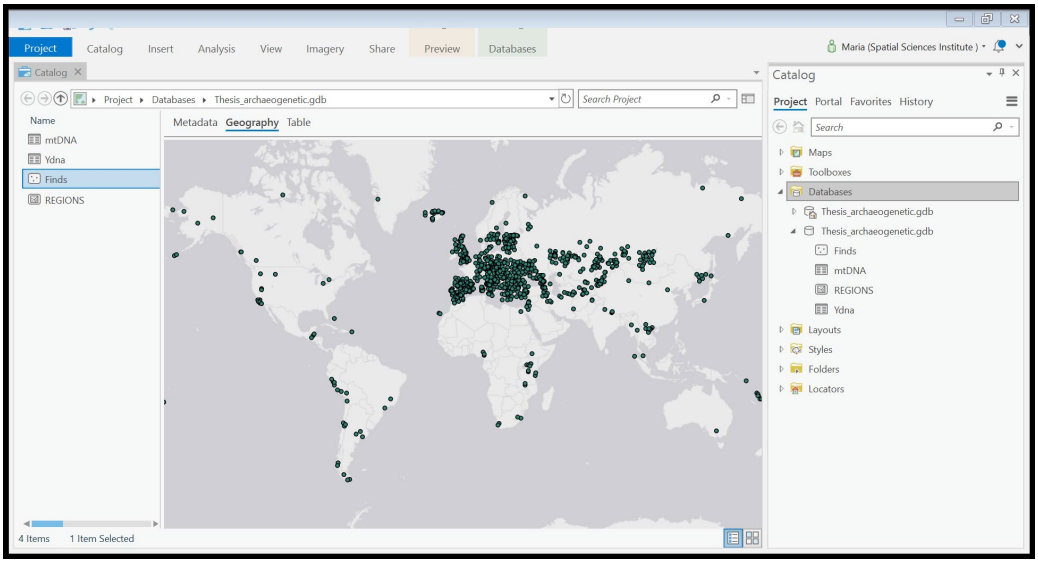

Figure 4. Geography view of feature class ‘archaeological_find’ points in Catalog

3.3.7. Genetic Data Tables

As introduced in section 3.1 on data needs and acquisition, the tables for mtDNA and Y-DNA contain fields for the DNA clades, their haplogroups, and the direct ancestor of the haplogroups. These progress from left to right from the younger, downstream subclade or clade to the haplogroup to which it belongs. In turn, the haplogroups' direct ancestor, or further upstream haplogroup, is listed. This additional field is included to assist in querying for related groups and keep the database design as simple as possible. No regional associations for the subclades or haplogroups of either mtDNA or Y-DNA are assumed, although there are frequent references to them in the literature. The primary keys for the 'mtDNA' and 'yDNA' tables remain the "mtDNA" and "yDNA" fields to connect the 'archaeological_find' table through the "mtDNA" and "yDNA" fields as foreign keys. If junction tables do become necessary for any future changes to the geodatabase and its uses, the Object ID fields could be used to set up primary and foreign keys with simpler numerical IDs rather than text. Data types for the two genetic data tables are all text, except for the Object IDs. An example from mtDNA showing several entries is given below in table 3.

Table 3. Sample of mtDNA genetic clades and upstream haplogroups

OBJECTID	haplogroup	mtDNA	ancestor
1	A	A	N
2	A	A1	N
3	A	A11	N
4	A	A14	N
5	A	A15	N
6	A	A16	N
7	A	A17	N
8	A	A1a	N
9	A	A2	N
10	A	A2a	N
11	A	A2b	N
12	A	A2c	N
13	A	A2d	N
14	A	A2h	N
15	A	A2i	N
16	A	A2p	N
17	A	A4f	N

3.4. Testing the Adequacy of the Data and Functionality of the Database

The fourth phase of the methods process will evaluate the data and test the database's performance. The tests of the data will determine if it is possible to complete some basic analyses with the sample size and the attributes included. These queries will test the relationship classes set up in the geodatabase, ability to select multiple attributes and limit to only the desired results, and the spatial join in place of a standard relationship between geographic data sets and tables or point data feature classes. These will also illustrate some benefits of working within a GIS that can render maps with flexible options. The details of these short tests are as follows:

- A spatial join will be run on the 'region' and 'archaeological_find' data sets, then a SQL query will locate the regions where Q haplogroup-descended specimens are found in the data.

- The second example will test the relationship of ‘mtDNA’ and ‘archaeological_find’ and explore the ability to show the temporal-spatial distribution of an early mtDNA haplogroup. First, using a query the database will find all members of clades or sub-clades belonging to haplogroup M. Then using a classed color ramp in ArcGIS Pro’s symbology options, the various age ranges of those related ‘archaeological_find’ samples will be mapped in ArcGIS Pro across the ‘region’ data with an underlying basemap for context.
- The geodatabase will be queried again for the ‘yDNA’ relationship and for multiple attributes. The example query will find results restricted using the “earliest_date” field for two branches of a larger haplogroup. For this test, the “yDNA” belonging to R1a and R1b will be located in the ‘yDNA’ table, and related points from archaeological_find will be queried in turn to find those older than 1 CE. The results will be mapped and visualized to show the areas with concentrations of the two branches.

In addition, several other SQL queries to check the geodatabase’s functionality and success of the basic design will be run. The results will be presented in simple tables.

Chapter 4 Results

This chapter will describe the final design of the geodatabase and the reasoning behind major changes or modifications in section 4.1. A benefit of the geodatabase for spatial data over standard relational database structure is briefly described in section 4.2. This chapter will show the results of queries that demonstrate the functionality of the geodatabase, as implemented using Esri ArcGIS Pro, in section 4.3.

4.1. Final Database Design

The final physical structure of the geodatabase changed only slightly from the original concept, due to some differences in Esri's file structure compared to requirements of the standard normalized database structure. Some fields were removed from 'region' feature class obtained from Esri and one added into the 'archaeological_find' table. Another difference from the original design was that domains for better data integrity were introduced for the 'yDNA', 'mtDNA', and 'archaeological_find' tables. The use of spatial capabilities to avoid explicit keys linking the 'archaeological_find' to 'region' feature classes was successful and maintained.

4.1.1. ERD Changes

Overall, the differences in the structure of the conceptual ERD described in Methods Chapter 3 versus the end result were minor. The final form of the ER diagram is shown below in Figure 5. Due to the changing nomenclature for Y-DNA, a "yDNA_notes" field was created to document regarding data for 'archaeological_find' in case expansions or changes were needed in future. However, unneeded fields from the 'region' feature class included by Esri that could be derived from its "geometry" were eliminated, namely "shape_length", "shape_area", "square_miles" and "square_kilometers".

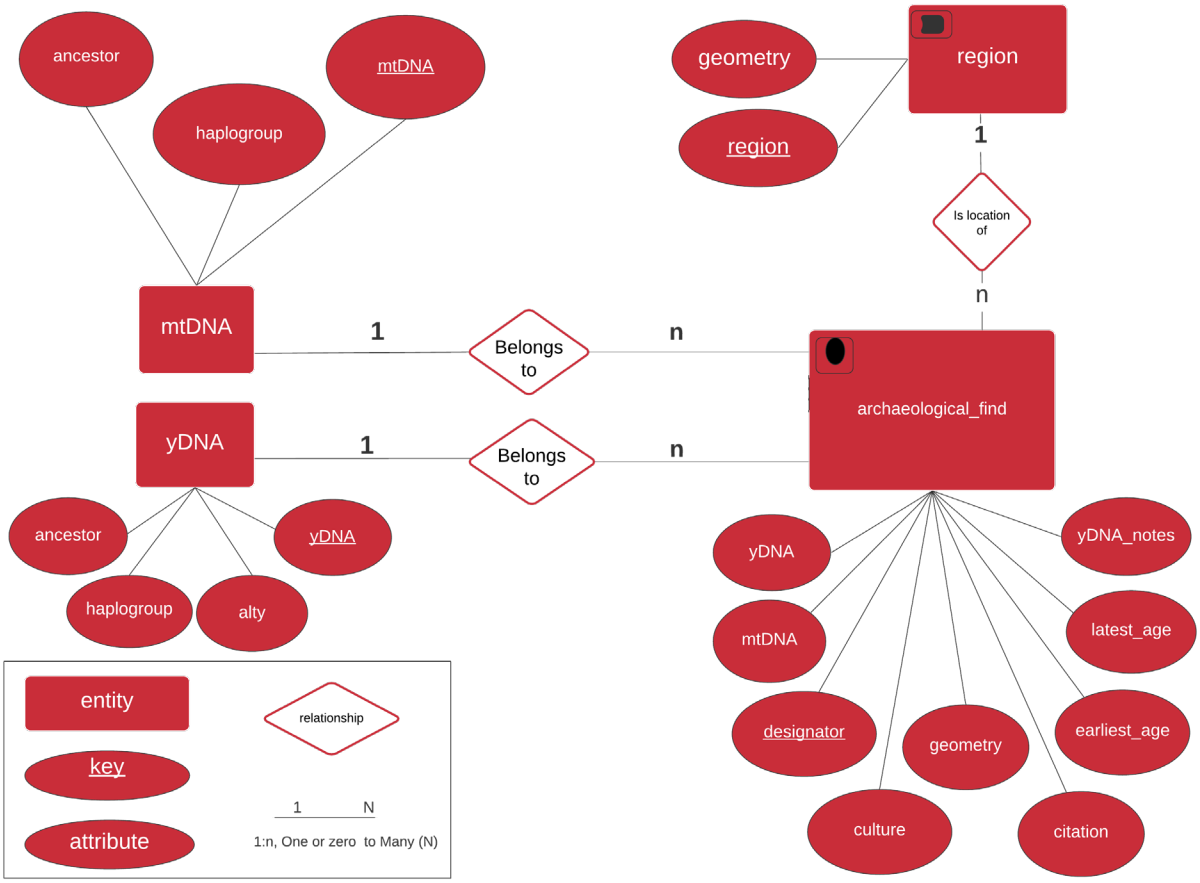

Figure 5. Conceptual ERD as implemented

4.2. Moving Away from a Standard Relational Database

Using a GIS for spatial database creation and management as well as analysis allows us to make use of processes such as spatial joins. The spatial join locates entries that fall within specified regions or boundaries, creating a layer with these relationships. This layer then can be further queried. Any other geographic tables potentially added in future to the geodatabase would not require a relationship set up using keys, as is necessary in a standard relational database if the database creator takes advantage of this capability. The results can be saved and added as a table or feature class within the geodatabase, but the spatial join itself does not remain a permanent

part of it. For this reason, the ‘region’ table represented in the drawing of the ERD provided above in Figure 5 reflects the relationship of ‘archaeological_find’ to ‘region’. The inclusion of the ‘region’ table within the database, but without the need for directly relating the tables using a key field, is described further in 4.2.1 and illustrated in a physical diagram of the final geodatabase form in Figure 6.

