RUSSIAN-ARMENIAN UNIVERSITY
INSTITUTE OF MATHEMATICS AND HIGH TECHNOLOGIES
DEPARTMENT OF BIOENGINEERING AND BIOINFORMATICS

MOLECULAR PHYLOGEOGRAPHY OF ARMENIANS BASED ON COMPLETE MITOCHONDRIAL GENOME DATA

BY

HRANT HOVHANNISYAN

DISSERTATION

SUBMITTED FOR THE DEGREE
OF CANDIDATE OF BIOLOGICAL SCIENCES (PhD)
(Գ 00.02 – Biophysics and Bioinformatics)

Supervisor: Prof. Levon Yepiskoposyan

YEREVAN – 2017

Contents

ABBREVIATIONS

aa – amino acid

ABC – Approximate Bayesian Computations

aDNA – ancient DNA

AMH – Anatomically modern huma

API – Application Programming Interface

ATP – Adenosine thriphosphate

b.p. – base pair

BSP – Bayesian Skyline Plot

DNA – deoxyribonucleic acid

GD – Genetic distance

HLA – Human Leukocyte Antigens

HVR(S) – Hypervariable region (sequence)

KYA – Kilo-Years Ago

LGM – Last glacial maximum

LSP – Light Strand promoter

HSP – Heavy Strand promoter

MDS – Multidimensional scaling

MLE – maximum likelihood estimation

MRCA – Most recent Common Ancestor

mtDNA – mitochondrial DNA

NRY – Non-recombining region of Y-chromosome

PCA – Principal Coordinates Analysis

rCRS – revised Cambridge Reference Sequence

RFLP – Restriction fragment length polymorphism

SNP – Single Nucleotide Polymorphism

INTRODUCTION

Being a natural corridor bridging Europe and the Middle East, the Armenian Highland played a key role in the Neolithic migrations accompanied colonization of Europe from the Near East. According to archaeological and paleoanthropological data, the colonization of the Armenian highland by modern humans occurred in the late Pleistocene [King et al., 2003]. However, climatic fluctuations during the Last Glacial Maxium (LGM), which began about 25 kya, resulted in the lack of continuous habitation in these areas [Dolukhanov et al., 2004] until the onset of the post-glacial period (ca. 11.5 kya), the favorable climatic conditions of which along with a variety of landscapes, flora and fauna, an abundance of fertile plains and water supply conduced the Armenian Highland to be one of the most attractive regions for permanent human dwellings [Tarasov et al., 2000; Dolukhanov et al., 2004]. It is considered that the recolonization of the Armenian Highland by the Near Eastern farmers began in the Neolithic [Dolukhanov, et al., 2004; Herrera et al., 2012], and some researchers even have directly linked the origin and development of agriculture with this region [Gandilyan, 1997; Hovsepyan et al., 2008; Barnard et al., 2011]. The archaeological data allow distinguishing several important periods in the development of farming in the Armenian Highland: Neolithic (6-5 kya), Chalcolithic (5-4 kya), Bronze and Iron Ages (4-3 kya) [Hovsepyan, 2008].

Though the territory of the Armenian Highland is now being actively explored by archaeologists, so far an unanimous concept characterizing the whole complex of Neolithic migrations in this region does not exist. This gap cab be bridged by the study of the genetic structure of populations autochthonous to the Armenian Highland.

Genetic structure of the Armenians was studied based on Y-chromosomal, mitochondrial DNA (mtDNA) and autosomal genetic systems [Nasidze et al., 2001, 2003, 2004; Weale et al., 2001; Harutyunyan et al. 2009; Herrera et al., 2012, Rootsi et al., 2012]. However, the results of these studies strongly depend on the nature of

the markers used and the length of the DNA fragments analyzed, which are not always compatible. For example, data on the mtDNA and Y-chromosomal variability in the Armenians have revealed that the population is more geographically structured according to the Y-chromosomal markers than by mtDNA, which is mainly caused by the significant influence of the cultural practice of patrilocality [Harutyunyan et al., 2009]. In addition, significant genetic differences were found between the populations inhabiting the southern mountainous (Karabakh, Syunik) and eastern regions of the Armenian Highland in comparison with the territories of the Ararat Valley and the northern, western and central regions of historical Armenia [Weale et al., 2001]. Large-scale study of four Armenian populations (Ararat Valley, Gardman, Lake Van and Sasun), the geographic areas of which roughly cover the territory of historical Armenia, showed that their patrilineal genetic structure is mainly represented by Y-chromosomal haplogroups associated with the Near Eastern farmers – R1b1b1-L23, J2-M172, J1-M267, G-M201 and E1b1b1c-M123 [Herrera et al., 2012]. The results indicate the prevalence of the haplogroup J2 in the western part and the haplogroup R1b1b in the eastern region of the Armenian Highland. The predominance of Neolithic Y-chromosomal lineages in the Armenian gene pool along with paucity of Palaeolithic archaeological remains suggests that the region was sparsely settled before the arrival of early farmers [Herrera et al., 2012]. Furthermore, the geographical distribution of specific Y-chromosomal haplogroups can mark different migratory events associated with the spread of agriculture. Particularly, the haplogroup R1b1b indicates the migration of early farmers into Europe via Anatolia, while the haplogroups J2 and G are associated with population movements through the Caucasus. Thus, traces of the original Neolithic migrations from the Near East have been preserved in the male gene pool of the Armenian Highland, and molecular genetic data allowed the identification of several migrational waves and directions. However, matrilineal genetic portion of the Armenian gene pool, which also has a graet potential in unrevealing the ethnogenesis of the population, is still poorly invesigated.

The accumulated data in the field of mitochondrial population genomics (more than 21 thousand mitochondrial genomes sequenced) have greatly improved our understanding of the topology of the human mtDNA phylogenetic tree and allowed refining the classification of mtDNA based on the regional characteristics of the mitochondrial genome evolution. However, one of the major gaps in the genome-wide mtDNA phylogeny is insufficient knowledge of mitochondrial gene pools of the West Asian populations, including the Caucasus in general, and Armenia in particular. Today, there is a limited number of genetic studies performed on the Armenians using mtDNA markers. Moreover, to the best of our knowledge, only one study addressed the issue of the variability of entire mitochondrial genomes in Armenians (among four other population samples representing the Azerbaijanis, Georgians, Turks and Iranians). The findings revealed unusually high genetic diversity of the groups studied leading to the conclusion of a close genetic relationship between the three South Caucasus populations, despite their affinity to different language families [Schonberg et al., 2013]. Unfortunately, the small size of the samples (28-30 donors in each group) does not allow conducting a comprehensive phylogeographic analysis of the variability of mitochondrial gene pools and molecular dating of mtDNA phylogenetic clusters, as well as reconstructing the genetic history of populations of the region.

To briefly summarize genetic studies based on mitochondrial DNA so far performed on the Armenians (more detailed description is provided in Literaute review chapter), one has to highlight several major problems and drawback of those studies. First, they mostly have considered the Armenians as a single group. Only a limited number of studies [Weale et al., 2001; Herrera et al., 2012], based on the Y-chromosomal markers, have subdivided Armenians into distinct geographic groups and the results of those studies have conclusively reinforced the previous idea that analysing Armenians aș a single group might bring to misleading conclusions [Weal et al., 2001]. Another striking drawback that has repeatedly come up in those studies is the small sampe set, in most of the cases not exceeding 50 samples, which can lead

to marked biases. Additionaly, different studies were performed using different mtDNA markes, thus making it practically impossible to pull all available data together for performing meta-analysis in order to overcome the issue of restricted datasets. Moreover, samling of individulas in most of the cases has been perfomed in Yerevan, capital of Armenia, where the overall population is presented by decendant of many different regions of Historical Armenia.

In the last 10-15 years there have been a number of articles on the population genetics of Armenians, but there are still quite a few unanswered questions concerning the ethnogenesis, history of Armenians and their relationship with neighboring populations. Only with large-scale, well-structured projects on sufficiently diverse Armenian samples and comparative data sets of similar quality it will be possible to answer those questions.

Thus, to conlusivelly address theses fundamental questions on maternal population history of the Armenians, in our work we have collected, sequenced and analyzed a massive dataset (ca 1900 samples) of mitochondrial DNA, both complete mitogenomes and HVR1 seqeunces, representing several geographic regions covering the whole Armenian highland. Moreover, besides of contemporary mtDNA data, here for the first time we have also analyzed a large sample set (52 specimens) of newly colleceted and sequenced mitogenomes representing ancient Armenian DNA (aDNA), collected from different burrial sites in the Republic Armenia and spanning from the 7th millenium BC to the late Iron age and Medieval.

Additionally, as soon as our and other large scale mitogenomic studies require sophisticated bioinformatics analysis of massive datasets of genetic data, in order to facilitate effectivness of storage and analysis of these type of information, we have aimed to design and implement bioinformatics tools that will allow researchers analysing mitogenomic data in a fast, efficient and reproducible ways.

Aim and objectives

The principal aim of our work was to reconstruct the matrilineal genetic history of the Armenians using the partial and compelete contemporary and ancient mitochondrial genomes from geographically different Armenian populations that cover the whole area of the Armenian Highland.

The following specific objectives were identified:

1. To describe and analyze the phylogenomic variability of complete and partial contemporary and ancient mitochondrial genomes in nine Armenian regional groups using vigorous methods of bioinformatics.
2. To assess the level of genetic variability in all considered samples by analyzing diveristy parameters of HVRI sequence, coding and non-coding portions of complete mitogenomes.
3. To determine the rate of genetic affinity among different Armenian groups and between Armenians and neighboring populations, as well as assess the position of Armenians on the mitochondrial genetic landscape of the Near East.
4. To clarify the pattern of demographic changes of the Armenian population during their history through applying simulation modeling techniques on the two aforementioned types of data.
5. To construct the most plausible demographic model of the ethnogenesis of the Armenians using a combination of ancient and contemporary mtDNA data and applying Coalescence simulation modeling and Approximate Bayesian computations.
6. To design and implement a new human mitogenomic database and develop bioinformatics tools for enhancing the efficacy of large scale human mitogenomic studies.

Scientific and practical significance of the results.

In our study for the first time the Armenian mitochondrial genetic pool was examined using high quality complete and partial mitochondrial genomes from several distinct regional groups of Armenians, with overall dataset set about 1900 samples. Moreover, we have also analysed the first so far available ancient mitogenomes from the Republic of Armenia and Karabakh and using different methods of bioinformatics have compared contemporary and ancient Armenian groups, as well as compared them with more than a hundred populations worldwide. Using these huge and highly informative datasets, we have shown that all the Armenian populations are highly diverse in context of haplogroup composition, though being very similar compared to each other. Moreover, our study has revealed that Armenian matrilineal gene pool does not contain Central or East Asian genetic traces despite numerous recent migration of Asian groups through the Armenian Highland. Using methods of multivariate analysis we have shown that Armenian groups display close genetic affinity and that the Armenian cluster has an intermediate position between the Levant, Caucasus and European populations, thus indicating an autochthonous origin of the Armenians on the territory of the Armenian Plateau.

Further, we have also tested whether the Armenian groups, including ancient samples, are significantly different from each other and have revealed that most of the groups are statistically not different, pointing to genetic homogeneity of global Armenian population, which might be explained by the cultural practice of patrilocality.

We have also reconstructed the most plausible demographic history of all the Armenian groups (separately and in aggregation) using simulation modelling of effective population size *Ne*, that allowed tracing the fluctuations of population sizes throughout at least 40 thousand years, which has also revealed similar population history of all the Armenian groups. Moreover, the overall pattern of population size changes allowed making inferences on the role of the Armenian Highland as a refugial zone for modern humans during and after the Last Glacial Maximum.As

soon as all results of comparison between the contemporary and ancient datasests (similar haplogroup distributions, similar grouping patterns on the multivariate analysis plots and similar patterns of demographic fluctuations) were pointing to the assumption that modern Armenian populations have very close relationships with the ancient data, the ultimate hypothesis of our dissertation was to check whether there is a genetic continuity between aforementioned datasets. To address this question, we have used recently developed but well established methods of simulation modelling and approximate Bayesian computations (ABC). Based on it we have shown that, indeed, the ancient samples represent direct ancestors of modern Armenian populations, strikingly highlighting that Armenian maternal genetic history is at least 3.5-4.0 ky old.

Summarising, our work has conclusively addressed a wide spectrum of fundamental ethnogenetic questions concerning Armenians, which were not elucidated or had weaker support based on previous studies. On the other hand, in this work we have designed and implemented a new database for complete human mitochondrial genomes and empowered it with bioinformatics and data parsing tools, that allow researchers from several fields, i.e. population genetics, medical genetics or forensic science, performing large scale mitogenomics studies in a more convenient and effective way, thus also adding a practical significance to our work.