4.2.1. Final Schema as Implemented in ArcGIS Pro

The final physical form of the tables in the geodatabase was consistent with the earlier plan discussed in Methods but eliminated the extraneous fields for shape length, area, and measurements of square kilometers and miles. These were not used in the geodatabase and since they could be derived from the geometry of the feature class, they were unnecessary and redundant. Only “region” and “geometry” fields were retained in the final implementation of the design for ‘region’. As planned, for ‘archaeological_find’ and the DNA tables, Catalog established relationship classes with cardinalities and keys, and spatial operations were used instead to link ‘region’ and ‘archaeological_find’. Data types for all the remaining fields in all tables were maintained as in the schema described previously in 3.3.4. Details of these data types and the cardinalities of relationships are illustrated in the crow’s foot notation diagram for the final version of the physical database given in Figure 6 below.

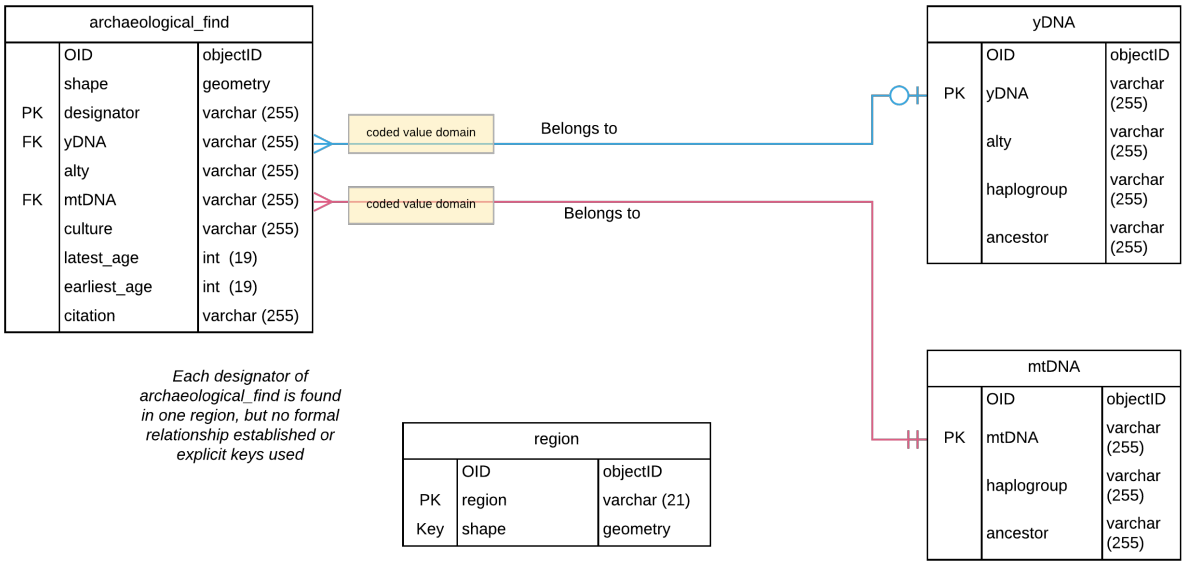

Figure 6. Schema drawing with data types for geodatabase

4.2.2. Aspects of Relational Database Structure

ArcGIS Pro was able to establish the relationships shown in Figure 7. Although the spatial tools made some aspects of relational database unnecessary, previously discussed in 4.1.1., Esri's geodatabase format should still have a structure to allow successful SQL retrievals. The "Clause" mode for queries presents a very intuitive interface for users unfamiliar with formal query language through guided selections of parameters. However, an example of the SQL involved in the actual operation can be seen by switching away from the default "Clause" to the "SQL" mode in a Definition Query. When performing a query on entries in 'archaeological_find' it is possible to retrieve not only the DNA clades, either 'mtDNA' or 'yDNA', but also the related haplogroups and their direct ancestors from the two DNA tables. A geodatabase provides the structure to use an established relationship classes to facilitate queries on normalized tables. Examples of this functionality are shown later in section 4.3.

In ArcMap and ArcGIS Pro, “Create Relationship Class” is one of Esri’s Data Management Tools. The tool asks for the origin and destination tables, the origin primary and foreign keys, whether the relationship is simple or composite, and the cardinality. After it runs, a relationship class will appear in the geodatabase itself. An example immediately follows in Figure 7.

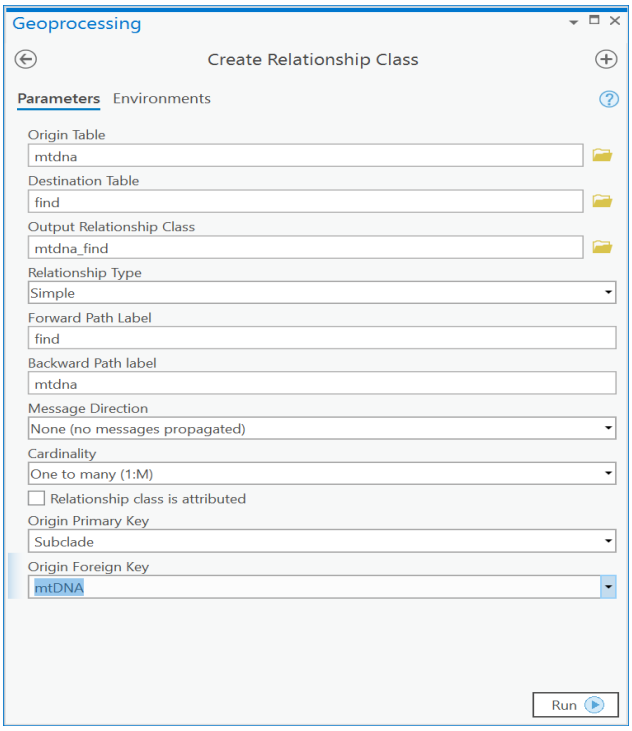

Figure 7. Create Relationship Tool and parameters

The successful creation of a relationship and its properties can be verified in Catalog as shown in the following two figures, Figure 8, and Figure 9.

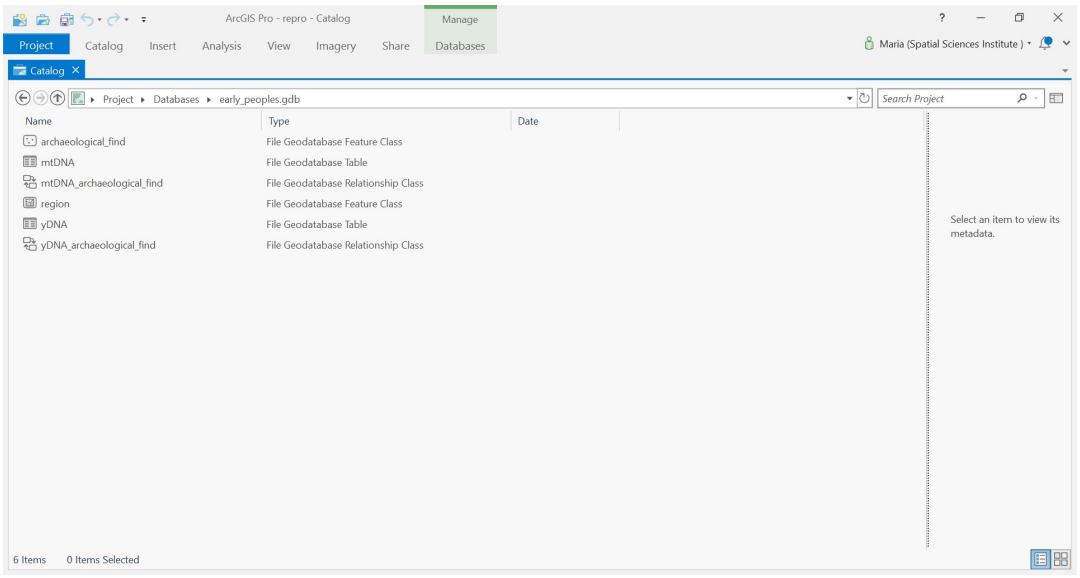

Figure 8. Relationship Classes as viewed in Catalog

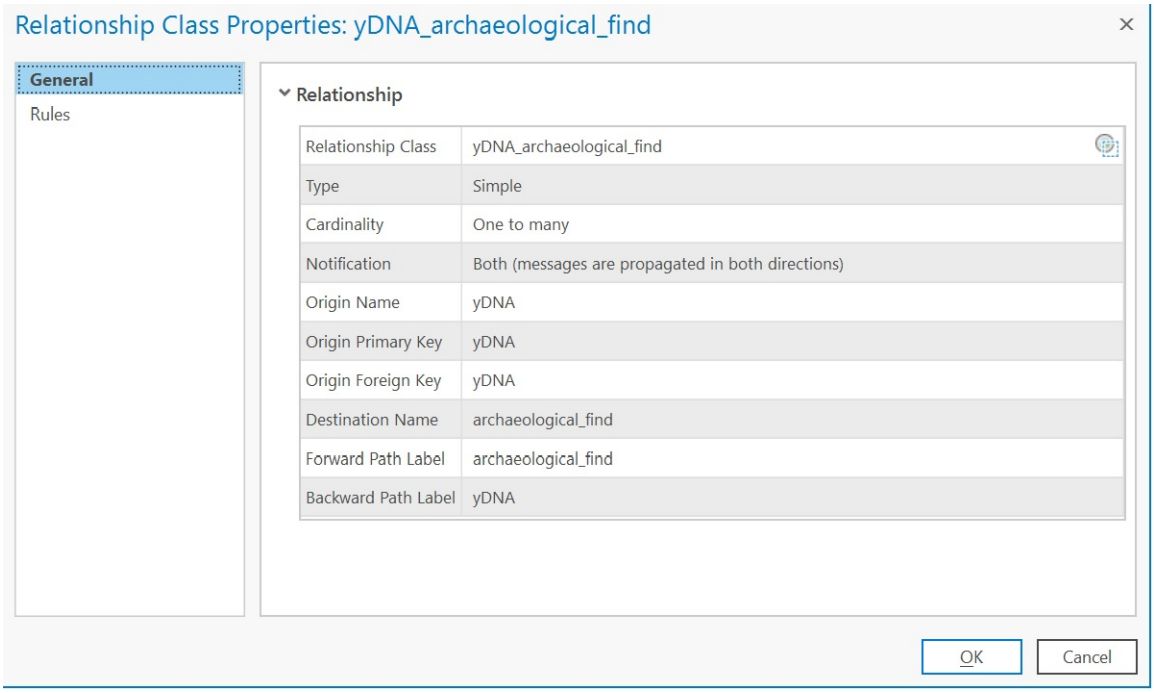

Figure 9. Relationship Class properties as seen in Catalog

4.3. Resulting Queries and Visualizations

4.3.1. Query with Spatial Join Test – DNA Tables and Region

The first result in testing the functionality of the geodatabase is the use of the spatial join to query without a formal relationship of a geographic table to other tables in the database. The example query uses a spatial join of the 'archaeological_find' feature class to the geographic reference feature class 'region'. The results of this spatial join were then queried in ArcGIS Pro using "Definition Query" to locate the regions where subclades belonged to the Q haplogroup.