Approbation. Proceeding of the dissertation have been and will be presented at:
IX and X Annual Scientific conferences of Russian-Armenian University, 2014, 2015, Yerevan, Armenia; Computer Science and Information Technologies 2015, Yerevan, Armenia; Conference «Application Of Modern Scientific Methods And Technologies In Expertise Sphere», National Bureau of Expertise of RA, 16-17 June, 2015, Tsakhkadzor, Armenia, Conference «DNA polymorphism in human populations», 7-10 December, 2016, Paris, France.

Publications. The main results of the dissertation are reflected in 8 scientific papers.

Structure. The dissertation comprises 113 pages of computer-formatted English text, including 12 tables and 20 figures, and consists of the following sections: Introduction, Literature Review, Materials and Methods, Results and Discussion, Conclusion, Inferences, and References (including 167 sources).

CHAPTER 1. LITERATURE REVIEW

1.1. The description of the Armenian population based on the genetic studies

Situated at the crossroads of Europe and the Middle East, the Armenian Highland during its long history has served as both a recipient and conduit for gene exchange between the two regions. While archaeological evidence for modern human, as well as the Neanderthal activity in Armenia, during the Palaeolithic exists [Arslanov et al., 2007; Pinhasi et al., 2008], the Last Glacial Maximum (LGM) likely made permanent settlements in the region infeasible [Dolukhanov et al., 2004] until the glacial recessions between 16 and 18 kya [Akçar et al., 2007]. Though archaeological evidence for Mesolithic sites in Armenia are sparse [Kartal, 2003], the improving climate during this period allowed the Armenian highlands to gradually transform into a region characterized by abundant water supply and wealth of fertile plains [Redgate, 2000]. These conditions as well as its proximity to the Fertile Crescent catalyzed the region's emergence as one of the earliest agricultural areas (ca. 8 kya) during the Neolithic Revolution [Hovsepyan et al, 2008]. Archaeological investigations suggest a prominent role for Armenia in the trade of culture and ideas during the Neolithic. Furthermore, the Armenian highlands might be considered as a region with the earliest known development of leather footwear [Pinhasi et al., 2010] and viticulture [Barnard et al., 2011] - technologies that would later acculturate across the Near East and eventually enter Europe.

Armenians have been living around the area Armenian Highland for thousands of years. The history of Armenians is very long and starts from times that we do not possess many records about [Redgate, 2000]. From the beginning the Armenians were not nomadic tribes but rather a sedentary population utilizing the benefits of agriculture. This is important when considering various hypotheses of the origin of Armenians. There is no conclusive opinion regarding the exact origin of Armenians, whether they were the inhabitants of the Armenian Highland from the very beginning or whether they were migrants from a different area, whether their language

developed in the area or it was brought by some others who came later and left their language. A number of hypotheses exist regarding the issue of the origins of Armenians, each one of them relying on certain mythological, historical, archeological, anthropological and linguistic evidence; however, not all of the hypotheses have the similar level of strength. However, today the most reliable and trustworthy methodology for defining this or other issues about the population history is the population genetics.

Genetic structure of the Armenians was mainly studied mainly using Y-chromosomal and autosomal genetic markers [Nasidze et al., 2001, 2003, 2004; Weale et al., 2001; Harutyunyan et al. 2009; Herrera et al., 2012, Rootsi et al., 2012, Hovhannisyan et al., 2014, Pagani et al., 2016, Mallick et al., 2016], which describe paternal and biparental genetic heritages of the studied population, respectvely.

The majority of the studies, however, considered Armenians as one group in the context of other populations and did not subdivide the sample. This can lead to the rough understanding of the genetic diversity of Armenians and their relationships with other ethnic groups. Since the development and widespread use of modern genetic technologies, population genetics shifted its focus from phenotypic markers to true genetic polymorphisms. Population genetic studies on Armenians have been performed since 2000 both in the wider context of ethnic groups and focusing on local differences within geographical subgroups of Armenians. These studies were conducted predominantly using Y-chromosomal and autosomal markers. The description of several of the most important of them is given below.

In 2001, Weale et al. have analyzed 734 samples of Armenians using 11 single SNPs and 6 microsatellites of Y-chromosome. This is the first study that has provided proper categorization of Armenians living in different geographic areas. Dividing the whole sample into six geographic groups of Armenians yielded an important result: the analysis has revealed a significant geographic stratification of Armenian population and their relative isolation from each other.

In recent years much more SNPs in non recombining portion of Y chromosome have been discovered, thus refining the distinct Y chromosomal lineages. One recent study has genotyped more than 400 samples from four groups of Armenians for 70 binary Y chromosomal markers and 17 microsatellites loci [Herrera et al., 2012]. The most frequent major haplogroups detected in Armenians were R (38% in Ararat valley, 36% in Gardman, 33% in Van, 34% in Sasun) and J (38%, 36%, 43% and 27% in the same groups). Within the haplogroup R, the majority of Y chromosomes belonged to the lineage R1b1b2* (33%, 31%, 32%, 15%, respectively), a predominantly Near Eastern lineage. It is noteworthy that a different lineage of haplogroup R is found commonly in Europe, R1b1b1a. Surprisingly high frequencies of haplogroup T were observed in Sasun (20.1%), while the frequency of this haplogroup is low in other Near Eastern populations and possibly originates in Levant. The results of that study suggest Neolithic origin of Armenian population, coinciding with the spread of agriculture in the area [Kushnareva, 1997; Hovsepyan et al., 2008].

Population genetic studies using autosomal genetic markers have also been performed on Armenians. Molecular genetic analysis of Armenians based on HLA markers has been performed recently [Matevosyan et al., 2011] using more than 4000 samples for low resolution typing and 100 samples for high resolution HLA typing. The most frequent high resolution HLA alleles detected in Armenians were HLA-A*0201 (15.5%); HLA-B*5101 (17.5%), DRB1*1104 (11.5%). In that report Armenians were divided according to geographical regions of Armenia, Karabakh and Diasporan communities. Authors noted general homogeneity of the Armenian population, although some structuring was present, such as proximity of Karabakh and Syunik.

Yet, the Armenian mtDNA genetic structure, which directly reflects the maternal portion the gene pool, is studied relatively poor.

To summarize all population genetic studies so far performed on the samples of Armenian ancestry, it is necessary to point out one major problem that mostly they have considered Armenians as a single group. Only a couple of studies [Weale et al., 2001; Herrera et al., 2012] have subdivided the Armenian population into several geographic groups and the results of those studies have conclusively reinforced the previous point that it is a mistake to consider Armenians as genetically homogeneous [Weal et al., 2001]. Another issue that has repeatedly come up in those studies is the small number of samples used, which can bias the results. The way those studies have been performed, each with different set of markers, makes it very difficult to perform meta-analysis to overcome the problem of small datasets.

1.2. Mitochondrial DNA as a versatile tool for ethnogenomic studies

1.2.1. The applications of molecular genetic markers in population genetics

The progress of molecular biology in past two decades is characterized by the extensive improvement of nucleic acid extraction, purification and especially sequencing methods. These advances provided novel molecular genetics tools for evolutionary, ecological and forensic studies, which allow resolving problems of these disciplines in more accurate way [Kivisild T., 2015, Stoneking M., 1993]. Rapidly accumulating data on human genetic variation, collected on different types of DNA markers, have been widely used for the reconstruction of genetic history of modern humans [Cavalli-Sforza et al., 2003]. Today, there are three types of molecular genetic markers used in population genetics: autosomal markers, inherited biparently, maternally inherited mitochondrial DNA (mtDNA) genome and the paternally inherited non-recombining portion of the Y chromosome. Among them, mtDNA markers are considered invaluable for human evolutionary and population genetics studies [Cavalli-Sforza et al., 2003]. Analysis of mtDNA variation has matured in the course of the past 20 years and has become a versatile tool in the study of our species for the time horizon of the last 100,000 years as well as our relationship to other species [Bandelt et al., 2006].

1.2.2. mtDNA structure and functions

Mitochondria are mammalian cellular organelles that have function of the oxidative phosphorilation and the formation of ATP. Two distinct genetic systems encode mitochondrial proteins: mitochondrial and nuclear DNA. mtDNA is a small 16,569 base pairs (b.p.) circle of double-stranded DNA which encodes 13 essential components of the respiratory electron transport chain, 2 ribosomal (12S and 16S) and 22 transfer RNA's (Fig. 1).

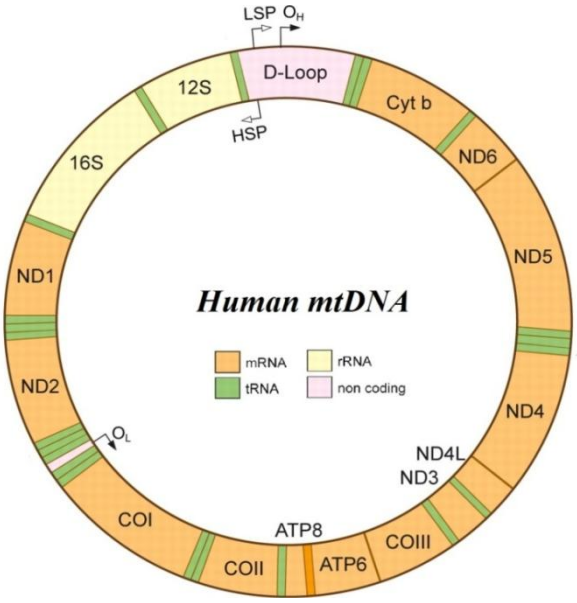


Figure 1. The structure of human mtDNA. Genes are indicated in orange, rRNA's – in yellow, tRNA's – in green, non-coding regions – in violet. LSP – light strand promoter, HSP – heavy strand promoter, O_L – origin of light strand, O_H – origin of heavy strand (Bandelt et al., 2006).

The mtDNA displacement loop (D-loop, or *control region*) is a 1.1-kb non-coding region which is involved in the regulation of transcription and replication of the molecule. The D-loop extends from position 16024 to position 576 of the mtDNA and is the largest region not directly involved in the synthesis of respiratory chain polypeptides. The D-loop contains three short regions which, in comparison to the rest of the genome, have a highly variable sequence at the population level: hypervariable sequence (HVS) HVS-I, HVS-II, and HVS-III, corresponding to HVR1, HVR2, and HVR3 in some sources [Brandstätter et al., 2004a]. The precise

definition of the different hypervariable sequences does vary from context to context. The forensic community traditionally took 16024–16365 to be HVS-I, 73–340 to be HVS-II, and 438–576 to be HVS-III [Brandstätter et al., 2004a]. By contrast, more recent population genetics studies [Brandstätter et al., 2004b] take wider ranges, particularly in HVS-I in order to capture the phylogenetically important positions 16390, 16391, and 16399 (HVS-I 16024–16400, HVS-II 44–340, and HVS-III 438–576). The first complete sequence of human mitochondrial DNA was published in 1981 [Anderson et al., 1981]. Early evolutionary studies have focused on polymorphic restriction sites, but with the rapid development of semi-automated complete mitochondrial genome sequencing, recent work has incorporated full-genome analysis.

Traditionally the human mtDNA is numbered with reference to the light strand according to the original so-called Cambridge Reference Sequence – CRS [Anderson et al., 1981] of mtDNA. In 1999, CRS was resequenced [Andrews et al., 1999] which has revealed 10 substitution errors, and the revised Cambridge reference sequence (rCRS) was considered as the etalon, so that the other obtained mtDNA sequences are to be compared with it.

1.2.3. Mitochondrial genome as a tool for population genetic studies

The phylogenetic analysis of human mitochondrial DNA is based upon the comparison of mtDNA sequences within and between different human populations. Mitochondrial genome has several unique features, which make it an appropriate genetic system for population genetics studies. Mammalian mtDNA is transmitted as a haploid molecule from mother with low frequency of heteroplasmy [Hutchison et al., 1974, Hauswirth et al., 1986], thus the maternal mtDNA lineage of a particular organism doesn't change, excluding *de novo* accumulated mutations. The mutation rate of the mtDNA is 5 to 10 times faster than that of the nuclear genome [Pikó et al., 1976, Brown et al., 1979], reaching 1.64×10^{-7} substitution/nucleotide/year for HVRI [Soares et al., 2009] mainly because mitochondria do not have repair enzymes neither for errors occurring during the replication nor for the damages of the DNA [Cann,

1983]. A specific set of mtDNA mutations defines the mtDNA haplogroups. The last build of global phylogenetic tree of mtDNA variation [van Oven, 2015] includes ca. 30 major haplogroups assigned by letters from A to Z with numerous sub-clades (Fig. 2).

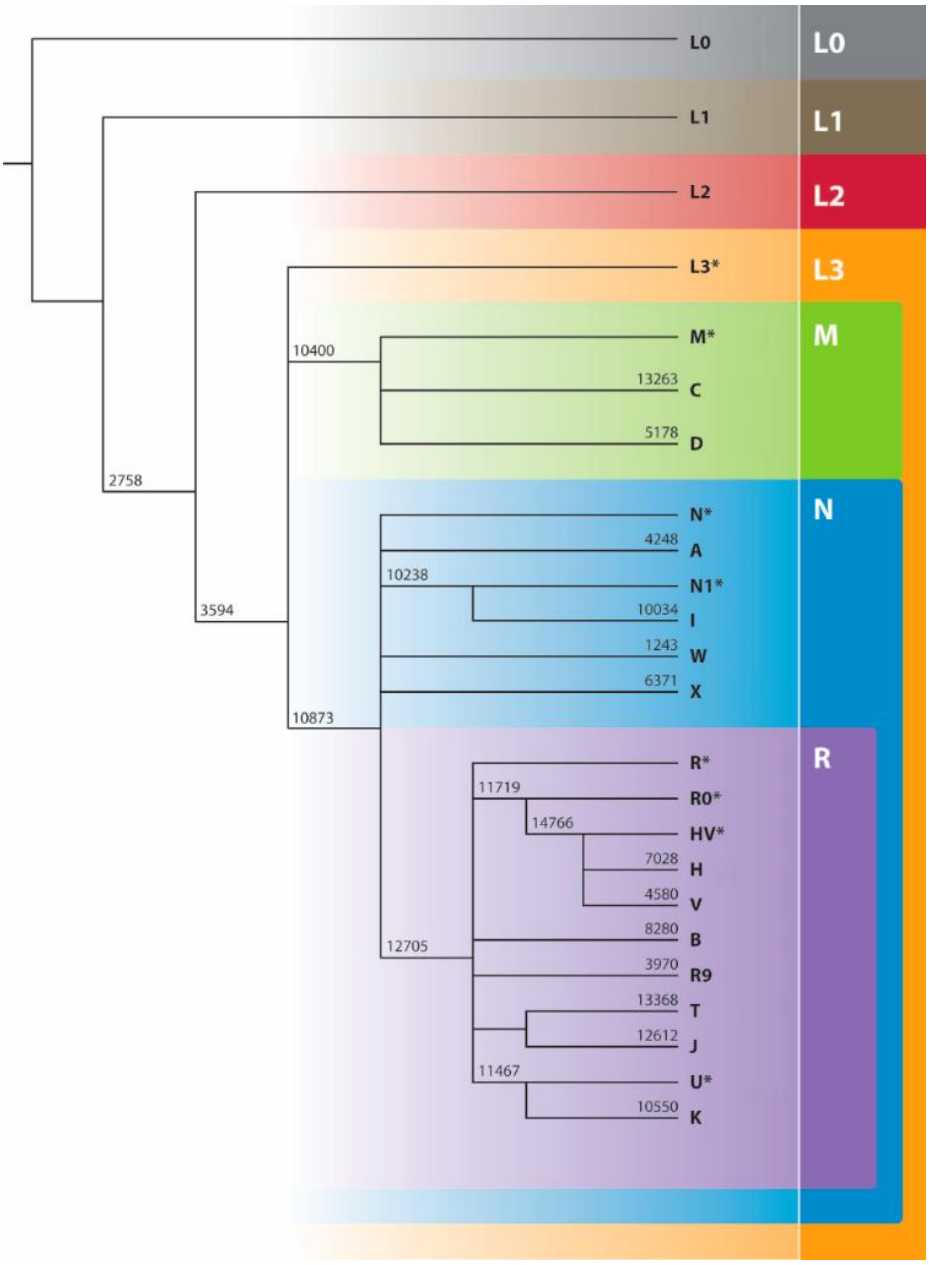


Figure 2. The schematic phylogenetic tree of major mtDNA haplogroups. The numbers on the tree nodes indicates mutations, defining the haplogroups. [Behar et al., 2007].

According to their frequency distribution, mtDNA haplogroups are considered to be specific for the particular geographic region, which however doesn't reflect the

area of origin of the haplogroup. For instance, haplogroups H, J and T considered as West European, C, D and G – Central Asian, etc.

When researchers began to work on human mtDNA in the 1980s after the publication of the famous Cambridge reference sequence of Anderson and colleagues, they first employed only a few restriction enzymes, using them to estimate very simple trees. The genealogical resolution of these trees was so low that they gave quite misleading results, which supported either an 'out-of-Asia' origin for modern humans, or even the multiregional model [Bandelt et al., 2006]. After the application of higher-resolution restriction analysis approaches on human mtDNA [Cann et al., 1987], first implemented at the A. Wilson's (UC Berkeley, USA) group, mitochondrial genome became both famous and notorious when used as an evidence supporting the 'out-of-Africa' model. Finally, in 2000th, the development of fully automated sequencing technologies allowed the massive sequencing of whole mtDNA genomes, which has increased the resolution of mtDNA phylogeny many-folds, opening a new phase of mtDNA research.

1.3. Previous studies on Armenian matrilineal genetic structure

In parallel with the development of DNA sequencing techniques, mitochondrial genome became widely used tool for population genetic research.

Human mitochondrial DNA is inherited maternally [Hutchison et al., 1974], thus does not recombine with paternal molecule [Brown et al., 1979, Michaels et al., 1982], has a fast mutation rate [Pikó, et al 1976] and is presented by multiple copies in cell [Hagström et al., 2014, Merriwether et al., 1991].

These unique features made mtDNA a versatile tool for evolutionary and population genetics studies in the end of last century, when genome-wide or whole genome population studies were a daydream for researchers [Avise et al., 1987, Hartl et al., 1997, Cavalli-Sforza, 1998, Cann et al., 1987, Avise et al., 1984, Kivisild, T., 2015].

This matrilineally transmitted genetic system has been also used to study populations of the Near East and Caucasus, including Armenians [Nasidze et al., 2001, Nasidze et al., 2004]. However, none of those studies were focused on investigation of Armenian gene pool in particular.

The very first studies focusing on matrilineal genetic legacy of the Near Eastern [Nasidze et al., 2001] and Caucasus [Nasidze et al., 2004)] ethnic groups were performed in early 2000th, utilizing only Hyper Variable Region I (HVRI) – a small portion of mtDNA.

From this perspective, a few population genetics studies comprising Armenian datasets were performed.

In this context, the first pivotal study of mtDNA gene pool of the Near Eastern ethnic groups was performed in 2000 by Richards et al. [Richards et al., 2000] where using the technique of founder analysis the authors assessed different migrations that shaped the process of Europe colonization. Here, the authors have utilized HVSI sequences of large sample sets from different populations of Near East (n=1234), North Caucasus (n=208) and Europe (n=2804). As a representative of the Near Eastern ethnic group, relatively large set of Armenian data (n=191) was included in the study. The haplogroup frequency analysis of the Armenians and some other neighboring ethnic groups has shown interesting results. It was revealed that haplogroup U5, which presumably originated in the Near East and evolved in Europe, was restricted in numerous groups from the Near East, except Armenians, Turks, Kurds, Azeris and Egyptians, i.e. in populations located in peripheral areas of the region. Moreover, in Armenian and Azeri gene pools the U5 haplogroup samples derived from European lineages, suggesting back migration from Europe to the Near East through the Caucasus and the Armenian Highland.

Another study of HVRI has been performed in 2001 by Nasidze et al. and involved 353 samples from 9 populations from the Caucasus [Nasidze et al., 2001b]. Within this sample set 42 Armenians were included. The analysis of genetic diversity showed higher values in the Caucassus populations compared to Europeans but lower

than in the Middle East. According to the results, the Caucassus populations clustered together while Europeans formed a separate group. The authors concluded that both the Armenian and Azerbaijani populations underwent a language replacement process in their history. However, the average genetic distance of Armenians was even slightly smaller to other Indo-European populations than to those from the Caucassus, which contradicts the idea of language replacement in case of the Armenians. Moreover, it is important to mention that Iranians, a major Indo-European speaking group of the region, was not included in the comparative data sets. Azerbaijanis, on the contrary, were much closely related to other Caucasus populations than to Turkic-speaking groups, thus enabling to conclude the occurrence of language replacement in that case.

Continuing this series of studies of populations from the Caucasus and the Near East, Nasidze et al. [2004] combined the previously reported Y-chromosomal and mtDNA data [Nasidze et al., 2001b] and added more published data on the populations from the Caucasus and the Near East. However, no new Armenian samples were included. The additional populations allowed more detailed analysis of genetic relationships between the populations and displayed a geographic rather than linguistic influence on the genetic structure of the populations. As for the previous work, the similar pattern of relatedness between Armenians, the Indo-European speaking and Caucasian speaking populations was found, which can be explained by the same sample sets of these populations.

More recent paper of Schönberg et al. [2011] was focused on the investigation of mtDNA gene pool of the Caucasus and West Asian populations using high-throughput sequencing data of mtDNA genomes, which included 147 individuals, representing Georgian (n=28), Azerbaijani (n=30), Iranian (n=30), Turkish (n=29) and Armenian (n=30) ethnic groups. This study has shown a high level of diversity of the populations studied, exceeding that within all of Europe and only slightly lower than the West Asian mtDNA diversity, which might indicate an old age of human populations from this region. All Georgian, Armenian and Iranian individuals had

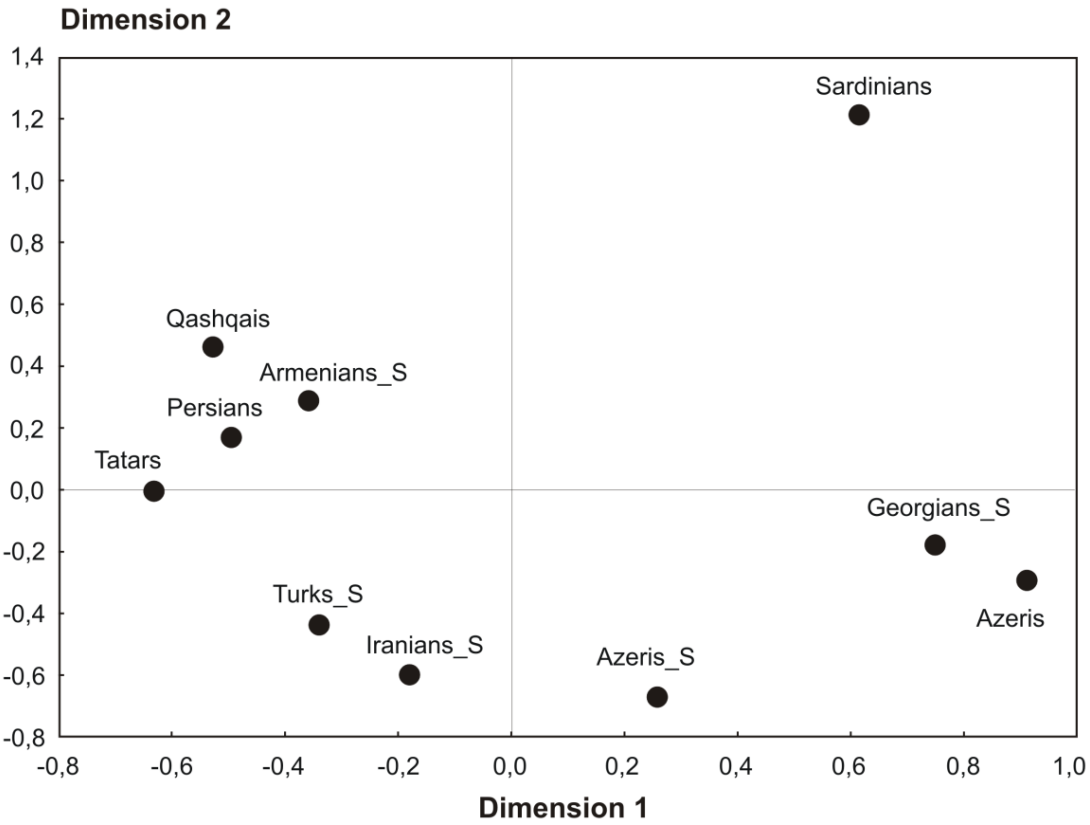
different sequences, while two Azeri individuals shared the same sequence, and 27 haplotypes were identified among 29 individuals of Turkish origin. Complete mtDNA genome sequences indicate that the three South Caucasus groups are genetically similar, although they represent three different language families: Indo European-speaking Armenians and Turkic-speaking Azeri cluster with the Caucasian speaking Georgians.

Here, haplogroup distribution of the populations was assessed. For the Armenians, the most frequent haplogroups were U, H and J, comprising 26.7%, 20% and 16.7%, respectively. Noteworthy, that Asian-specific haplogroups A, C, D and F were completely absent from the dataset of Armenians investigated in the study, while in the samples from Azerbaijan and Turkey these haplogroups were found in relatively high proportions – 13.3% and 17.1%, respectively, indicating that ancient migrations of Central Asian ethnic groups that took place ca a millennium ago have had a substantial contribution into the maternal gene pools of the mentioned populations.