In the early ages included in the geodatabase, Q-descended subclades were found in East, Central, and West Asia, all regions of Russia, East Europe, and both North and South America. Today, descendants of pre-Colombian populations in the Americas almost entirely – as much as 90% - belong to the larger Q haplogroup (Grugni et al. 2019) with the remainder of the modern distribution of the haplogroup principally in northeast Asia. This example of a spatial join query displays the ancient distribution of Q haplogroup Y-DNA entries in nine regions, Northern America, Central America, South America, Eastern Europe, European Russia, Central Asia, Eastern Asia, and Southern Asia. This is shown by the solid-colored regions below in Figure 10.

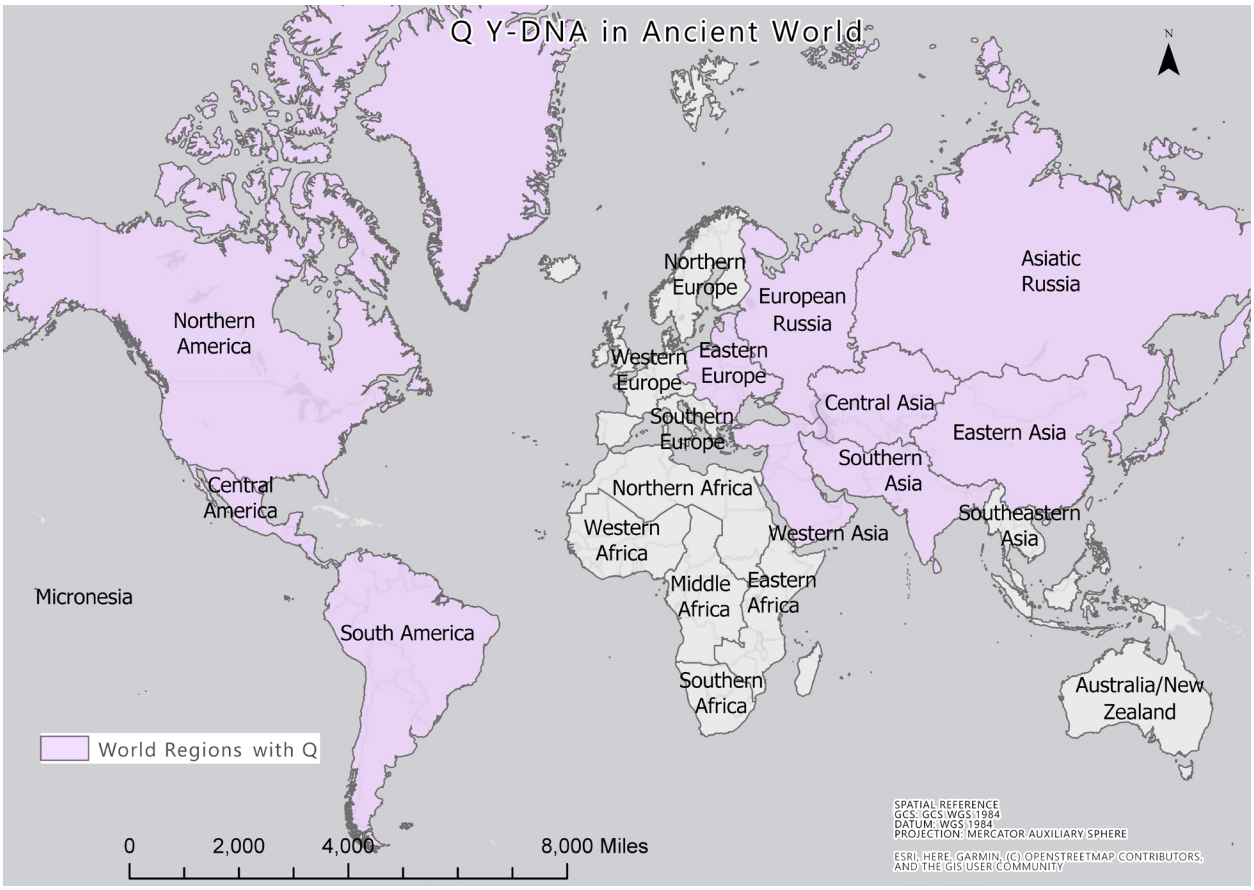

Figure 10. Spatial join query, Q Y-DNA haplogroup

4.3.2. Relationship Class Test – Mapping Related Data

The second test of the geodatabase verifies that the relationship class structure allows retrieval of related data from a table for queries on the 'archaeological_find'. The relationship class between 'mtDNA' table to 'archaeological_find' table is used to locate descendants of the mtDNA haplogroup M, the matrilineal or "motherline" marker, from 'archaeological_find' and display their distribution. The tables were normalized, so 'archaeological_find' contains only the mtDNA clade codes in the field "mtDNA", which serves as a key to the other data in the 'mtDNA' table. The 'mtDNA' table first was restricted by a definition query to the M haplogroup, then related records in the 'archaeological_find' feature class were selected. After

this step, the multiple entries in ‘archaeological_find’ highlighted only belonged to clades belonging to haplogroup. Furthermore, the query selected all clades for the M haplogroup, demonstrating that using the relationship class between ‘mtDNA’ table to ‘archaeological_find’ was successful. For clearer visualization, those related records were saved as a separate layer, shared below in Figure 11.

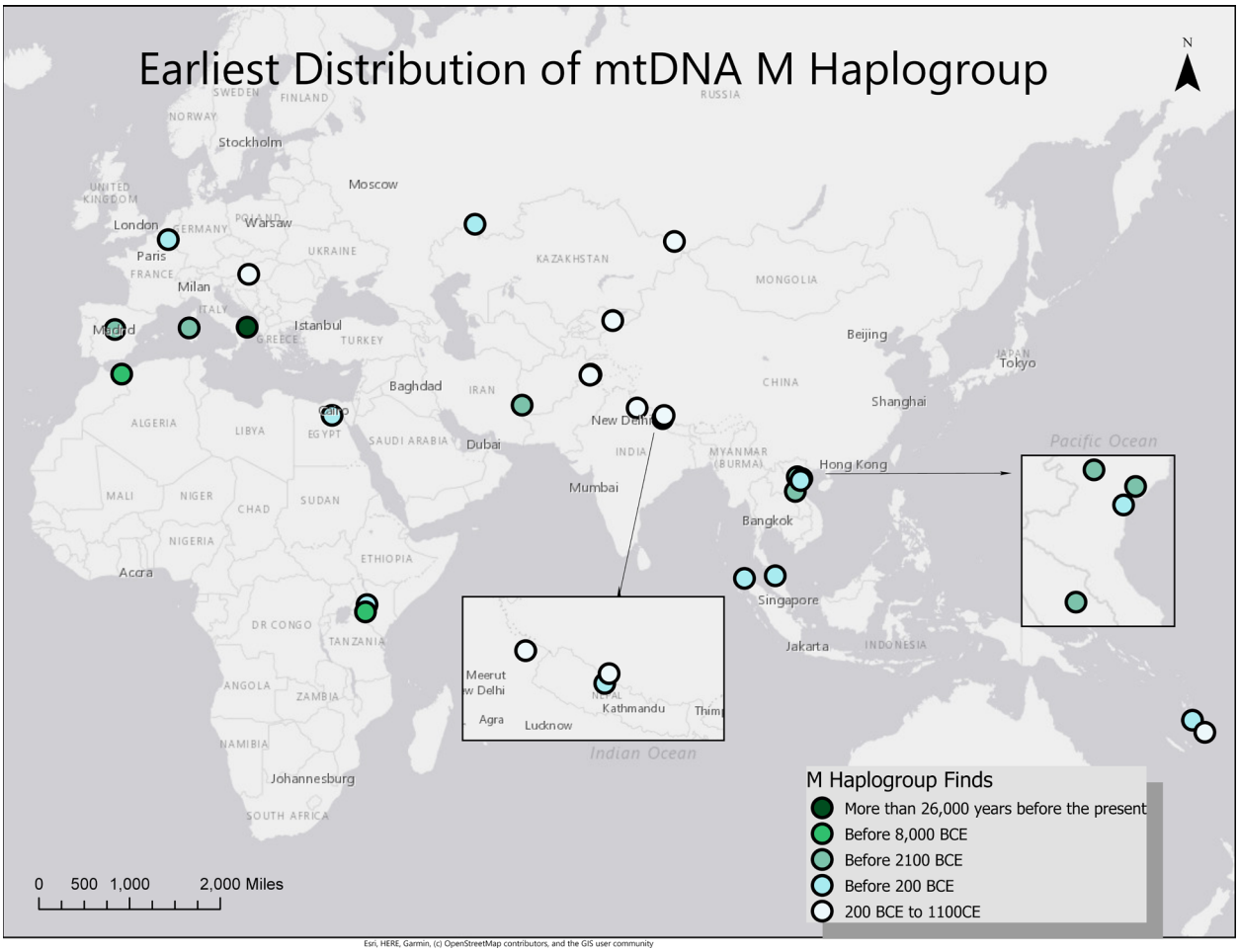

Figure 11. Query of finds and related mtDNA table, M mtDNA haplogroup

4.3.3. Test of Querying against Multiple Fields – Y-DNA Examples and Date Fields

The geodatabase allows for queries of multiple fields to find results that meet a limited set of criteria. The geodatabase was queried to find only members of clades of the R1a and R1b Y-DNA haplogroups, the patrilineal or "fatherline" markers, for a map showing where the respective haplogroup branches were concentrated. Because R1a and R1b distributions today differ in frequency between eastern and western parts of Eurasia (Underhill et al. 2015), the example also checks whether the geodatabase structure can accommodate queries to investigate and compare distributions of the haplogroups in earlier millennia.

For each haplogroup, the 'yDNA' table first was queried to find either R1a or R1b. Second, the related records in 'archaeological finds' were selected. The query was then further restricted to only return 'archaeological_find' entries dated prior to 1 CE, using where 'earliest_age' is less than or equal to 1 CE, or "earliest_age <= **1**" as displayed in Esri's SQL box. This example also served as a test of the relationship class on the 'yDNA' table to 'archaeological_find'. Each query result was exported as a separate file.

The R1a and R1b results are shown in Figure 12 following this paragraph. R1a and R1b samples are visualized with blue and pink, respectively, with darker areas appearing where one of the two branches were more concentrated, and purplish areas appearing where there likely were concentrations of both branches. These maps illustrate that R1a and R1b did show different distributions from one another in the past, similar to what is seen today.