Despite the methodological advantages that were used in this study (i.e. complete mtDNA sequencing and sophisticated bioinformatics approaches), it had some limitations. First, the Armenian, Azerbaijani, Turkish and Iranian samples were collected in the capitals of these countries, without clarification of deep ancestral information of the donors. Second, the size of datasets was quite restricted, and didn't exceed 30 individuals for each population, which hardly could describe the overall genetic diversity of the whole ethnic group. The Armenian sample was presented by 30 individuals from Yerevan; the authors didn't mention the origin of those people even on the grandparental level, which is important when analyzing Armenians, since the population of Yerevan represent various distinct geographical regions of Armenian Highland, which in many cases don't match geographic limits of the Republic of Armenia.

The Armenian full mtDNA sequences presented in the paper of Schönberg et al. [2011] were used in the recent paper of Derenko et al. [2013], though, only 10

samples of Armenian descent were added in this work. This study indicates that all populations from the Caucasus region (Armenians, Azeris, and Georgians) are quite



dispersed (Fig. 3) though their genetic proximity has been demonstrated by Schönberg et al. [2011].

Figure 3. Multidimensional scaling (MDS) plot based on Fst statistics calculated from complete mtDNA sequences for population samples from Iran, Anatolia, Caucasus, and Europe. The Populations from Schönberg et al. 2011 labeled with "S" after underscore. The figure is taken from Derenko et al., 2013.

Nevertheless, these 10 additional samples were obtained from the territory of Central Iran and regional residence of their matrilineal ancestors wasn't determined.

Another paper, which included the Armenian mtDNA data as of a representative population of the South Caucasus, was published by Yunusbayev et al. [2013]. Here, the authors have conducted the analysis of uniparental Y-chromosomal, mtDNA and autosomal markers. The mtDNA samples of the Armenians (n=37), as well as samples of other ethnic groups were genotyped for HVSI, HVSII and coding region loci using restriction fragment length polymorphism analysis (RFLP). In this study, the principal coordinate analysis (PCA) plot based on mtDNA variation of studied

groups placed the Armenians closer to the Caucasus ethnic groups than to the populations of the Middle East. The Caucasus populations did not differ in common mtDNA haplogroup frequencies to an extent that would allow the discrimination of geographic subregions or language groups, indicating the proximity of matrilineal genetic component of the Caucasus groups. Additionally, the ADMIXTURE analysis, based on autosomal SNP's data, has shown that the gene pool of the Armenian population predominantly consists of three components – the Middle Eastern, Caucasus and South Asian, however one should consider that the Middle Eastern and Caucasus components were not clearly differentiated in this analysis in general. Considering that the ADMIXTURE approach was applied to the autosomal data, which doesn't include mtDNA markers, it also was of our interest to determine if the matrilineal genetic components of Armenians are the same as for autosomes, which was not carried out in the study of Yunusbayev et al [2013]. On the other hand, as soon as this work wasn't focused particularly on the Armenians, especially on the maternal genetic component, the same shortcomings that present in other papers described above concerning Armenian datasets are also present here.

The results mentioned above show that Armenian matrilineal genetic portion was described only in the context of studies on other Middle Eastern and Caucasus populations. All studies that were carried out with inclusion of Armenians didn't contain appropriate representative dataset: the number of individuals did not exceed 50, except the paper of Richards et al. [2000] and the origin of those individuals was not clearly determined. These data do not allow making any precise conclusions and inferences on the Armenian matrilineal genetic structure. In summary, it points out that the matrilineal genetic history of the Armenian population is still poorly investigated. Considering the matter that since ancient times and up till now the Armenians occupy the territory of the Armenian Highland, which served as key migration crossroad for numerous populations, the investigation of the Armenian mtDNA gene pool is crucial not only for the reconstruction of the Armenian history, but also for understanding the genetic component of these ancient migrations, which

undoubtedly has affected the Armenian genetic structure. Hence, the current state of the study of the Armenian maternal genetic pool is insufficient for clear understanding of these questions and the place of the Armenians on the genetic landscape of the Middle East.

1.4. Simulation modeling of genetic data in addressing the questions of population history

In evolutionary biology, population simulation plays an increasingly important role in helping researchers to test various genetic models describing the origin of genetic diversity and DNA sequence patterns [Evanno et al., 2005, Dunning et al., 1995, Beaumont et al., 2002, Schaffner et al., 2005, Dupanloup et al., 2002]. Computer simulations are excellent tools for understanding the evolutionary and genetic consequences of complex processes whose interactions cannot be analytically predicted [Yuan et al., 2012]

The general principle is to generate *in silico* data sets (known as pseudo data) of genetic polymorphism under specified scenarios describing the evolutionary history and genetic architecture of a species. For example, in understanding human evolution, one scenario might consider population expansion with a specific geographic origin, ancestral and descendant population sizes, and dispersal rates within and between continents [Malaspinas et al., 2016].

The past decade has witnessed an increased interest in unravelling the evolutionary history of populations, and simulations are crucial for inferring such histories. In particular, simulations have been embraced for inferring demographic expansion and migration in humans [Malaspinas et al., 2016].

A common problem is to evaluate the plausibility of alternative hypotheses and to estimate demographic and genetic parameters under the best-supported model. Typically, data are generated under alternative models of evolutionary history, and relevant population-genetics statistics (such as F_{ST}, number of alleles and F_{IS}) are used to summarize each data set, creating a distribution of possible values under each

scenario. Summary statistics obtained from observed data are matched to these distributions using a variety of ad hoc methods or, more recently, using approximate Bayesian computation (ABC) [Beaumont et al., 2009, Tanaka et al., 2006, Guillemaud et al 2010, Estoup et al., 2012], which is a general statistical approach that has revolutionized the use of simulations for statistical inference [Sunnaker et al., 2013]. Simulations have been used for statistical inference in evolutionary biology, ecology, conservation and epidemiology. Usually, population genetic simulation studies have three main applications – predictive, statistical inference and evaluation of statistical genetics methods.

All of them comprise several steps which are generally described in Figure 4. Simulation programs, i.e. simulators, differ greatly in the evolutionary and demographic scenarios that they consider.

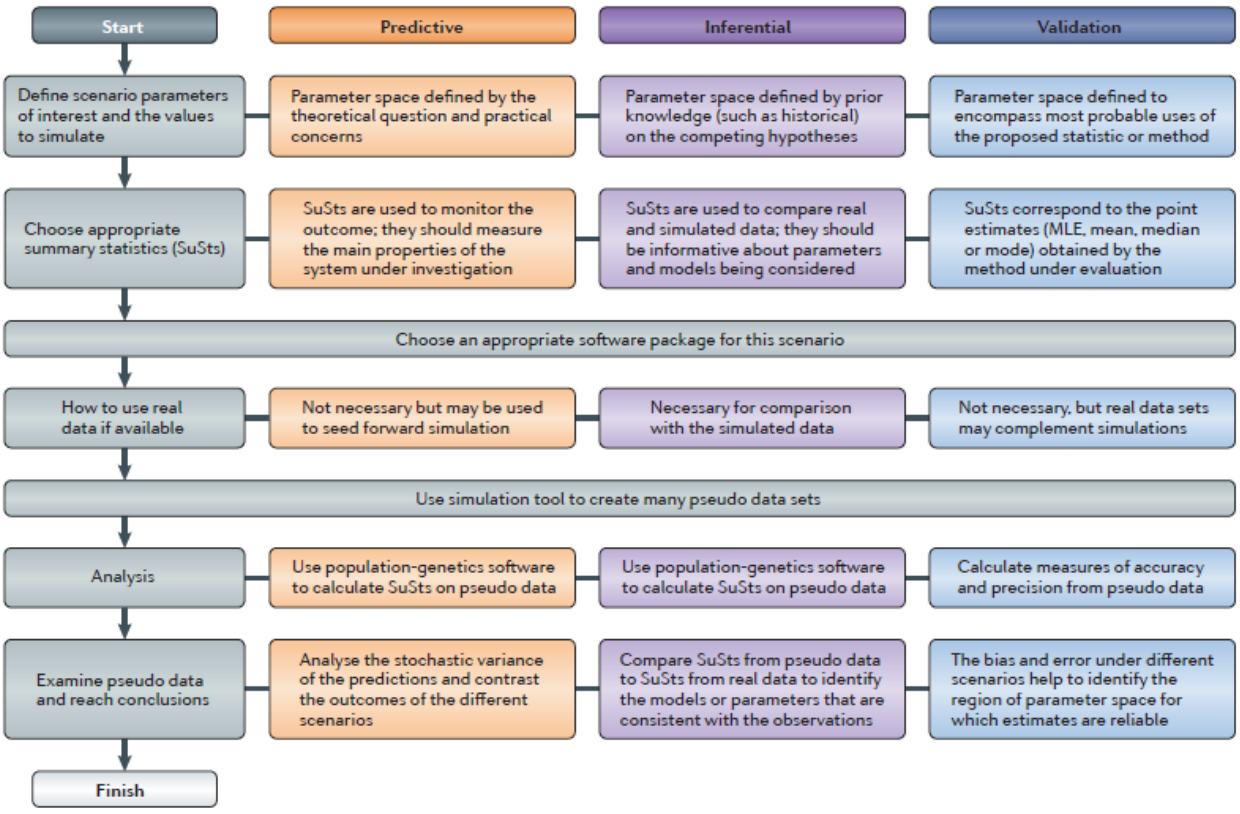


Figure 4. Designing predictive, inferential and validation simulation studies. The left-hand column indicates the steps in simulation study design, and the other columns show the similarities and differences in designing predictive, inferential and validation studies. Grey boxes designate actions that are applicable to all; colored boxes denote differences. MLE – maximum likelihood estimation [Hoban et al., 2012].

Computer simulations are essential for explaining the origin and maintenance of genetic variation and they have applications in many disciplines and the past decade has witnessed much progress in this area, and one can expect simulations to become a standard tool for the study of genetic variation.

Coalescence simulation modeling

Three main categories of genetic data simulation algorithms are backward-time, forward-time, and resampling approaches. The backward-time approach, also known as coalescent simulation [Kingman, 1982], starts from the observed sample in the present generation and works backward — that is, starting from a population of individuals, this approach first traces all alleles to a single ancestor, dubbed the most recent common ancestor (MRCA), and then works forward up to the current generation, introducing mutations or other genetic information into the generated genealogy. The forward-time approach is designed to start from an initial population and track its evolution under various genetic models, over multiple generations, with samples usually drawn from either the final or one of the last several generations. The resampling approach works from existing genomic data sets such as the HapMap data (http://www.hapmap.org/). A simple version of this approach generates samples by bootstrap resampling from the existing data. In the following subsections, we demonstrate the principles for each of these three simulation categories, discussing the primary features of each of these simulator types, as well as their differences.

The coalescent theory was originally developed in the 1980s [Kingman, 2000] and then was extended by many researchers to allow recombination, selection, and other complex evolutionary models. Coalescent simulation in population genetics can be generally divided into two processes. The first is the coalescent process, describing the ancestral history of a sample of individuals originating from the MRCA. Figure 5 shows Kingman's *n*-coalescent process, where four individuals in the present day are coalesced to the MRCA in the past under three coalescent events, and where the expected time interval is

$$E(T_k) = \frac{4N_e}{k(k+1)} \quad (1)$$

Here N_e is the effective population size, k = (1, 2, 3), and T_k is the time interval between the *(k-1)*-st and *k*-th coalescent events. The second process is the permutation process, which describes when and how alleles mutate over time across the genealogy. After these two processes are run, each sample is output as a sequence, with each allele represented by the ancestral state or the derived state.

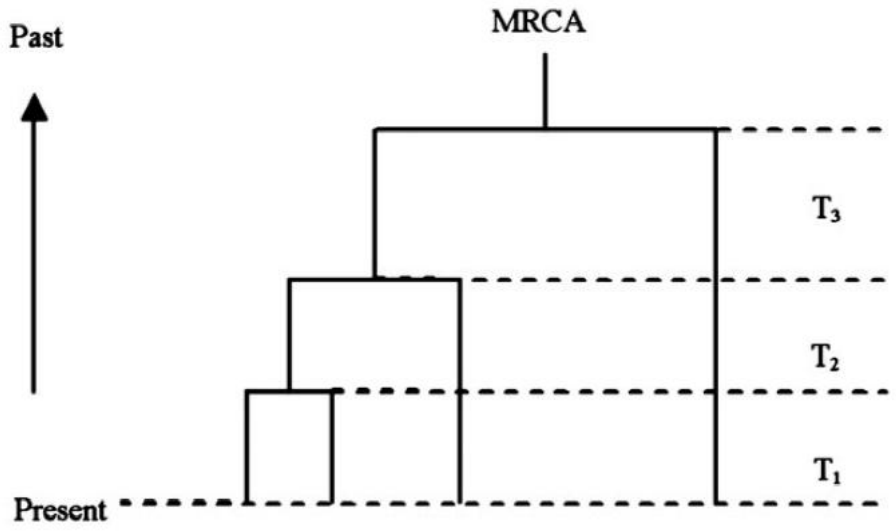


Figure 5. Kingman's *n*-coalescent process [Yuan et al., 2012]

The coalescent process is usually implemented based on the Wright-Fisher (W-F) model (named after Sewell Wright and Ronald Fisher). The W-F model is simply described as follows. Assume that we have a population of *N* diploid individuals and that initially there are *i* copies of allele *a* and *(2N – i)* copies of allele *A* at a particular locus. Thus, the probability of getting *j* copies of allele a in the next generation by randomly sampling from the current population is given by Equation (2) below. Using this equation, we can calculate distributions of allele frequencies in successive generations.

$$\Pr(j|i) = \binom{2N}{j}\left(\frac{i}{2N}\right)^{j}\left(1-\frac{i}{2N}\right)^{2N-j} \quad (2)$$

In the backwards (coalescence) process beginning from the current generation *T*, supposing that the sequence of population sizes is *NT, NT-1, NT-2, NT-3,...*, the (Markovian) transition probability from generation *t* to *t – 1* is as follows [Slatkin, 2001]:

$$\Pr(i_{t-1}|i_r) = \binom{2N_{t-1}}{i_{t-1}}\left(\frac{i_t}{2N_t}\right)^{i_{t-1}}\left(1-\frac{i}{2N_t}\right)^{2N_{t-1}+i_{t-1}} \quad (3)$$

where i_t denotes the copies of allele a at generation $_t$.

Since the coalescent process traces individuals to their MRCA, it is designed to simulate the trajectory (path) H = {iT, iT-1, iT-2, iT-3,…, i2, i1, i0}, where $i_T = i,...,$ $i1 = 1$, and $i_0 = 0$ [Slatkin, 2001]. This trajectory reflects that the simulation starts from the current generation and goes backward in time until the generation where allele *a* is lost. At this point, the individuals in the current generation have been coalesced to a single ancestor (i.e., MRCA), exclusively composed of alleles *A*. The probability of the sample trajectory *H* is given by:

$$\Pr(H) = \prod_{t=1}^{T}\Pr(i_{t-1}|i_t) \quad (4)$$

In the coalescent process, recombination can be flexibly simulated and it is usually implemented under a finite-site model [Hudson, 2002], in which the number of sites (markers) between which recombination can occur is finite. Recombination events can split chromosomes being simulated into a number of segments. Each segment will be modeled by a genealogy tree (i.e., ancestral history) through the coalescent process. Consider a population of *n* chromosomes, with population recombination rate $\rho = 4N_e r$, where *r* is the recombination rate between the ends of the chromosome. Let *R* denote the number of recombination events Figure 5 in the

history of the population. The expectation of R can be expressed as follows [Hudson and Kaplan, 1985]:

$$E(R) = \rho\left(\sum_{i=1}^{n-1}\frac{1}{i}\right) \quad (5)$$

In the permutation process, mutations are usually simulated according to a Poisson process. Each mutation is assumed to occur uniformly and independently on the genealogy tree. Assume the population mutation parameter is $h = 4N_e l$, where $_l$ is the mutation rate for the chromosome being simulated. The expected number of mutations on the *i*-th branch of the genealogy tree is expressed as [Hudson and Kaplan, 1985]:

$$E(M_i) = \theta t_i \quad (6)$$

where t_i is the length of the *i*-th branch.

Theoretically, this simulation scheme gives an excellent framework for population genetics studies in terms of sampling properties and sample statistics. Moreover, this approach is computationally efficient since it only traces the history of the observed sample backward in time. Thus, the coalescent approach is very widely used, with a number of powerful simulators developed under this framework [Kuhner et al., 1998, Ronquist et al., 2003, Stamatakis, 2006].

Approximate Bayesian Computation

Approximate Bayesian computation (ABC) constitutes a class of computational methods, also known as likelihood-free techniques, rooted in Bayesian statistics. A common incarnation of the Bayes theorem relates the conditional probability (or density) of a particular parameter value θ given data D to the probability of D given θ by the rule:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} \quad (7)$$

where $p(\theta|D)$ denotes the posterior, $p(D|\theta)$ the likelihood, $p(\theta)$ the prior, and $p(D)$ the evidence (also referred to as the marginal likelihood or the prior predictive probability of the data). In all model-based statistical inference, the likelihood function is of central importance, since it expresses the probability of the observed data under a particular statistical model, and thus quantifies the support data lend to particular values of parameters and to choices among different models. For simple models, an analytical formula for the likelihood function can typically be derived.

However, for more complex models, an analytical formula might be elusive or the likelihood function might be computationally very costly to evaluate. ABC methods bypass the evaluation of the likelihood function, by comparing the summary statistics parameters of simulated and observed data. In this way, ABC methods widen the realm of models for which statistical inference can be considered.

The ABC methods are mathematically well-founded, but they inevitably make assumptions and approximations whose impact needs to be carefully assessed. Furthermore, the wider application domain of ABC exacerbates the challenges of parameter estimation and model selection. ABC has rapidly gained popularity over the last years and in particular for the analysis of complex problems arising in biological sciences as population genetics, ecology, epidemiology, and systems biology [Leuenberger et al., 2010, Nunes et al., 2010, Marin et al., 2012].

1.5. Databases for human mtDNA

Mitochondrial DNA is inherited maternally [Hutchison et al., 1974], thus does not recombine with paternal molecule [Brown et al., 1979, Michaels et al., 1982], has a fast mutation rate [Piko et al., 1976] and presented by multiple copies in cell [Hagström et al., 2014, Merriwether et al., 1991].

These unique features made mtDNA a versatile tool for evolutionary and population genetics studies in the end of last century, when genome-wide or whole genome population studies were a daydream for researcher.

After the publication of the milestone Cambridge reference sequence by Anderson and colleagues [Anderson et al., 1981] researchers began studying human mtDNA using low resolution restriction fragment length polymorphism analysis with only a few restriction enzymes, utilizing them to estimate very simple phylogenetic trees. After the application of higher-resolution restriction analysis approaches, mitochondrial genome has become popular tool for population geneticists. In 2000th, the development of fully automated sequencing technologies allowed massive sequencing of whole mtDNA genomes. Since then numerous studies have been published where the authors have sequenced hundreds or even several thousand (Fig. 6) [Gunnarsdóttir et al., 2011, Zheng et al., 2012., Behar et al., 2012] human mitogenomes in order to address medical issues [Brandon et al., 2006], different questions of population genetics [Duggan et al., 2014], phylogeography [Derenko et al., 2014], demography [O'Fallon et al., 2011], etc.

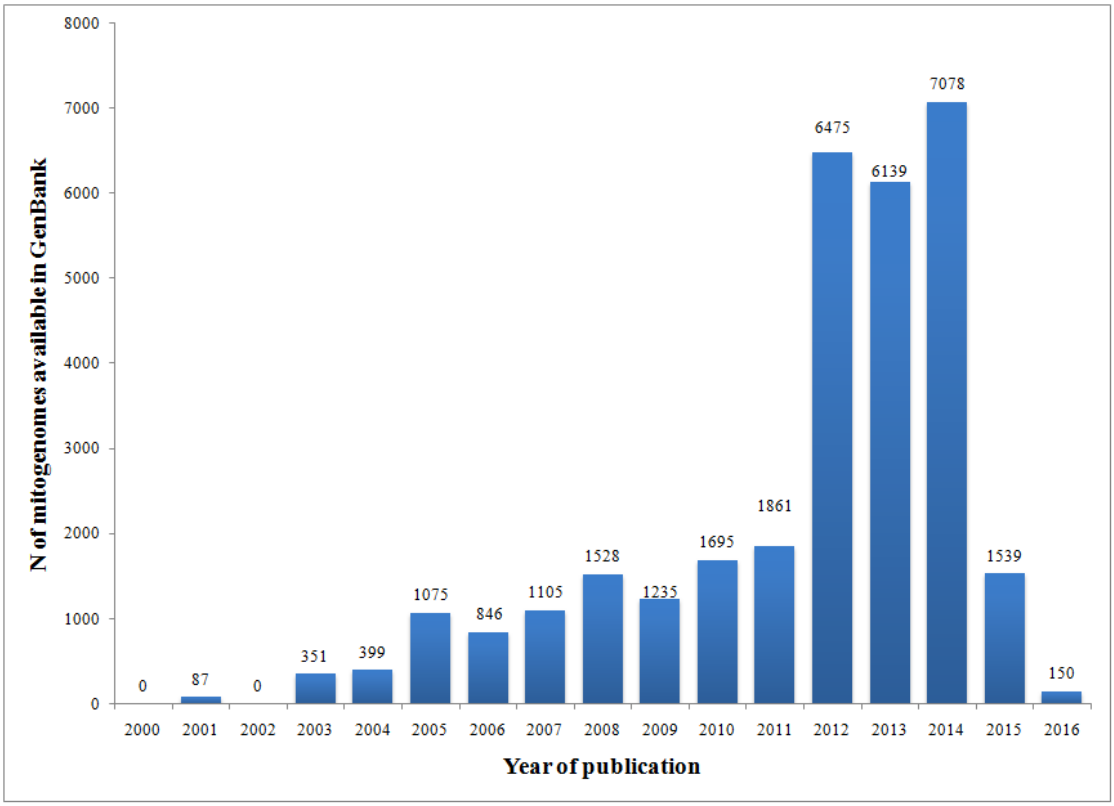


Figure 6. Dynamics of the number of deposited complete or near-complete human mitochondrial genomes in GenBank per year. The search was conducted using the query available in Mitomap (http://www.mitomap.org/bin/view.pl/MITOMAP/MitoSeqs).

According to the most recent Phylotree [van Oven et al., 2009] to the date, February 18th 2016, 24,275 complete human mitogenomes have been sequenced and published in ca 300 papers and projects, defining the current nomenclature of global human mitochondrial DNA variation.

However, from the perspective of data amount, the information on the overall available data of human mitogenomes in .fasta format (ca 520MB) doesn't exceed one entire human haploid genome in the same format. Nevertheless, numerous features of human mitogenomes, such as mtDNA haplogroups, their geographic distribution, population affiliation, ethnicity of individuals screened and other highly relevant characteristics for population genetics studies make it problematic for large-scale mtDNA data management and further analysis. For the purpose of integrating these different types of data, several attempts were made to design human mtDNA-specific databases in the recent two decades, and to date there are several resources that are developed for this aim.

HvrBase

One of the first databases for these specific kind of data was *HvrBase* [Burckhardt et al., 1999] launched in 1999. It contained aligned sequences of the hypervariable regions I and II (HVRI and HVRII) with available information on the individuals (humans or apes) from whom the sequences were obtained. The authors implemented a 'search' function, allowing to retrieve the sequences matching a key-word defined by users, enabling them to query the following types of information: name in publication or GenBank, author and publication, species, population, continent and origin. Additionally, sequences were searchable for certain motifs. The collection of human data at the time of publication of the *HvrBase* comprised 5846 and 2302 HVRI and HVRII sequences, respectively. For 2061 samples, the both of HVR regions were sequenced. *HvrBase* was not updated since 2000, and, moreover, today, when researchers utilize complete mitochondrial genomes for population genetics studies, the data on HVR1 and HVR2 are not widely used anymore. In 2006, by

adding more sequences from primates including humans, the database was updated to *HvrBase++* [Kohl et al., 2006] – the improved and extended version of *HvrBase*. The *HvrBase++* database comprised not only the data on hyper-variable regions but also the mitochondrial genomes and nuclear sequences from several chromosomal loci. Human data were presented by 20,037 sequences. Table 1 displays a human HVRI dataset gathered from 103 publications which encompasses sequences from 89 countries and 220 ethnic groups.

Table 1. Human HVRI datasets over six continents.