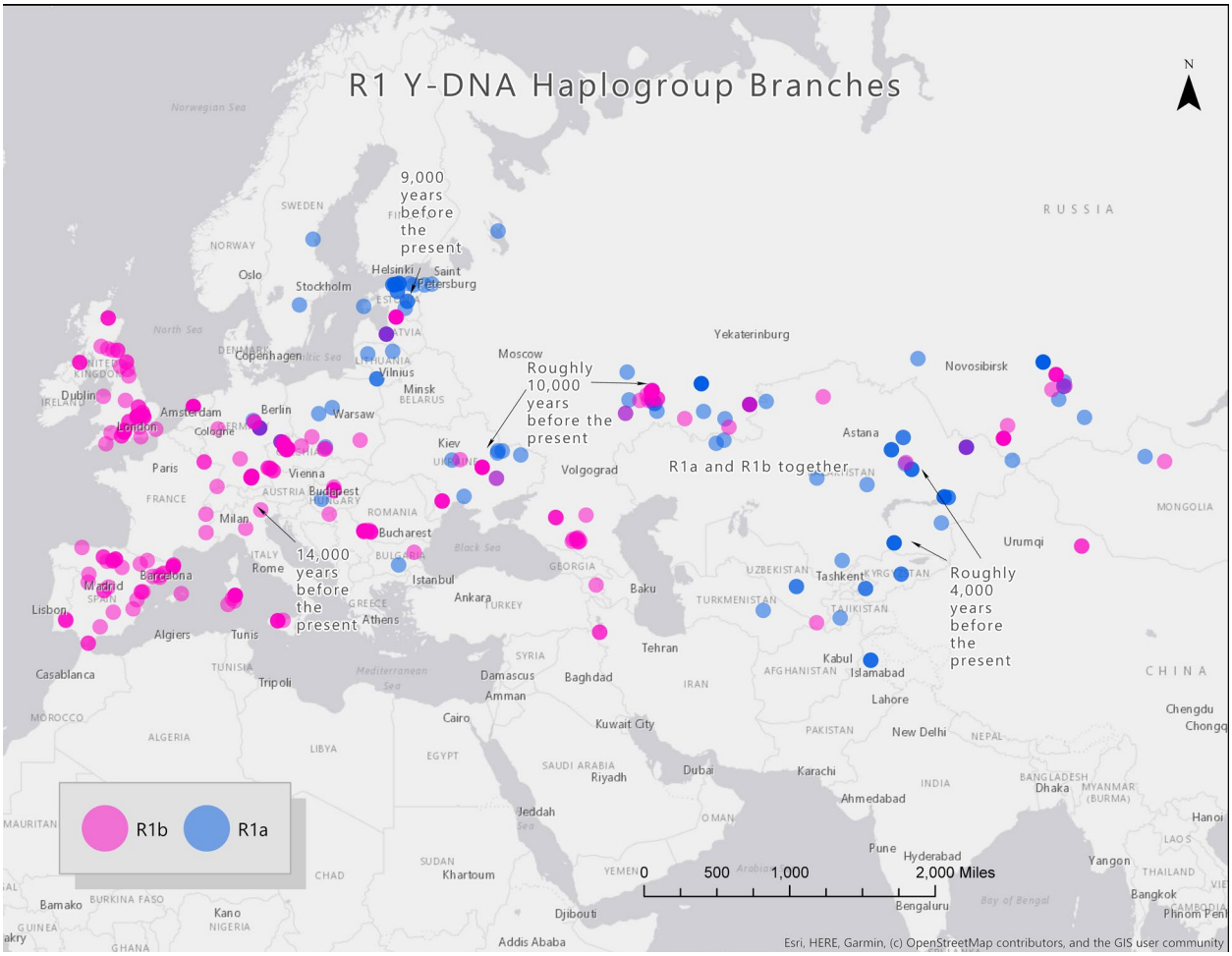

Figure 12. R1a and R1b distributions before 1 CE

R1a in ancient times had a stronger presence in eastern Europe and central Asia, to the northeast of the Altay mountain range and south to parts of present-day Mongolia, than R1b. R1a also appears to possibly have been notable in the Baltic areas and parts of Scandinavia prior to 1 CE. Today, R1a groups mostly appear in the eastern regions of Europe and still are noted in Central and South Asia (Underhill et al. 2015).

The most westerly areas of Europe today have a high percentage of R1b-related Y-DNA groups (Underhill et al). R1b query results show that in ancient times R1b was already

predominant over R1a in westernmost areas of Europe. Although R1 groups are not the predominant ones seen in Scandinavia today, but those commonly noted in the region now are usually of the R1b branch versus R1a. It appears this may not have been the case in ancient times, as the Baltic and parts of Scandinavia show many samples for R1a but not R1b. In ancient times R1b also was seen, alongside R1a, in the steppe areas of eastern Europe, and as far east as the environs of the present-day city of Urumqi.

4.3.4. Make Query Table – an Alternative to Standard Joins and Definition Queries

A more general query tested locating the oldest finds by using "Make Query Table" tool. This is an alternative to creating joins or definition queries in the map view that can be used straight from the analysis toolbox. A map layer can be created rather than solely a data table of results if one of the tables includes a geometry field, or "Shape" as named in Esri. Selecting fields is not necessary in the tool except when using Model Builder; the selections below serve as illustration only. A screenshot of the "Make Query Table" interface follows immediately in figure 14. This example shows that by using SQL "archaeological_find.earliest_age <= -8000", the GIS could retrieve all the finds at least 10,000 years old. A sample of the results table from the query is provided in table 4, showing some of the oldest archaeological_finds and attributes, immediately following figure 13. Table 4 has omitted OID and geometry fields for space considerations; 'yDNA' was returned successfully in the query but the top ten oldest selection of results shared here are all from females.

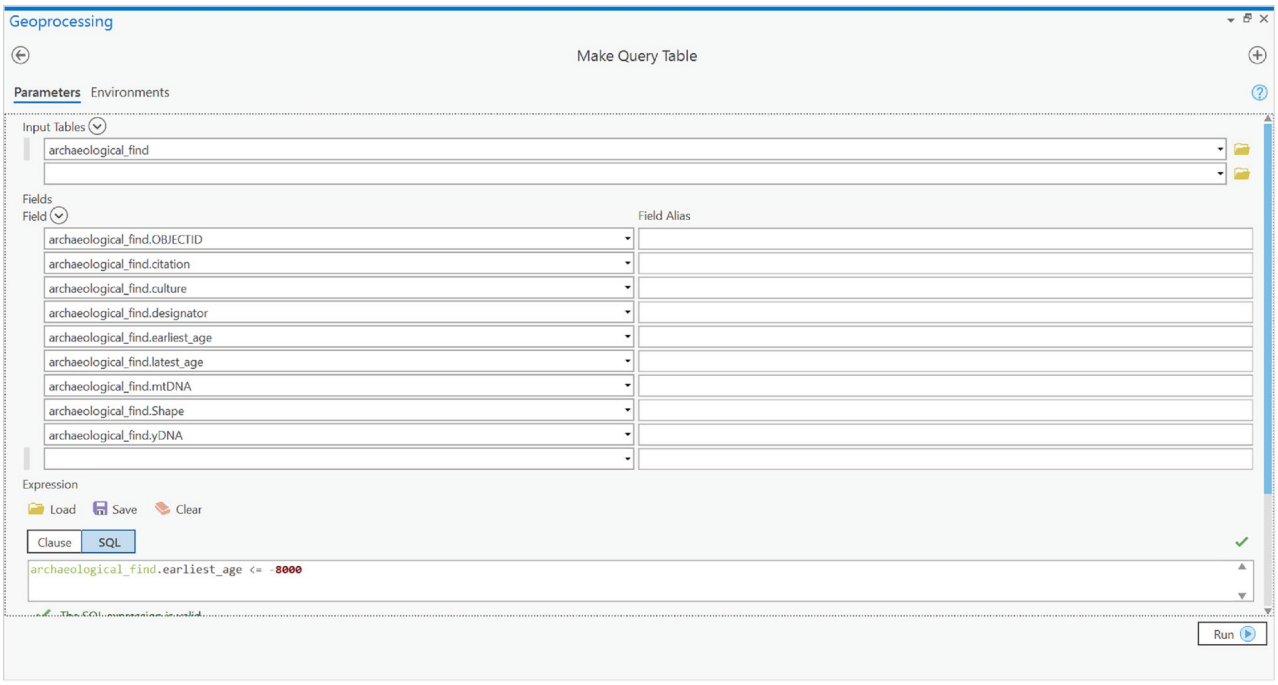

Figure 13. Make Query Table Processing Parameters interface

Table 4. Sample of Make Query Table results for samples older than 10 kya

earliest_age	designator	culture	mtdna	citation
-38000	Tianyuan	Tianyuan	B	http://www.pnas.org/content/110/6/222
-29250	Krems WA3	Gravettian	U5	http://www.nature.com/nature/journal/v
-25800	Ostuni1	Gravettian	M	http://www.nature.com/nature/journal/v
-16800	El Miron	Magdalenian	U5b	http://www.nature.com/nature/journal/v
-15000	AfontovaGora3	AfontovaGora3	R1b	http://www.nature.com/nature/journal/v
-13000	TAF012	Iberomaurusian	U6a	http://science.sciencemag.org/content/e
-12250	Oriente C	Late Epigravettian	U2'	https://www.sciencedirect.com/science/
-10000	CB13	CB13	K1a	http://mbe.oxfordjournals.org/content/e
-9650	USR1	Denali complex	C1b	https://www.nature.com/articles/nature

Several other queries were carried out using the “Make Query Table” to check the structure of the ‘yDNA’, ‘mtDNA’, and ‘archaeological_find’ tables. This tool does not appear to allow for adjustments in the SQL mode to switch from inner and outer or left and right

queries. Instead, tables should be selected in the appropriate order when filling out the parameters, joining can be accomplished by matching the corresponding fields. For example, the SQL

```
archaeological_find.yDNA = yDNA.yDNA And (yDNA.ancestor = 'BT' Or
yDNA.ancestor = 'A00' Or yDNA.ancestor = 'B' Or yDNA.ancestor = 'A1' OR
yDNA.ancestor = 'A1b' OR yDNA.ancestor = 'CT')
```

yielded results for the male finds from all cultures and eras descended from these earliest Y-DNA haplogroups in the geodatabase. A check of the 'archaeological_find' table "Y-DNA" field confirmed that these results included the clades or other Y-DNA groups that descended from these groups present in the geodatabase. The table created by the query, with "citation" field omitted for space, is given here in table 5.

Table 5. Make Query Table results from earliest Y-DNA haplogroups

OBJEC	culture	designator	yDNA	ances	haplo	alty	mtDN	earlie	latest
1511	Hunyadihalom	I2783	CT	BT	BT		T2b	-4228	-3963
2043	Straubing(Early Bronze Age	WEHR_1474	CT	BT	BT		V	-2029	-1772
2044	Straubing(Early Bronze Age	WEHR_1564	CT	BT	BT		V	-2029	-1772
1279	Pre-Pottery Neolithic	ZMOJ=BON014	C	CT	CT		K1a	-8300	-7800
1357	Neolithic	I0706	C	CT	CT		K1a	-6300	-5900
1376	Hoabinhian	La368	C	CT	CT		M5	-6000	-5850
1379	Starcevo	I3498	C	CT	CT		U8b	-5837	-5659
1407	Early Neolithic Kitoi	DA357	C	CT	CT		A+1	-5500	-5000
1466	Lengyel	I1899	C	CT	CT		T2b	-4800	-4500
2631	Medieval	R1285	C	CT	CT		T2c	771	1490
2680	Medieval Nomad	DA106	C	CT	CT		C4b	1000	1250
2702	Kipchak	DA23	C	CT	CT		F1b	1045	1095
1289	Nachikufu(?)	I2966	BT	A1b	A1b		L0k	-8000	-3000
2041	Straubing(Early Bronze Age	WEHR_1414	BT	A1b	A1b		K1a	-2029	-1772
2247	Philistine	ASH008.A0101	BT	A1b	A1b		H2c	-1000	-900
2215	late Stone to Metal Age	4/A	B2b	B	B		L1c	-1260	-1020
1365	early Stone to Metal Age	2/SE I	B	BT	BT		L0a	-6020	-5740
2342	Pastoral Neolithic	I8804	A1b1b2	A1	A1	A-L427	L4b	-751	-411
2416	Ballito Bay B	Ballito Bay B	A1b1b2	A1	A1	A-L427	L0d	-199	18
2433	Ballito Bay	Ballito Bay A	A1b1b2	A1	A1	A-L427	L0d	-36	119
2341	Pastoral Neolithic	I8758	A1b1b2a	A1	A1	A1b(xA1b	L0a	-751	-415
2429	Later Stone Age	I9133	A1b1b2a	A1	A1	A1b(xA1b	L0d	-50	200
2505	Hunter-gatherer	I9028	A1b1b2a	A1	A1	A1b(xA1b	L0d	300	0
2736	Pastoral Neolithic	I89194	A1b1b2b	A1	A1	A1b1b2b;	L4a	-8000	10000
2676	Medieval	urm035	BCDEF	BT	BCDEF		H2a	900	1200

Another test query involved finding all finds prior to 1 CE belonging to the V haplogroup:

```
archaeological_find.mtDNA = mtDNA.mtDNA And mtDNA.haplogroup = 'V' And
archaeological_find.latest_age <= 0
```

This SQL returned the pre-Common Era finds in 'archaeological_find' with all the clades belonging to the upstream haplogroup V, shown immediately below in Table 6.

Table 6. Make Query Table restricted date results for mtDNA haplogroup V

designator	culture	mtDNA	yDNA	haplogro	ancesto
I2012	Roessen	V1a		V	HV
N27	Lengyel(BKG)	V14		V	HV
N19	Funnel Beaker (TRB	V14		V	HV
I10564	Afanasievo	V1a		V	HV
I7290	Bell Beaker	V3		V	HV
I5367	Bell Beaker	V10		V	HV
I2365	Bell Beaker	V3	R1b1a1a2a1a2b1	V	HV
I7638	Early Bronze Age	V10	I2a2a	V	HV
OTTM_156	Middle Bronze Age	V1b	K2b2a2	V	HV

The last two queries for V haplogroup and early Y-DNA ancestors A00, A1, A1b, B, BT, and CT also yielded results with geometry that could be seen in the map view. This part of the tool worked well as described in Esri literature but did not display the types of joins it was performing. As stated earlier, there does not appear to be a way to use the SQL window to change to OUTER RIGHT JOIN or make similar specifications on the kind of join. This may be preferred by many users, but others may be disappointed by the limit. (The tool was not tested using the Python scripting environment; it is possible that different kinds of joins are allowed there.)

Esri also indicates that the table created by the tool should be exported to retain it permanently, but this generated an empty table on two test queries. In these instances, a

workaround was achieved by selecting the original table results. Then, one may create a layer from the selection, and finally using data output save either a shapefile or table.

The query results that could be exported successfully did not seem to differ in any meaningful way from the ones that failed. It is unclear why this happened, but ArcGIS Pro occasionally failed to initially load the entire set of rows, and the attribute table would have to be reopened. It is likely that use of a OneDrive directory to output analysis results had a role in the inconsistent behavior. Working locally when possible, then uploading results to cloud-based storage for back ups or sharing is advised.

Chapter 5 Conclusions and Discussion

This chapter discusses the results of the geodatabase project, lessons learned, and ideas for future related work. Section 5.1 provides an overview of the successes and shortcomings of the geodatabase project. Reasons for insufficient data in some areas of the world are discussed in depth in 5.2, along with the potential for these issues to be addressed. Parts of the geodatabase project design and process that worked well or potentially could be improved will be addressed as a "lessons learned" in section 5.3, including the use of domains in some fields to limit error. These possibilities for future changes, especially concerning open source alternatives to confront budget and/or license constraints, will be discussed in the final section 5.4.

5.1. Overview

The design and implementation of the geodatabase in Esri's ArcGIS Pro was a success, and multiple queries and visualizations could be easily made with the geodatabase. The relationship classes made it possible to find related information and used normalized tables in the GIS. The initial creation of tables and feature classes in the geodatabase went smoothly. Domains with coded values were not difficult to implement. ArcGIS Pro did sometimes crash in very simple processes of editing when using domains but did not usually provide an error to aid troubleshooting. However, no work was lost, and the domains and other aspects of geodatabase structure were intact when the GIS was restarted.

A key issue that arose early limited the utility of the geodatabase in global analysis of some of the key questions in studying early human evolution and migrations; namely, unequal sample sizes for different continents. This was one of two external issues that presented challenges for this project and were distinct from design and technical questions. The other issue

was the evolving system of nomenclature for Y-DNA clades and haplogroups, and a working solution was quickly found. These external issues are expected to partly resolve themselves and become far less of a challenge with time, although there are multiple reasons for unequal sampling and the issue is complex.

The growth of the field of genetic genealogy has created a happy dilemma in which progress is happening so quickly that a consensus has not been achieved in some details such as nomenclature or naming conventions. The mtDNA naming conventions appear to be more stable, but in Y-DNA at least three methods have been noted to name the mutations that mark a departure of a clade or subclade from its parent branch. Two are commonly in use, and it remains to be seen which nomenclature eventually may come to predominate completely.

Until that time, the geodatabase makes use of the ISOGG more commonly seen in the papers used in data collection. The problem of Y-DNA nomenclature being in flux was overcome by using this method as the primary and using resources to "translate" the alternative nomenclature sometimes used in papers into this primary. As described in chapters three and four on methods and results, an "alty" field in the 'yDNA' table records common alternative names for ISOGG clade or subclade names. A field was eventually added to the 'archaeological_find' table called "ydna_notes" to record any additional details that potentially could be used to update the table if advances or changes in the genetic research call for a revision of the geodatabase tables to allow for better queries concerning Y-DNA clades.

Future work for this project involves a continuation of data collection and expansion of the geodatabase. Because of the likelihood that some potential users would be excluded because of the licensing costs required to create and maintain an Esri enterprise geodatabase needed to

allow for certain sharing and access controls, open source alternatives should be explored. This will be discussed later in the chapter.

5.2. Data Gaps

The greater availability of samples from certain regions is partly because of their climate or environment, such as being found in bogs, and consequently harboring better suitability for DNA analysis. There is now a greater amount of work being carried out in regions of Asia (Fu et al. 2014; 2013), Siberia, and present-day China, for example, than what might be inferred from the papers available for incorporation into the geodatabase. Most papers are published in English to share internationally, but by searching academic work published in French, Mandarin, or Russian, it may be possible to increase the amount of usable data. Some of the difference in availability of samples from certain regions also results from cultural and legal circumstances. Advances in technology may close the gap due to challenges processing samples from some environments, such as the growing availability of DNA extracted from archaeological finds near the tropics (Slatkin and Racimo 2016). However, the cultural and legal reasons that impede, or in some cases completely prohibit, work on human remains are unlikely to change soon.

5.2.1. Acknowledging the Data Gap

Although samples from African and Pacific regions with the necessary DNA analysis were found throughout the course of the project, the data remained heavily skewed toward northern regions of Asia and Europe. North American sample sizes were also not as abundant. Africa is crucial for understanding the earliest human ancestors and their migrations (Pääbo 2014), yet many samples with published analysis are relatively recent compared to the contents of the geodatabase. Without an acceptable sample size that includes a full spatio-temporal range,

it is not possible to state with confidence that fossil DNA either confirms or contradicts the work of genetic analysis in resolving many questions of human migration and evolution.

As an example of the data gap problem, one of the sample queries illustrated in figure 13, on page 57, shows the maternal DNA or mtDNA clades belonging to haplogroup M. The earliest appearing clades on the map are found in two regions of Africa and southeastern Europe, close to Asia Minor. This suggests a picture somewhat at odds with other research on haplogroup M's early branching off and its earliest dispersal (Metspalu 2004). This research explores today's Asian distribution of the haplogroup and posits that M haplogroup may have arisen in southwestern Asia, although Metspalu (2004) admits possibility of an origin in east Africa. The analysis based on this geodatabase appears to support an east African origin for this haplogroup, but this is weakened as it must be acknowledged that better representation in the data for southern Asia might reveal evidence for even earlier appearances there than east Africa.

Although not ideal, the sample sizes for Africa, the Levant, most of Southwest Asia, and Central Asia are acceptable for some analysis of specific clades or larger haplogroups, and investigation into spatio-temporal patterns. The samples from the Americas also allow for some investigation. However, the lack of available samples from Pacific regions, especially Australia, only allows very limited analysis at this stage. The ability to examine global spatio-temporal patterns and fully investigate movements of genetic markers across the world will require better representation in some regions.

5.2.2. Technology and Global Sample Availability

Since being founded more than two decades ago, the field of ancient DNA research continues to grow but shares a common problem with forensics, that the amount of DNA available within the samples is often limited (Rohland and Hofreiter 2007). As discussed in

Chapter 3, the techniques used to extract and analyze human DNA have steadily improved over the last twenty years, so that samples found in climates that were once not ideal for this analysis now may be processed successfully. Sampling from tropical locations has lagged behind temperate and arctic regions partly due to better preservation of ancient DNA, but with recent excavations such as Mota Cave in Ethiopia, perhaps more revealing discoveries will emerge from Oceania and Africa (Slatkin and Racimo 2016).

With time, researchers may be able to run analysis on many more finds that previously had not been categorized for mtDNA or Y-DNA haplogroups due to difficulty in extraction of uncontaminated DNA with the required quality and quantity. For example, two famous individuals found in Australia near Lake Mungo were omitted from this geodatabase due to controversies over the proper identification of their DNA lineages. DNA analysis on the finds was revisited recently (Heupink et al. 2016), and although these results were not included at this time, it is likely that these later identifications of mtDNA or Y-DNA will be found to be reliable. This hopefully will lead to more samples being conclusively typed and yield much-needed diversity in the collection of analyzed samples. Continued efforts to collect data on archaeological finds and incorporate them into a geodatabase will create a constantly improving resource for investigation of archaeogenetics and all the fields making use of it.

5.2.3. Cultural, Legal, and Moral Considerations

The size of available analysis to incorporate can be expected to grow in most regions, but in some the handling of human remains is subject to strict limitations. Although ethics figure largely in any scientific endeavor involving human remains, the cultural and religious practices of some peoples in the Americas and Australia control what is permissible and what is not in the treatment of finds that could have belonged to an ancestor. Laws and cultural practices do not

necessarily prohibit the study of human remains; nevertheless, some institutions may not participate in this because of the legal requirements and sensitivity. It is highly likely that the size of samples from these continents will remain smaller in proportion to those from others.