Continent	Lineages	N	Number of countries	Populations	Languages
Europe	2033	4358	17	25	31
Africa	1046	1680	25	47	47
North America	824	1581	7	34	9
South America	267	473	7	11	19
Asia	2867	4778	23	102	67
Australia/Oceania	224	473	10	12	28
World	7036	13343	89	220	194

The collection comprised 13 873 HVRI and 4940 HVRII sequences. Additionally, authors included 1376 complete mitochondrial genomes, 205 sequences from X-chromosomal loci and 202 sequences from autosomal chromosomes 1, 8, 11 and 16. In order to reduce the introduction of erroneous data into *HvrBase++*, the authors have developed a procedure that monitored GenBank for new versions of the current data in *HvrBase++* and automatically updated the collection. For the stored sequences, supplementary information on the donors, such as geographic origin, population affiliation and language could be retrieved. As a new key feature, *HvrBase++* provided an interactive graphical tool to easily access data from dynamically created geographical maps. Now the HvrBase++ is available through

http://hvrbase.cibiv.univie.ac.at, though it is outdated and has several bugs that do not allow for the appropriate and efficient usage of the database.

mtDB

Following HvrBase++, the database mtDB was launched in 2005 [Ingman et al., 2006] and provided users with complete mitogenomic data on population scales; also including the data for medical relevancy. By the 1 March 2007, there were 2697 complete mitogenomes from different geographic regions taken from 33 published papers (Table 2). mtDB had four main types of functionality. First, it allowed downloading of all mtDNA sequences either by individual molecules or by sets of population. All data were grouped into ten major geographic regions based on the population affiliation of the individuals. Data from the same populations were also downloadable as batches of individual files. All sequences were linked to the original papers where they were first published and to the corresponding GenBank accession numbers.

Table 2. Number (N) of complete human mitogenomes obtained from corresponding geographic regions

Geographic region	**N**
Africa	287
Middle East	45
Europe	1192
North America	11
South America	14
Asia	922
South Asia	125
Australia	32
Melanesia/Micronesia/ Polynesia	69
Total	**2697**

Secondly, mtDB had special search function for locating polymorphic sites. At the moment of mtDB publication, 3311 polymorphic loci were identified and tabulated. The table comprised a separate line for any polymorphic site with a number of sequences that contain each particular variant at that site, location of the locus comparing to rCRS, codon number and position, and the mutational shift in the amino acid (aa). Clicking on the number of a particular variant, users are able to obtain a list of the sequences containing that particular mutation (insertions relative to CRS were discarded) and then consequently download them.

Ultimately, to search the data on mtDNA haplogroups of interest, users are provided with the phylogenetic tree with illustrated clickable haplogroups. When clicking, the database shows all the sequences matching the haplogroups with the further ability to adjust the query according to the sequence length and geographic region.

Finally, a 'search' function for mitochondrial haplotypes is implemented in the mtDB. Users can search up to 10 loci simultaneously by specifying the ‘position’ and ‘nucleotide’. Again, those sequences could then be downloaded from the database.

mtDB was outdated since March 1, 2007 and the lack of convenient functionality of sequence retrieval and data parsing limited the usage of this database.

HmtDB

The database was created in 2005 [Attimonelli et al., 2005] and relaunched in January 2012 [Rubino et al., 2012]. To date, the *HmtDB* is one of the most convenient databases of human mtDNA: it is regularly updated and has numerous functional features, which make it a powerful tool for population-based mitogenomic studies. Moreover, by including the mtDNA data from patients with different disorders, the database provides users with the information that has direct medical relevance. *HmtDB* stores mitogenomic data annotated with the population and variability information (Table 3 and 4, respectively).

Table 3. Total amount (N) of mitogenomic data in HmtDB.

Continent name	Individuals	Number of genomes	Complete genomes	Coding regions only
Africa	Normal	3091	2943	148
	Patient	71	71	0
America	Normal	2336	2231	105
	Patient	26	24	2
Asia	Normal	6385	6324	61
	Patient	1116	1116	0
Europe	Normal	7205	6711	494
	Patient	1708	1571	137
Oceania	Normal	1537	1523	14
	Patient	0	0	0
Undefined	Normal	6818	6716	102
Continent	Patient	634	631	3
All continents	**Normal**	**27372**	**26448**	**924**
	Patient	**3555**	**3413**	**142**

Table 4. Number (N) of variable site in HmtDB

Continent name	N
Africa	3505
America	3461
Asia	5058
Europe	4473
Oceania	1554
All continents	**7694**

The annotations of sequences are curated manually, giving the data a higher degree of accuracy. The authors have designed a '*Classify'* tool that allows the

database to predict the haplogroups based on *Phylotree* for all mitogenomes stored in the database or even for new sequences provided by users.

HmtDB provides three main categories of usage. First, users can browse the database by a multi-criterion query system. The possibilities for parsing the data include searching by unique identifiers (*GenBank* accession numbers and internal *HmtDB* identifiers), by continents and countries, by haplogroups, sex and age of the individuals, disease types and others (Fig.7), as well as by combinations of queries.

Second, users can analyze their own mitogenomes using the '*Classify'* tool (for complete genomes) or by the '*Fragment-classifier*' tool (for partial sequences). These tools perform predictions and assignments of haplogroups to complete or partial mitogenomes. The result includes an output list of most likely haplogroups, a table with their mutated positions, numerous statistics of the variable sites, and possible disease associations of particular variants with different disorders.

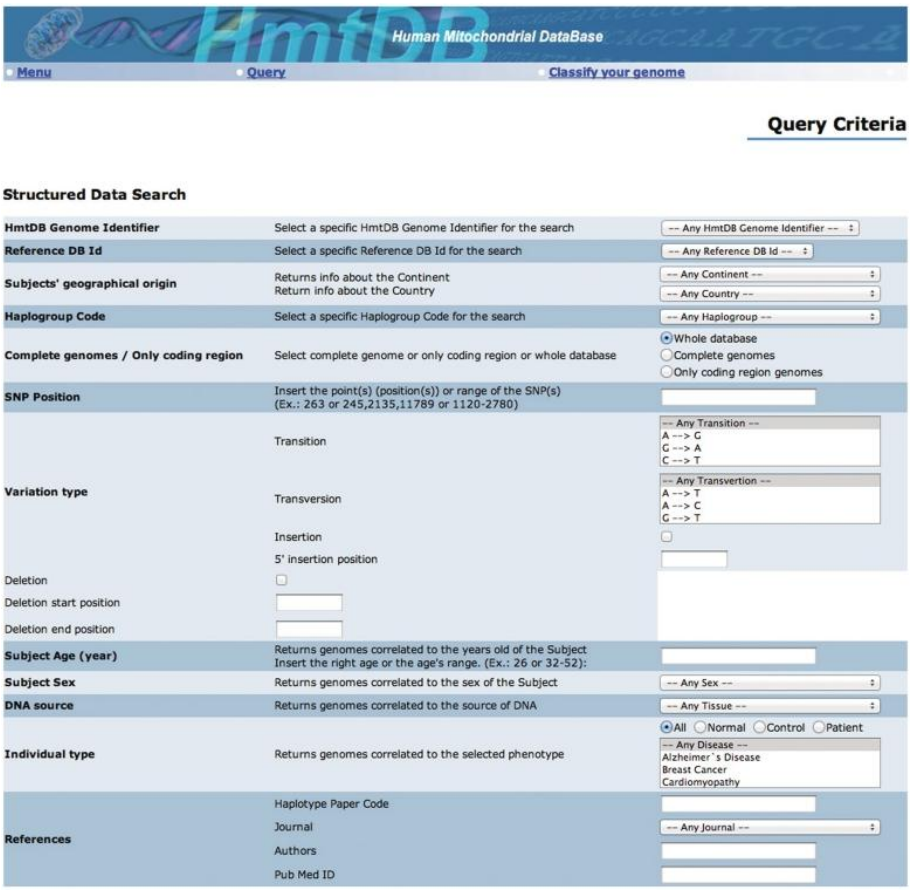


Figure 7. Multi-criterion query system of HmtDB

However, the computational performance of the '*Classify'* tool allows a simultaneous analysis of only a few mitogenomes (it takes ca 17 minutes to assign haplogroups for 10 complete mitogenomes), while for multiple genomes, the '*Classify'* tool lags behind other haplogroups assignment tools like *Haplofind* [Vianello et al., 2013] (ca 70 seconds for 100 mitogenomes).

Third, users can download sequence alignments with reference genomes as well as variability data.

Mitomap

The very first database for human mtDNA was *Mitomap*, which initially appeared in print form in the paper of Kogelnik et al. [1995]. Then, in 1996, it was developed into an online database, containing published human mtDNA variation along with geographic and disease specific variants. Today, *Mitomap* is a manually curated, regularly updated and functionally rich resource of high-quality human mtDNA data for researchers, clinicians and genetic counselors.

Mitomap has three main categories for usage. First, it contains background information about the human mitochondrial DNA. This section might be a very useful resource for researchers who begin mtDNA studies, since it has comprehensive information about the numerous characteristics of human mitogenomes, such as general representation of mtDNA, *aa* translation table, haplogroups and their frequencies, major rearrangements, illustrations of mtDNA and many others. Additionally, users within this database can find the information on other mtDNA-specific databases, tools and several useful resources. Moreover, the section provides users with up-to-date information (comprising ca 5800 references) on papers published since 1997 where mtDNA was used. These features make Mitomap an extremely important resource for mitogenomic studies, especially for beginners.

Second, *Mitomap* stores the annotated listing of mtDNA variants from both healthy individuals and patients; From the database users can comb the data for either the control or coding region variants. The frequencies of the variants are calculated

from the set of 30589 complete or near complete human mitogenomes retrieved for *GenBank*. The web-interface of the *Mitomap* variant search function is shown on Fig. 8. The users can retrieve the information on the loci, nucleotide change, codon position and number, *aa* change, number of *GenBank* records where the particular variant is found and relevant references. Additionally, the end-users are allowed to download the data in different file formats.

Third, the *Mitomap* team has developed the *Mitomaster* analysis tool and currently provides the Application Programming Interface for it. The main function of *Mitomaster* is identification of polymorphic positions, calculation of variant statistics (positions, GB frequencies, references, etc.) and assignment of haplogroups to complete or partial mitogenomes (Fig. 9).

Show 10 entries Locus MT-ATP8

Position	Locus	Nucleotide Change	Codon number	Codon Position	Amino Acid Change
8369	MT-ATP8	C-A	2	1	non-syn:P-T
8369	MT-ATP8	C-T	2	1	non-syn:P-S
8371	MT-ATP8	C-T	2	3	syn:P-P
8374	MT-ATP8	A-G	3	3	syn:Q-Q
8375	MT-ATP8	C-T	4	1	syn:L-L
8376	MT-ATP8	T-C	4	2	non-syn:L-P
8379	MT-ATP8	A-G	5	2	non-syn:N-S
8380	MT-ATP8	T-C	5	3	syn:N-N
8381	MT-ATP8	A-G	6	1	non-syn:T-A
8382	MT-ATP8	C-T	6	2	non-syn:T-I

Showing 1 to 10 of 181 entries (filtered from 9,228 total entries)

Previous

Figure 8. Mitomap variant search, querying variants of ATP8 gene (10 polymorphic positions are shown).

The query might be performed using sequences, *GenBank* identifiers or single nucleotide variants or SNVs. It should be noted that Mitomaster performs relatively fast; as it takes ca 74 seconds to analyze 10 complete mitogenomes.

Other human mtDNA databases and database-like resources, such as *Phylotree* [van Oven et al, 2009], *Empop* [Parson et al., 2007], *MitoAge* [Toren et al., 2015], *mtDNA community* [Behar et al., 2012] and *MamMiBase* are not included in our review since some of them are currently unavailable (*MamMiBase)*, others are essentially just data repositories, rather than functional databases (*Phylotree and mtDNA community*) and the rest are not designed specifically for population genetics purposes (*Empop* and *MitoAge*).