In the U.S., the Native American Graves Protection and Repatriation Act (NAGPRA) holds force. The software for databasing and managing museum holdings, EMu by Axiell, includes tools for North American users to comply with legal requirements linked to indigenous people's artifacts and physical remains, such as repatriation after allowed studies as specified in NAGPRA ("EMu – Collections Trust" n.d.; "Axiell Go" n.d.). Australian public bodies are similarly bound to comply with the Return of Indigenous Cultural Property Program (Feikert 2009).

Although these laws and those of other nations with indigenous populations may not necessarily prevent any scientific examination of remains, scientific testing may be delayed or outright prohibited while authorities determine if the remains belong under control of a certain group. The beliefs, customs, and leadership of the group deemed to have authority will determine the outcome. In some cases, this means reburial (or other cultural funerary practice) by the group once analysis is done, but in others, DNA analysis would not be permitted whatsoever. For various reasons, some organizations such as the Australian Capital Territory (Feikert 2009) choose not to hold artifacts or remains at all, and so do not participate in the legal guidelines.

The famous case of the "Kennewick Man" found in 1996 by teenagers in Washington state, is one such instance. Initially, some researchers believed that the remains were possibly related to other ancient peoples from Eurasia that had made the journey to North America in prehistoric times, but later disappeared as a distinctive group - or perhaps left no descendants at all. As such, claims to the remains by the tribes of Columbia Plateau and Nez Perce were not

recognized by scientists (Goldberg 2006). However, the Army Corps of Engineers halted testing, agreeing with the Native American tribes, and a court case followed. Numerous appeals came to the courts, with the right to scientific access finally being hinged upon the need for proof that "human remains bear some relationship to a *presently existing* tribe, people, or culture..." as the US Court of Appeals for the Ninth Circuit explained in Bonnichsen vs. United States (2004). The tribes declined to take the case to the Supreme Court, and instead chose to focus their efforts on making changes to NAGPRA.

Whether because of the legal burdens or respect for the spiritual beliefs and emotions involved, it is likely that many researchers will not seek to work on human remains where it would not be acceptable to the peoples who believe themselves to be their descendants or relations. Aside from these present-day ethical and legal considerations, other moral considerations related to past abuses certainly impact international efforts to examine human remains. Disagreements over ownership and proper care of artifacts, mummies, and other remains abound, often a result of colonial practices or flagrant disregard for local law or sensitivities by past collectors.

To protect their cultural heritage and ensure full access to finds for their own researchers, many institutions, particularly those in former colonies (Porr and Matthews 2019), may resist loaning remains to foreign laboratories, especially for DNA extraction. The process requires destruction of at least a small part of the remains, which both damages the find and limits the number of times it can be repeated. However, not all institutions are equipped to carry out this testing, and so work on some samples may be delayed.

These reservations about excavations and analysis on human remains in many areas are complex and understandable (Porr and Matthews 2019). They are unlikely to be resolved to the

satisfaction of all with a stake in the future of human archaeological finds. For this reason, the availability of a fully representative sample of ancient humanity will not be realized soon, although improved techniques in retrieval from archaeological finds open to investigation will bring it much closer to fruition.

5.3. Lessons Learned

Overall, working in Esri's geodatabase format using ArcGIS Pro was uncomplicated and suitable for the needs of the project. Data management tasks, such as creating and working with domains went well. Writing and running queries was not difficult, although some minor issues were noted. The file geodatabase remained a very appropriate choice for work with a project at the present size with one individual responsible for entries and edits.

5.3.1. Domains

Working with domains in the 'archeological_find' helped ensure that values were appropriate and made work with the related DNA tables go smoothly. For this project, domains keep the values of mtDNA and Y-DNA clades to recognized formats. Because so many clade and sub-clade names were possible in this geodatabase, the domains were auto-generated using a readily available script, Esri ArcGIS Pro's 'Table-to-Domain' tool. The source tables were genetic haplogroups tables created to form the basis of the Y-DNA and mtDNA tables in the geodatabase, compiled from names and relationships of the clades and haplogroups described in references for genetic genealogy ("MtDNA – Results (MtDNA – Mutations) – FamilyTreeDNA Learning Center" n.d.) and DNA resource pages ("Welcome to ISOGG... | International Society of Genetic Genealogy" n.d.).

To reduce errors being introduced at the very beginning, the DNA tables to be used as the foundation of the domains in the 'archaeological_find' table were checked for unique values

with filtering options in Excel. Next, this column of clades was double-checked for any entries that did not match known conventions or formats. Once the verification was complete, the tool was used to create domains for Y-DNA and mtDNA clades in the 'archaeological_find' table.

The large number of coded values made the drop-down menu quite long. Nevertheless, it was not cumbersome since the GIS could locate the section of the drop-down when the first characters were entered. The use of numbered types with domains was considered in order to shorten these lists, since the DNA sub-clades, clades and sub-haplogroups could be organized by the older upstream group. However, the use of numbers to represent more than four or five groups did not seem to be a user-friendly solution; it would not be reasonable to expect someone to mentally recall correspondences for so many values or refer to external notes. Continuing with the large number of coded values with their widely understood alphanumeric system seemed preferable. If Esri ever offers subtypes using other codes or a built-in lookup option, this could be implemented without frustrating an editor or possibly introducing more error.

5.3.2. Tool Use in the File Geodatabase

Some tools and features were lacking for the file geodatabase. For some functions, an enterprise geodatabase is required. The cost of the enterprise geodatabase license was prohibitive, and the file geodatabase was adequate for the purpose of the project at this stage. One example of unavailable tools is the "Make Query Layer" tool. The "Make Table Query" tool was available and provides comparable results. The geometry field can be included in "Make Query Table" tool to give a mappable result or can be omitted to give data tables only.

The relationship classes in Esri worked as expected, allowing for normalized tables. The only drawback to this was that executing some queries in ArcGIS Pro was less straightforward than they would have been using DBMS tools. One issue encountered was inconsistent behavior

in one tool using SQL. The SQL window in the "Make Query Table" tool has an option to verify that the SQL entered is valid. In at least one instance, the SQL was returned as "valid", yet the query failed, giving the "An invalid SQL statement was used" error. Conflicting messages as these do not help establish whether an alternative form of SQL may be used, if the user did not construct the query properly, or some other issue was at fault. Nevertheless, by using the aforementioned "Query Table" or by using simple spatial joins and/or definition, query results could be obtained for queries requiring data stored in the separate tables.

5.4. Future Possibilities

Use of the file geodatabase format was simple and more affordable than an Esri enterprise geodatabase. As discussed in 5.3, it was appropriate for its purpose. Nevertheless, for a hypothetical future project involving multiple data owners and a much larger volume of data, some options of the enterprise geodatabase might be desired. There are greater options for use of SQL in direct data management, which is not recommended with the file geodatabase at present, according to Esri's help pages ("Supported Databases—ArcGIS Pro | Documentation" n.d.). As mentioned in 5.3.2, not all tools are available in the file geodatabase. Enterprise would also allow for multiple users to access and edit using versioning. As the project grows, a move to enterprise might be more attractive.

5.4.1. Open Source Options?

The costs of Esri licensing can be prohibitive for many projects, and enterprise licensing adds further expense. Esri offers many benefits, such as a high standard of visualization, a multitude of tools and algorithms readily available for analysis of all kinds, curated data, and resources for training and troubleshooting. This certainly justifies the costs for many organizations, but all budgets do not allow an Esri option no matter how desirable.

5.4.1.1. Quantum GIS – QGIS 3.x

Use of the geodatabase in its present form does not require expensive licensing. For most non-commercial purposes, there is a $100 fee for student or personal use license that can be easily obtained from Esri. Even this minimal fee is not necessary to use the current archaeogenetic database. Access to the file geodatabase in its current form is possible with recent Quantum GIS (QGIS) versions, certainly all after 3.0 can read .gdb files. QGIS is an open source project and is completely free to download and use with the full suite of tools that have been developed for it. QGIS has added more training and user manual information to its website since its initial appearance, and multiple videos on usage for new users can be found on YouTube free of charge. QGIS also can work with a PostgreSQL database coupled with the PostGIS extension. This offers some choices for porting the project to open source in the future if desired.

5.4.1.2. PostgreSQL

If use of an enterprise option should become desirable, but Esri licensing is still not possible, then PostgreSQL with PostGIS for spatial capabilities could be a viable alternative. FME Safe Software offers a solution to help convert a geodatabase to PostgreSQL, but this is also a commercial solution. For an open source solution making use of GDAL, the code:

```
ogr2ogr -f "PostgreSQL" PG:"dbname=mydbname user=postgres" myFileGDB.gdb
```

can be used. User Burham also noted that “FileGDB” had to be installed for this to work, but “ogrinfo –formats” could be used to verify its installation (“Ogr2ogr - How to Import ESRI Geodatabase Format .Gdb into PostGIS” n.d.).

As PostgreSQL with PostGIS has continuously updated its offerings and released new versions since 2012, it is likely that this solution, or a similar one, can still work with file geodatabases created with more recent ArcGIS Pro versions. As of 2020, a package available on

GitHub (Iannou 2021) has been used to convert the full file geodatabase, including domains, for use with PostGIS.

Other open source databases with spatial capabilities exist, so further investigation into these options and the benefits or problems associated with them would be needed before migrating the geodatabase (Badea and Badea 2018). Both users of Esri and of QGIS can connect to these alternatives, as well as use the present geodatabase format. Should the project achieve more success and require better options for easy sharing or other needs than the file geodatabase allows, these open source enterprise database solutions are promising. However, work in the file geodatabase format in this project has allowed for simple data management, along with a highly convenient integration with a top-notch GIS and all that it offers.

In conclusion, building the archaeogenetic geodatabase using Esri's ArcGIS Pro was successful. The geodatabase structure was able to easily execute queries like those made in a prototype relational database. Data management tools worked well, which remains important as the project grows with the publication of more analyzed DNA samples. The spatial capability allows for the easy incorporation of special newly published spatial datasets, such as vector files of historic geographic features. Should the project eventually require a move away from Esri software, there are options to migrate the geodatabase without requiring it to be rebuilt. The archaeogenetic geodatabase has proven to be an expandable data repository for spatial analysis.

References

Achilli, Alessandro, Ugo A. Perego, Hovirag Lancioni, Anna Olivieri, Francesca Gandini, Baharak Hooshiar Kashani, Vincenza Battaglia, et al. 2013. "Reconciling Migration Models to the Americas with the Variation of North American Native Mitogenomes." *Proceedings of the National Academy of Sciences* 110 (35): 14308–13. https://doi.org/10.1073/pnas.1306290110

Arias, Leonardo, Roland Schröder, Alexander Hübner, Guillermo Barreto, Mark Stoneking, and Brigitte Pakendorf. 2018. "Cultural Innovations Influence Patterns of Genetic Diversity in Northwestern Amazonia." Edited by Connie Mulligan. *Molecular Biology and Evolution*, August. https://doi.org/10.1093/molbev/msy169.

"Axiell Go." n.d. Accessed October 24, 2020. http://help.emu.axiell.com/v6.0/en/Topics/EMu/Axiell%20Go.htm.

Badea, Ana, and Gheorghe Badea. 2018. *CONSIDERATIONS ON OPEN SOURCE GIS SOFTWARE VS. PROPRIETARY GIS SOFTWARE*.

Baumann, Bailey. n.d. "Finding Environmental Opportunities for Early Sea Crossings: An Agent-Based Model of Middle to Late Pleistocene Mediterranean Coastal Migration," 69.

Beyin, Amanuel. 2011. "Upper Pleistocene Human Dispersals out of Africa: A Review of the Current State of the Debate." *International Journal of Evolutionary Biology* 2011.

Burrows, Cynthia. n.d. "Developing an Archaeological Specific Geodatabase to Chronicle Historical Perspectives at Bethsaida, Israel," 90.

Cavalli-Sforza, Luigi L. and Cavalli-Sforza. 1994. "THE GREAT HUMAN DIASPORAS: A HISTORY OF DIVERSITY AND EVOLUTION." 136578. 1994. https://repository.library.georgetown.edu/handle/10822/545601.

"Converting GIS Vector Data to KML | Keyhole Markup Language." n.d. Google Developers. Accessed March 22, 2021. https://developers.google.com/kml/articles/vector.

"EMu – Collections Trust." n.d. Accessed October 24, 2020. https://collectionstrust.org.uk/software/emu/.

Feikert, Clare. 2009. "Repatriation of Historic Human Remains: Australia." Web page. July 2009. https://www.loc.gov/law/help/repatriation-human-remains/australia.php.

Fu, Qiaomei, Heng Li, Priya Moorjani, Flora Jay, Sergey M. Slepchenko, Aleksei A. Bondarev, Philip L. F. Johnson, et al. 2014. "Genome Sequence of a 45,000-Year-Old Modern Human from Western Siberia." *Nature* 514 (7523): 445–49. https://doi.org/10.1038/nature13810.

Fu, Qiaomei, Matthias Meyer, Xing Gao, Udo Stenzel, Hernán A. Burbano, Janet Kelso, and Svante Pääbo. 2013. "DNA Analysis of an Early Modern Human from Tianyuan Cave, China." *Proceedings of the National Academy of Sciences* 110 (6): 2223. https://doi.org/10.1073/pnas.1221359110.

Gimbutas, Marija. 1963. "The Indo-Europeans: Archeological Problems." *American Anthropologist* 65 (4): 815–36.

Grugni, Viola, Alessandro Raveane, Linda Ongaro, Vincenza Battaglia, Beniamino Trombetta, Giulia Colombo, Marco Rosario Capodiferro, et al. 2019. "Analysis of the Human Y-Chromosome Haplogroup Q Characterizes Ancient Population Movements in Eurasia and the Americas." *BMC Biology* 17 (1): 3. https://doi.org/10.1186/s12915-018-0622-4.

Haak, Wolfgang, Guido Brandt, Hylke N. de Jong, Christian Meyer, Robert Ganslmeier, Volker Heyd, Chris Hawkesworth, Alistair W. G. Pike, Harald Meller, and Kurt W. Alt. 2008. "Ancient DNA, Strontium Isotopes, and Osteological Analyses Shed Light on Social and Kinship Organization of the Later Stone Age." *Proceedings of the National Academy of Sciences* 105 (47): 18226–31. https://doi.org/10.1073/pnas.0807592105.

Heupink, Tim H., Sankar Subramanian, Joanne L. Wright, Phillip Endicott, Michael Carrington Westaway, Leon Huynen, Walther Parson, Craig D. Millar, Eske Willerslev, and David M. Lambert. 2016. "Ancient MtDNA Sequences from the First Australians Revisited." *Proceedings of the National Academy of Sciences* 113 (25): 6892–97. https://doi.org/10.1073/pnas.1521066113.

Hrdlička, A. 1907. *Skeletal Remains Suggesting or Attributing to Early Man in North America*. Washington, DC: Government Printing Office.

Hrdlicka, A. 1936. *The Coming of Man from Asia in the Light of Recent Discoveries*. Washington, DC: Smithsonian Institute.

Hublin, Jean-Jacques, Abdelouahed Ben-Ncer, Shara E. Bailey, Sarah E. Freidline, Simon Neubauer, Matthew M. Skinner, Inga Bergmann, Adeline Le Cabec, and Stefano Benazzi. 2017. "New Fossils from Jebel Irhoud, Morocco and the Pan-African Origin of Homo Sapiens." *Nature* Jun 7;546(7657):289-292. https://doi.org/10.1038/nature22336.

Iannou, GM. (2017) 2021. *Cartologic/Fgdb2postgis*. Python. Cartologic. https://github.com/cartologic/fgdb2postgis.

Krzewińska, Maja, Gülşah Merve Kılınç, Anna Juras, Dilek Koptekin, Maciej Chyleński, Alexey G. Nikitin, Nikolai Shcherbakov, et al. 2018. "Ancient Genomes Suggest the Eastern Pontic-Caspian Steppe as the Source of Western Iron Age Nomads." *Science Advances* 4 (10): eaat4457. https://doi.org/10.1126/sciadv.aat4457.

Morrish, Seán William, and Debra F. Laefer. 2010. "Web-Enabling of Architectural Heritage Inventories." *International Journal of Architectural Heritage* 4 (1): 16-37. https://doi.org/10.1080/15583050902731056.

"MtDNA – Results (MtDNA – Mutations) – FamilyTreeDNA Learning Center." n.d. Accessed February 6, 2021. https://learn.familytreedna.com/user-guide/mtdna-myftdna/mt-results-page/.

"Ogr2ogr - How to Import ESRI Geodatabase Format .Gdb into PostGIS." n.d. Geographic Information Systems Stack Exchange. Accessed February 27, 2021. https://gis.stackexchange.com/questions/83016/how-to-import-esri-geodatabase-format-gdb-into-postgis.

"Opening File Based Geodatabases in QGIS 2.4 • North River Geographic Systems Inc." 2014. *North River Geographic Systems Inc* (blog). September 8, 2014. https://www.northrivergeographic.com/archives/opening-file-based-geodatabases-qgis-2-4.

Pääbo, Svante. 2014. *Neanderthal Man: In Search of Lost Genomes*. Basic Books.

Porr, Martin, and Jacqueline Matthews. 2019. *Interrogating Human Origins: Decolonisation and the Deep Human Past*. Routledge.

Rasheed, Haroon ur, Muhammad Jawad, Shahid Nazir, Saadia Noreen, and Allah Rakha. 2017. "MtDNAmap: Geographic Representation of MtDNA Haplogroups." *Forensic Science International: Genetics Supplement Series* 6 (December): e516–17. https://doi.org/10.1016/j.fsigss.2017.09.208.

Reich, David, Nick Patterson, Desmond Campbell, Arti Tandon, Stéphane Mazieres, Nicolas Ray, Maria V. Parra, et al. 2012. "Reconstructing Native American Population History." *Nature* 488 (7411): 370–74. https://doi.org/10.1038/nature11258.

"RFC 7946 - The GeoJSON Format." n.d. Accessed March 22, 2021. https://tools.ietf.org/html/rfc7946.

Rohland, Nadin, and Michael Hofreiter. 2007. "Comparison and Optimization of Ancient DNA Extraction." *BioTechniques* 42 (3): 343–52. https://doi.org/10.2144/000112383.

Slatkin, Montgomery, and Fernando Racimo. 2016. "Ancient DNA and Human History." *Proceedings of the National Academy of Sciences* 113 (23): 6380–87. https://doi.org/10.1073/pnas.1524306113.

Stringer, C.B., and P. Andrews. 1988. "Genetic and Fossil Evidence for the Origin of Modern Humans." *Science* 239 (4845): 1263–68.

Stringer, Chris, and Peter Andrews. 2005. *The Complete World of Human Evolution*. London & New York: Thames & Hudson.

"Supported Databases—ArcGIS Pro | Documentation." n.d. Accessed September 21, 2020. https://pro.arcgis.com/en/pro-app/help/data/databases/dbms-support.htm.

Twumasi, Bright Osei. 2002. “Modelling Spatial Object Behaviours in Object- Relational Geodatabase.” International Institute for Geo-Information Science and Earth Observation.

Underhill, Peter A., and Toomas Kivisild. 2007. “Use of Y Chromosome and Mitochondrial DNA Population Structure in Tracing Human Migrations.” *Annual Review of Genetics* 41 (1): 539–64.

Underhill, Peter A., G. David Poznik, Siiri Rootsi, Mari Järve, Alice A. Lin, Jianbin Wang, Ben Passarelli, et al. 2015. “The Phylogenetic and Geographic Structure of Y-Chromosome Haplogroup R1a.” *European Journal of Human Genetics* 23 (1): 124–31. https://doi.org/10.1038/ejhg.2014.50.

Viciani, D., F. Geri, N. Agostini, V. Gonnelli, and L. Lastrucci. 2018. “Role of a Geodatabase to Assess the Distribution of Plants of Conservation Interest in a Large Protected Area: A Case Study for a Major National Park in Italy, Plant Biosystems.” *An International Journal Dealing with All Aspects of Plant Biology* 152 (4): 631-641,. https://doi.org/10.1080/11263504.2017.1308974.

Wei, Lan-Hai, Ling-Xiang Wang, Shao-Qing Wen, Shi Yan, Rebekah Canada, Vladimir Gurianov, Yun-Zhi Huang, et al. 2018. “Paternal Origin of Paleo-Indians in Siberia: Insights from Y-Chromosome Sequences.” *European Journal of Human Genetics* 26 (11): 1687–96. https://doi.org/10.1038/s41431-018-0211-6.

“Welcome to ISOGG... | International Society of Genetic Genealogy.” n.d. Accessed February 6, 2021. https://isogg.org/.

Wilson, Allan C., and Rebecca L. Cann. 1992. “The Recent African Genesis of Humans.” *Scientific American* 266 (4): 68–75.

“Working with File Geodatabases (.Gdb) Using QGIS and GDAL | Geospatial @ UCLA.” n.d. Accessed October 25, 2020. https://gis.ucla.edu/node/53.