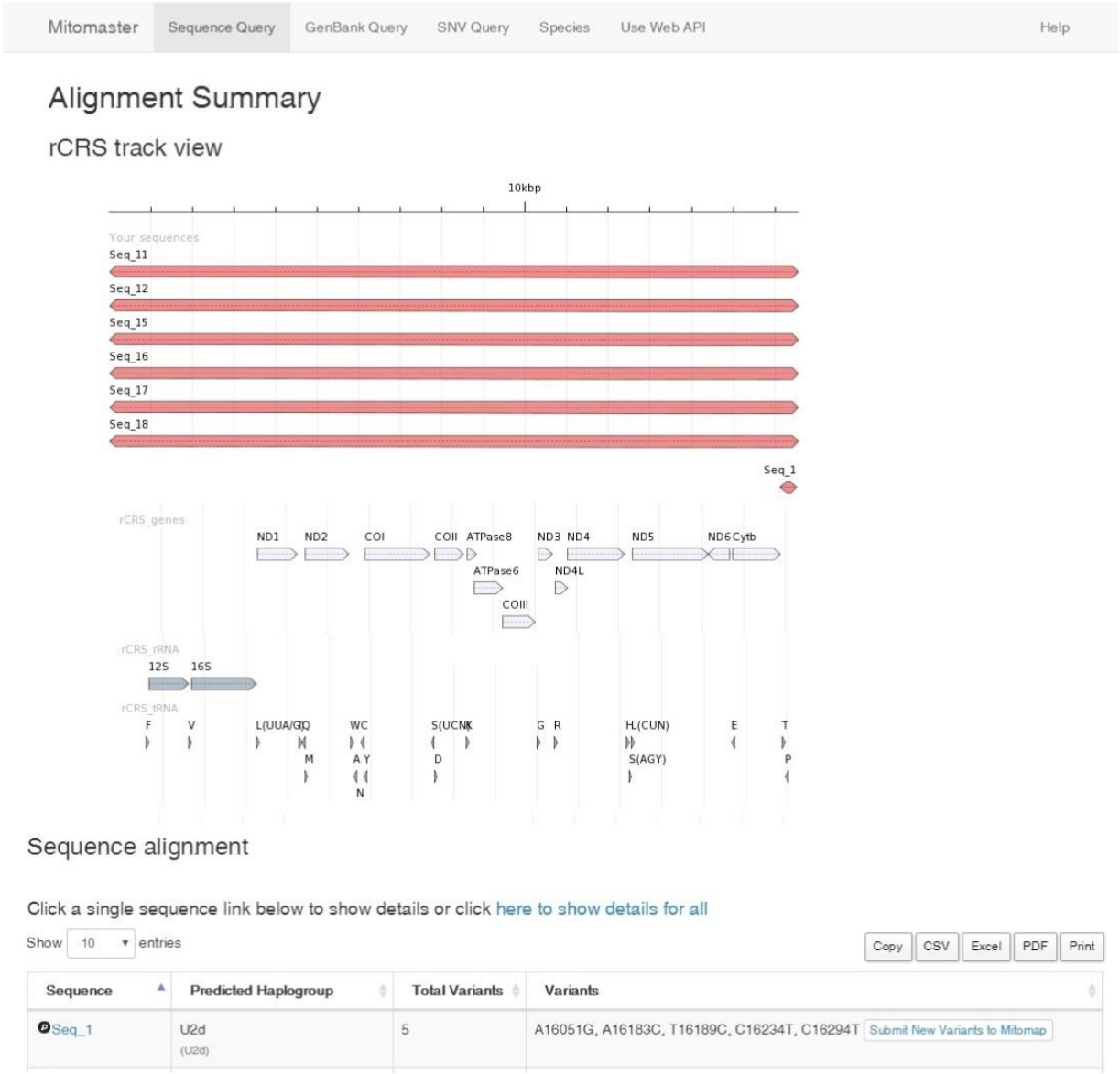


Figure 9. Mitomaster's output result for six complete mitogenomes and one HVRI

Summarizing, we have reviewed the main population genetics-oriented databases for human mitochondrial DNA storage and management.

The *HvrBase*++ and *mtDB* databases were launched more than ten years ago but have not been updated since, while the number of new mtDNA partial and complete sequences has increased substantially. Besides being outdated, the *mtDB* did not have appropriate functional characteristics for managing mtDNA data. For example, a lack of the data on human mtDNA haplogroup and sorting/grouping functions were restricting the usage of *mtDB*. Additionally, when the current next-generation technologies allow massive sequencing of entire mitochondrial genomes on large population scales, the data on HVRI and HVRII of mtDNA available in these aforementioned databases fail to fulfill the requirements of modern population genetics.

On the other hand, the *HmtDB and Mitomap* databases are regularly updated and have numerous options for complex data searching – a powerful querying system, where user can search data according to the mutated position, haplogroup, geographic region, tissue, sex, etc, and contain mtDNA haplogroup assignment tools and convenient downloading functions.

Nevertheless, there still exist some functional characteristics that are not implemented in the mentioned databases that would significantly enhance the efficacy of their usage. For instance, the possibility of downloading the information in different file formats would be a useful feature for databases, as numerous programs that analyze sequence data may require specific data formats such as *fasta*, *nexus*, *phylip*, etc. Moreover, since mtDNA is extensively used in human evolutionary studies, it might also be useful to implement algorithms permitting the partitioning of multiple alignments into segments, which would allow the assignment of specific mutation rates to corresponding sites of the molecule. Additionally, this will enable researchers to study different regions of the molecule, such as protein coding genes, rRNAs, tRNAs, etc. on the population scale, thus creating additional possibilities in human mtDNA research. One of the major points that should be

taken into account by mtDNA database developers is the growing number of ancient mtDNA samples, which have peculiar characteristics and are highly important for human population genetics studies.

CHAPTER II. MATERIAL AND METHODS

2.1 Sample collection, data generation and comparative datasets collection

In our work, we have used two types of mtDNA data – modern and ancient.

Modern data were presented by several large datasets covering almost the whole area of historical Armenia. These datasets comprise unpublished and published mtDNA data (HVRI and complete mitogenomes) from both general Armenian population and Armenians from distinct geographic regions within the Armenian Highland. The spatial distribution of unpublished complete mitogenomes and HVRI samples is presented in Fig. 10 and 11, respectively.

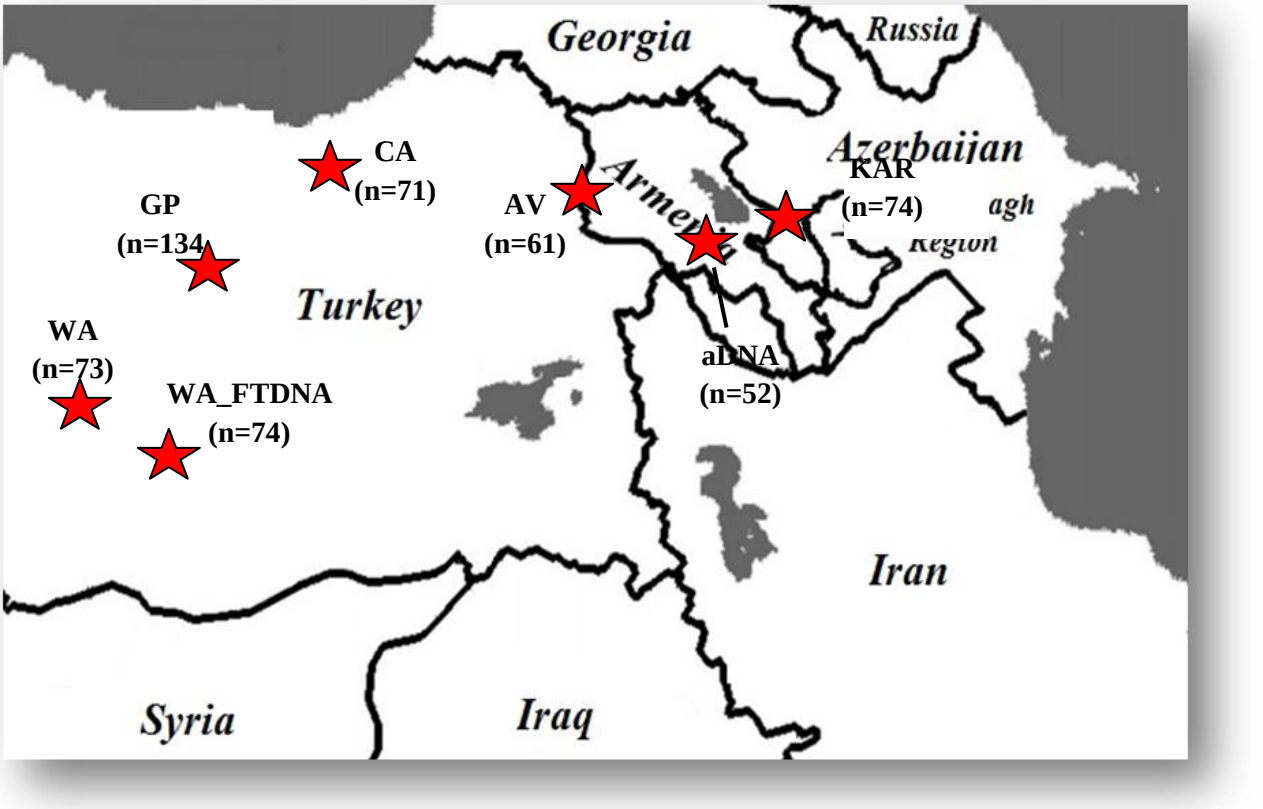


Figure 10. Geographic distribution of Armenian complete mitogenomes which were collected, sequenced and analyzed in frames of this study for the first time. Red stars show to geographic locations of obtained samples. Names of populations corresponding to the abbreviations are given in the text below. Numbers in brackets indicate sample sizes.

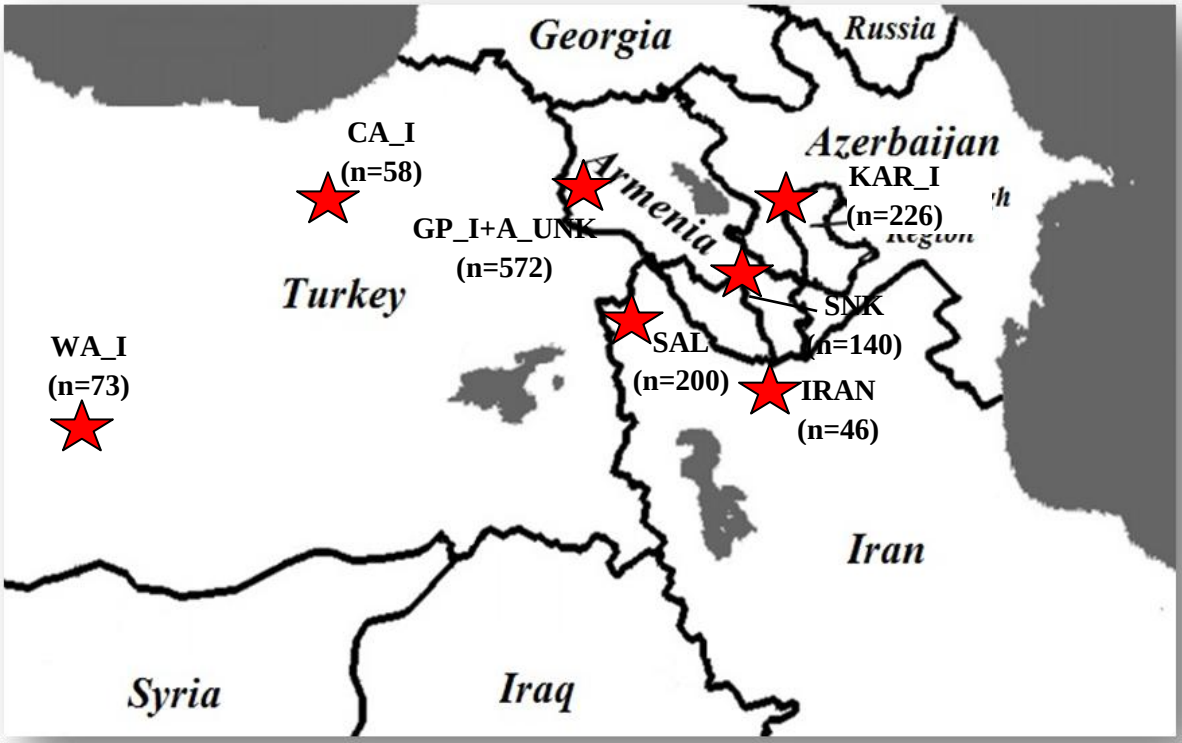


Figure 11. Geographic distribution of Armenian HVRI sequences which were collected, genotyped and analyzed in frames of this study for the first time. Red stars show to geographic locations of obtained samples. Names of populations corresponding to the abbreviations are given in the text below. Numbers in brackets indicate sample sizes.

All individuals that were subjected to DNA sampling represented geographic regions corresponding to the birthplace of their grandmaternal ancestors. All subjects provided written informed consent for the collection of samples and subsequent analysis.

To collect and subsequently analyze the data on mitogenomic variability of Armenians from different geographic regions, in 2012-2014 the staff of the Laboratory of Ethnogenomics (Institute of Molecular Biology, NAS RA) has organized several trips within Armenia and collected saliva samples of 289 individuals representing 4 Armenian regional groups – Armenians from (i) Karabakh (Kar, n=74), (ii) Ararat Valley (AV, n=61), (iii) Central Armenia (CA, n=71), (iv) Western Armenia (WA, n=73).

2 ml of saliva was collected from each individual in 15 ml tubes with 2 ml of DNA preservative buffer, consisting of 0.5% sodium dodecyl sulphate and 0.05 M EDTA.

All these samples were subsequently sent to the collaborating group of Dr. Boris Malyarchuk (Institute of Biological Problems of the North, Russian Academy of Sciences, Magadan, Russia) for complete mitochondrial DNA sequencing, which was performed by standard protocols as described elsewhere [Torroni et al., 2001].