Yang, Ning Ning, Stéphane Mazières, Claudio Bravi, Nicolas Ray, Sijia Wang, Mari-Wyn Burley, Gabriel Bedoya, et al. 2010. “Contrasting Patterns of Nuclear and MtDNA Diversity in Native American Populations.” *Annals of Human Genetics / Stories of Human Genetics*, September, 74(6):525-38. https://doi.org/10.1111/j.1469-1809.2010.00608.x

Appendix A Data Table

Description of the data sets and tables

region	Low resolution - global/Extent:West -179.999989 East 179.999989 North 83.623600 South -89.000000	Holocene (current epoch)	Esri shapefile = polygon	Yes	Esri data portal
mtDNA	Nonspatial	from early as 100 kya to present	table, nonspatial	Yes, but expanding	Created through Family Tree DNA, ISOGG
yDNA	Nonspatial	from early as 100 kya to present	table, nonspatial	Yes, but expanding	Created through Family Tree DNA, ISOGG
archaeological_find	Moderate resolution- Coordinate precision to 6 decimal points (using location names in sources)/ Extent West -175.110590 East 170.230494 North 70.776562 South -55.248573	43,000 BCE to 1450 AD	Esri shapefile = point	Yes	Multiple published papers- including Mathiesen et al. 2017; OSM public project served as reference to locate relevant papers quickly

Appendix B Published Data

Excel Table included in Mathiesen et al. (2017) as supplemental material

Analysis_Label	Culture	Sample_ID	Y-HG	mtDNA	Average	Date	Location	Country	Sex	Coverage	SNPs
Balkans_BronzeAge	Bulgaria_Ezero_EBA	Bul10	..	..	4957	3090-2924 calBCE	Sabrano	Bulgaria	F	0.066	66247
Balkans_BronzeAge	Bulgaria_Beli_Breyag_EBA	Bul6	I2a2	..	4450	3400-1600 BCE	Beli Breyag	Bulgaria	M	0.823	370439
Balkans_BronzeAge	Bulgaria_Beli_Breyag_EBA	Bul8	I	..	4450	3400-1600 BCE	Beli Breyag	Bulgaria	M	0.017	18337
Balkans_BronzeAge	Bulgaria_MLBA	I2163	R1a1a1b2	U5a2	3638	1750-1625 calBCE	Merichleri, K	Bulgaria	M	4.106	825494
Balkans_BronzeAge	Bulgaria_EBA	I2165	I2a2a1b1b	T2f	4908	3020-2895 calBCE	Merichleri, K	Bulgaria	M	5.55	857708
Balkans_BronzeAge	Bulgaria_EBA	I2175	I2a2a1b1	K1c1	5122	3328-3015 calBCE	Smyadovo	Bulgaria	M	0.527	423781
Balkans_BronzeAge	Bulgaria_BA	I2510	G2a2a1a2	H4a1	4758	2906-2710 calBCE	Dzhulyunitsa	Bulgaria	M	6.7	821681
Balkans_BronzeAge	Bulgaria_BA	I2520	H2	H	5132	3336-3028 calBCE	Dzhulyunitsa	Bulgaria	M	6.117	795071

Image of data as published by Prendergast et al. 2019

Lab ID	Site	Map no.	Latitude (°)	Longitude (°)	Archaeological association	Genetic cluster	Sex	mtDNA haplogroup	Y chromosome haplogroup
I12533	Prettejohn's Gully (GsJi11)	15	−0.545	36.106	Early pastoral?	PN outlier	M	K1a	E2(xE2b); E-M75
I12534	Prettejohn's Gully (GsJi11)	15	−0.545	36.106	Early pastoral?	PN outlier	F	L3f1b	N/A
I8874	Cole's Burial (GrJj5a)	14	−0.442	36.267	PN	PN cluster	M	L3i2	E1b1b1a1a1b1; E-CTS3282
I8809	Kisima Farm, A5/Porcupine Cave	2	0.458	36.709	PN	PN cluster	M	M1a1	E1b1b1b2b2a1; E-M293
I8820	Kisima Farm, A5/Porcupine Cave	2	0.458	36.709	PN	PN cluster	F	M1a1f	N/A

Appendix C Domains

List of mtDNA domain coded values

A2c	C1c	D1t	G3a	H28	H45	H7a	I3a	K2b	M13	M5b	Q1b	T2	U4	V14	X2b
A2d	C1d	D2a	G3b	H29	H46	H7b	I4a	K3	M1a	M65	Q2a	T2+	U4a	V17	X2c
A2h	C1e	D4	H18	H2a	H49	H7c	I5a	L0a	M1b	M70	R	T2a	U4b	V1a	X2d
A2i	C1g	D4a	H1a	H2b	H4a	H7d	I5b	L0d	M20	M7b	R+1	T2b	U4c	V1b	X2f
A2p	C4	D4b	H1b	H2c	H4d	H7f	I6	L0f	M21	M7c	R0	T2c	U4d	V3	X2i
A4f	C4a	D4e	H1c	H3	H5	H8c	J	L0k	M28	M8a	R0a	T2d	U5	V7a	X2l
A8a	C4b	D4h	H1e	H3+	H5'	H92	J1	L1c	M3	M9a	R1a	T2e	U5a	W	X2m
B	C4d	D4j	H1f	H30	H5+	H9a	J1	L2a	M30	N	R1b	T2f	U5b	W1	X2p
B2	C5c	D4m	H1g	H32	H5a	HV	J1+	L3b	M33	N1a	R2	T2g	U6a	W1-	X4
B2a	C7a	D4o	H1h	H33	H5b	HV-	J1b	L3d	M35	N1b	R2+	T2h	U6b	W1+	Z1
B2b	Cb1	D4q	H1i	H35	H5c	HV+	J1c	L3e	M3a	N9a	R3	T2k	U6d	W1c	Z1a
B2i		D5a	H1j	H3a	H5d	HV0	J1d	L3f	M3c	N9b	R30	U*	U7a	W1e	Z3a
B2y		F1a	H1k	H3b	H5n	HV1	J2a	L3h	M4		R5a	U1a	U7b	W3a	
B4a		F1b	H1n	H3c	H60	HV2	J2b	L3i	M49		R6a	U1b	U8a	W3b	
B5a		F1d	H1q	H3f	H65	HV4	K	L3x	M4a		R6b	U2	U8b	W5	
B5b		F1e	H1t	H3g	H66	HV6	K1	L4a			R7	U2'	U8c	W5a	
		F1f	H1u	H3h	H67	HV9	K1	L4b			S2a	U2+		W6	
		F2a	H2	H3t	H6a	I1	K1a	L5b				U2a		W6a	
		F2c	H2+	H3u	H6b	I1a	K1b					U2b		W6c	
		F2g	H23	H3v	H6c	I1b	K1c					U2c			
						I1c	K1d					U2d			
												U2e			

List of Y-DNA domain coded values

A00	E	G2a2b	I	I2a2	J	L	Q	R1a
A1	E1b1b1b2	G2a2b1	I1	I2a2a	J1	L1	Q1	R1a1
A1b1b2	E1b1b1b2a	G2a2b2a	I1a1b1	I2a2a1	J1a	L1a	Q1a	R1a1’2
A1b1b2a	E1b1b1b2b2a1	G2a2b2a1a	I1a1b3	I2a2a1	J1a2a1a2d2b2	L2	Q1a*	R1a1a
A1b1b2b	E1bE1b1b1a1a1b1	G2a2b2a1a1b1	I1a1b3b	I2a2a1a1a	J1a2b	M1b	Q1a1	R1a1a1
B	E2	G2a2b2a1a1b1a1a1	I1a2a1a2	I2a2a1a1a1	J2	N	Q1a1b	R1a1a1?
B2b	F	G2a2b2a1a1c1a	I1a3	I2a2a1a1a2	J2a	N1a	Q1a1b1	R1a1a1b
BCDEF	F*	G2a2b2a1c	I1b	I2a2a1a2a1a	J2a1	N1a1a1a1a4a1	Q1a2	R1a1a1b1a2
xBT	G	G2a2b2a3	I2	I2a2a1b	J2a1a2a2	N1c1a	Q1a2a	R1a1a1b1a2b1
BT	G1a	G2a2b2b	I2a	I2a2a1b1	J2a1d	N1c1a1a	Q1a2a1	R1a1a1b1a3
C	G2	G2a2b2b1	I2a1	I2a2a1b1b	J2a1h	N1c2	Q1a2a1a	R1a1a1b1a3a
C1a	G2a	G2a2b2b1a	I2a1a1	I2a2a1b1b1	J2a1h2	N3a3′5	Q1a2a1a1	R1a1a1b1a3a1
C1a2	G2a1	G2a2b2b1a1	I2a1a1a	I2a2a1b2	J2a2a	N3a3a	Q1a2a1a1	R1a1a1b1a3b
C1b	G2a1a1	G2a2b2b1a1a	I2a1a1a1a1a1a1e5~	I2a2a1b2a2	J2a8	NO	Q1a2a1b	R1a1a1b2
C1a2a	G2a2	G2b	I2a1a1a1b	I2a2a1b2a2a2	J2b	O	Q1a2a1c	R1a1a1b2a
C1b1a1a1	G2a2a	H	I2a1a2	I2a2a2	J2b2a	O1a	Q1a2b	R1a1a1b2a2a
C2b	G2a2a1	H1a1	I2a1a2a1a	I2a2a2a	J2b2a1	O1a1a1a	Q1a2b2	R1a1a1b2a2b
C2b1a1	G2a2a1a	H1a1a	I2a1b	I2b	K	O1b	Q1aa1c	R1a1c
C2b1a1a	G2a2a1a2	H1a1d2	I2a1b1	I2c	K2	O1b1a1a1b	Q1b1	R1a5
C2b1a1b2	G2a2a1a2a	H1a2a1	I2a1b1a	I2c1	K2b1	O1b1a1a1b1	Q1b2	R1b
C3	G2a2a1a2a1	H1b1	I2a1b1a1	I2c2	K2b1a3	O2a	Q1c	
C6	G2a2a1a2a1a	H2	I2a1b2	I2d	K2b2a2	O2a1c1b1a	R	
CT	G2a2a1a3	H3b		IJ		O3a	R*	
D	G2a2a1b	HIJ		IJK		O3a3b2	R1	
D1b2b	G2a2a1b1					P1	R1*	

List of Y-DNA coded value domains

R1b1
R1b1a
R1b1a1
R1b1a1a
R1b1a1a1
R1b1a1a2
R1b1a1a2a
R1b1a1a2a1
R1b1a1a2a1a
R1b1a1a2a1a1
R1b1a1a2a1a1b
R1b1a1a2a1a1c
R1b1a1a2a1a1c1a
R1b1a1a2a1a1c2b2a1b1a
R1b1a1a2a1a1c2b2b1a1a1
R1b1a1a2a1a2
R1b1a1a2a1a2a1
R1b1a1a2a1a2a1b
R1b1a1a2a1a2a5
R1b1a1a2a1a2b1
R1b1a1a2a1a2c
R1b1a1a2a1a2c1
R1b1a1a2a1a2c1a1
R1b1a1a2a1a2c1e2b3
R1b1a1a2a1a2c1e2b3a1
R1b1a1a2a1a2d
R1b1a1a2a1a2f
R1b1a1a2a2
R1b1a1a2a2c1
R1b1a1b
R1b1a1b1
R1b1a2
R1b1a2a
R1b1a2a1a
R1b1a2a1a1
R1b1a2a1a1b
R1b1a2a1a1c
R1b1a2a1a1c2b2b
R1b1a2a1a1c2b2b1a1
R1b1a2a1a2
R1b1a2a1a2*
R1b1a2a1a2b
R1b1a2a1a2c
R1b1a2a1a2c1
R1b1a2a1a2c1g
R1b1a2a1a2c1g1a1
R1b1a2a1a2c2
R1b1a2a2
R1b1a2a2c1
R1b1b
R1b1b2
R1b1b2a
R1b1b2g
R2
R2a
R2a3a
R2a3a2
R2a3a2b
R2a3a2b2b1
R2a3a2b2c
S1a
T
T1a
T1a1
T1a1a
T1a1a1b2
T1a2b