Additionally as a general Armenian population (GP) and Western Armenians (WA_FTDNA) we have used 134 and 74 complete mitochondrial genomes, respectively, from ethnical Armenians whose data were available in the Family Tree DNA public database (https://www.familytreedna.com).

It should be mentioned that this sample sets mostly describe population of western part of historical Armenia, as soon as the most part of Armenian DNA Project contributors are descendants from Western Armenia.

For the HVRI data, general Armenian (GP_I) population was represented by 400 unrelated individuals, whose ancestors inhabited different regions of Armenian Highland at least on the grandparental level. All these samples were collected in 2009-2010 from different areas of Republic of Armenia by our collaborators from Estonian Biocentre (Tartu, Estonia) and were genotyped for HVRI region by standard restriction fragment length polymorphism (RFLP) protocols described elsewhere [Macaulay et al, 1999] at the Estonian Biocentre. Additionally, 5 out of 400 samples were sequenced for complete mitochondrial genome at the same laboratory.

We have also used unpublished mtDNA HVRI sequences of 711 ethnical Armenians. All those samples, except those of representing Salmast region, were collected in 1999 by Prof. Levon Yepiskoposyan in different geographic regions of the Republic of Armenia; mtDNA HVRI RFLP genotyping of these samples was performed at the Centre for Genetic Anthropology, University College London in 2004 using standard protocols [Macaulay et al, 1999].

The samples represented the populations of Syunik (SNK, n=140), Central Armenia (CA_I, n=58), Iranian Armenians (Iran, n=46), Armenians from Karabakh (KAR_I, n=226), Western Armenians (WA_I, n=65). Other 176 individuals, who were genotyped for HVRI did not specify the region of maternal ancestry (A_UNK).

Additionally, we have utilized 200 Armenian samples representing the region of Salmast, which were collected from the descendants of Armenians from the region Khoy/Salmast (currently in the north-west of Iran) in 2010 and were genotyped for HVRI in 2010 at the Arizona University, Tucson, USA.

We have also used published mitogenomic data on Armenians taken from Schonberg et al., 2011 (n=30, sampling was performed in Yerevan) and Derenko et al., 2013 (n=10, samples represent Iranian Armenians).

To investigate mitochondrial DNA variation in ancient populations of Armenian Highland in 2014-2015 our team in collaboration with the Institute of Archaeology and Ethnography of National Academy of Sciences (Yerevan, Armenia) has collected bone and teeth specimens from burials located in the territory of the Republic of Armenian and Artsakh. The collection of ancient specimens represented different archaeological epochs spanning from Middle/Upper Paleolithic (20kay) to the late Bronze age (3kya).

Ca 90 collected specimens were subsequently delivered to Willerslev's group at the Center for GeoGenetics (Copenhagen, Denmark) for complete mitochondrial DNA sequencing. All laboratory techniques for aDNA analysis, including DNA extraction and library preparation, molecular screening, next-generation sequencing (using Illumina HiSeq2500 platform) and DNA authentification were performed as described in Allentoft et al. 2015.

Consensus sequence calling based on both rCRS and whole genome was carried out using in-home python script.

44 out 90 sequences had the coverage > 9 and thus were subsequently considered for further analysis.

Additionally, we have used 8 published ancient mitogenomes from specimens collected from the territory of Armenia which were reported in Allentoft et al., 2015.

Summarizing, in our work we have utilized 284 unpublished modern Armenian mitogenomes, representing 4 distinct geographic regions of Historical Armenia, 52 ancient Armenian mitogenomes, of which 8 were previously published, 208 published modern Armenian mitochondrial genomes and 1307 unpublished HVRI Armenian sequences, 734 of which representing 5 other Armenian regional groups and 572 describing general Armenian population.

Thus in total, in our project we have described and analyzed 1851 mtDNA samples of ethnic Armenians, which makes our population one of the most studied ethnic group so far from the mitochondrial DNA perspective.

To compare the Armenian mitochondrial genetic profile based on mtDNA haplogroup frequencies with those of other populations we have collected a large number of comparative datasets, representing ethnic groups from all over the world. This comprehensively and thoroughly collected database comprises ca 23 thousand samples representing 113 populations worldwide. More detailed information on comparative datasets is shown in Table 5.

Table 5. Comparative datasets used for haplogroup frequency analysis.

Reference	**Region**	**Population**	**Abbreviation**	**N**
Achilli et al., 2007	Europe	Italy (North)	Ita_N	346
Achilli et al., 2007, Alvarez-Iglesias et al., 2009	Europe	Spain	Spa	215
Al-Zahery et al., 2011	Middle East	Iraqis	Iraq	176
Al-Zahery et al., 2011	Middle East	Marsh Arab	Mar	145
Behar et al., 2006	Europe	Ashkenazi Jews	Ash	583
Behar et al., 2008	Africa	Ethiopian Jews	Eth_Jews	29
Behar et al., 2008	Africa	Moroccan Jews	Mor_Jews	149

Reference	Region	Population	Abbreviation	N
Behar et al., 2008	Caucasus	Azerbaijani Jews	Aze_Jews	58
Behar et al., 2008	Caucasus	Georgian Jews	Geo_Jews	74
Behar et al., 2008	Europe	Sephardi Jews	Seph_Jews	194
Behar et al., 2008	Middle East	Bedouin	Bed	58
Behar et al., 2008	Middle East	Iranian Jews	Iran_Jews	82
Behar et al., 2008	Middle East	Iraqi Jews	Iraq_Jews	135
Behar et al., 2008	Middle East	Palestinian	Palest	110
Behar et al., 2008	Middle East	Yemenite Jews	Yem_Jews	119
Behar et al., 2008	Southern Asia	Cochini Jews	Coch_Jews	45
Behar et al., 2008	Southern Asia	Mumbai (Bene Israel)	Mum	34
Behar et al., 2010	Africa	Moroccans (Berbers)	Mor_Berb	319
Behar et al., 2010	Africa	Morocco	Mor	363
Behar et al., 2010	Caucasus	Lezgin	Lez	46
Behar et al., 2010	Europe	Belorussia	Bel	275
Behar et al., 2010	Europe	Chuvash	Chu	55
Behar et al., 2010	Europe	Romania	Rom	346
Behar et al., 2010	Middle East	Cyprus	Cyp	183
Behar et al., 2010	Middle East	Egypt	Egy	188
Behar et al., 2010	Middle East	Jordan	Jor	198
Behar et al., 2010	Middle East	Lebanon	Leb	168
Behar et al., 2010	Middle East	Syria	Syr	215
Behar et al., 2010	Southern Asia	India (South)	Ind_S	183
Behar et al., 2010, Abu-Amero et al., 2007	Middle East	Saudi Arabia	S_Ar	319

Reference	Region	Population	Abbreviation	N
Behar et al., 2010, Kasperaviciute et al., 2004	Europe	Lithuania	Lit	216
Bekada et al., 2013	Africa	Algeria	Alg	239
Delfin et al., 2014	South_Eastern Asia	Philippines	Phil	357
Derenko et al., 2007	Europe	Kalmyks	Kal	110
Derenko et al., 2007	Middle East	Kurds	Kur	25
Derenko et al., 2007	Siberia	Altaians-Kizhi	Alt	90
Derenko et al., 2007	Siberia	Buryats	Bur	295
Derenko et al., 2007	Siberia	East Evenks	E_Evenk	45
Derenko et al., 2007	Siberia	Khakassians	Khakas	57
Derenko et al., 2007	Siberia	Khamnigans	Khamn	99
Derenko et al., 2007	Siberia	Shors	Sho	82
Derenko et al., 2007	Siberia	Telenghits	Teleng	71
Derenko et al., 2007	Siberia	Teleuts	Teleut	53
Derenko et al., 2007	Siberia	Tuvinians	Tuv	105
Derenko et al., 2007	Siberia	West Evenks	W_Evenk	73
Derenko et al., 2007, Irwin et al., 2010	Central Asia	Tajikistan	Taj	287
Derenko et al., 2007, Pakendorf et al., 2003	Siberia	Yakuts	Yak	153
Derenko et al., 2013	Middle East	Persians	Pers	181
Derenko et al., 2013	Middle East	Qashqai	Qas	112
Fadhlaoui-Zid et al., 2011	Africa	Lybia	Lyb	269
Finilla et al., 2001	Europe	Finns	Fin	192
Gunnarsdóttir et al., 2011b	South_Eastern	Sumatra	Sum	72

Reference	Region	Population	Abbreviation	N
	Asia			
Irwin et al., 2007, Updated from Tambets et al., 2004	Europe	Hungarians	Hun	323
Irwin et al., 2008	Eastern Asia	Vietnam	Vie	174
Irwin et al., 2008	Europe	Greek Cypriots	Gre	86
Irwin et al., 2008	Europe	Northern_Greeks	N_Greeks	318
Irwin et al., 2010	Central Asia	Afghanistan	Afg	98
Irwin et al., 2010	Central Asia	Kazakhstan	Kaz	255
Irwin et al., 2010	Central Asia	Kyrgyzstan	Kyr	247
Irwin et al., 2010	Central Asia	Turkmenistan	Turkm	249
Irwin et al., 2010,Behar et al., 2010, Quintana-Murci et al., 2004	Central Asia	Uzbeks	Uzb	578
Karachanak et al., 2012	Europe	Bulgarians	Bul	855
Kivisild et al., 2004	Africa	Ethiopia	Eth	270
Kivisild et al., 2004	Middle East	Yemen	Yem	118
Kong et al., 2003	Eastern Asia	Daur	Dau	45
Kong et al., 2003	Eastern Asia	Oroqen	Oro	44
Kong et al., 2003, Derenko et al., 2007	Eastern Asia	Mongol	Mon	95
Malyarchuk et al., 2003	Europe	Czech	Cze	179
Malyarchuk et al., 2008	Europe	Slovaks	Slo	207
Malyarchuk et al., 2010	Europe	Tatars	Tat	73
Malyarchuk et.a. 2002	Europe	Russia	Rus	198
Mielnik-Sikorska et al, 2013	Europe	Poland	Pol	404
Mielnik-Sikorska et al, 2013	Europe	Ukraine	Ukr	159

Reference	Region	Population	Abbreviation	N
Mikkelsen et al., 2012	Africa	Somalie	Som	190
Pala et al., 2009	Europe	Italy (Tuscany)	Ita_T	322
Pliss et al., 2006	Europe	Latvians	Lat	299
Poetsch et al., 2003	Europe	Germans	Ger	287
Quintana-Murci et al., 2004	Southern Asia	S_Pakistan	S_Pak	133
Rakha et al., 2010, Quintana-Murci et al., 2004	Southern Asia	N_Pakistan	N_Pak	318
Richard et al., 2007	Europe	France	Fra	868
Richard et al., 2007	Europe	French Basque	Fre_Basque	80
Scheible et al., 2014, Derenko et al., 2007	Eastern Asia	Koreans	Kor	384
Schonberg et al., 2011	Caucasus	Azeris	Aze	30
Shlush et al., 2008	Middle East	Druze	Dru	311
Soares et al., 2008, Soares et al., 2011	South_Eastern Asia	Indonesia	Indo	41
Stoljarova et al., 2016	Europe	Estonia	Est	114
Tanaka et al., 2004	Eastern Asia	Japan	Jap	1312
Tillmar et al., 2010	Europe	Sweds	Swe	282
Trejaut et al., 2005	South_Eastern Asia	Taiwan_aboriginal	Taiwan	640
Updated from Tambets et al., 2000, Schonberg et al., 2011	Middle East	Turkey	Tur	412
Wen et al., 2004	Eastern Asia	Naxi	Nax	45
Wen et al., 2004	Eastern Asia	North Han	N_Han	226

Reference	Region	Population	Abbreviation	N
Wen et al., 2004	Eastern Asia	South Han	S_Han	614
Wen et al., 2004	Eastern Asia	Tu	Tu	41
Wen et al., 2004	Eastern Asia	Tujia	Tuj	94
Wen et al., 2004, Qian et al., 2001	Eastern Asia	Lahu	Lah	82
Wen et al., 2005	Eastern Asia	Miaozu	Mia	142
Yao et al., 2004	Eastern Asia	Uygur	Uyg	47
Yunusbayev et al., 2013	Caucasus	Abazins	Aba	105
Yunusbayev et al., 2013	Caucasus	Abkhazians	Abk	136
Yunusbayev et al., 2013	Caucasus	Adyghey	Ady	155
Yunusbayev et al., 2013	Caucasus	Balkars	Bal	140
Yunusbayev et al., 2013	Caucasus	Chechens	Che	176
Yunusbayev et al., 2013	Caucasus	Cherkessians	Cherk	123
Yunusbayev et al., 2013	Caucasus	Dargins	Dar	110
Yunusbayev et al., 2013	Caucasus	Ingush	Ing	103
Yunusbayev et al., 2013	Caucasus	Kabardin	Kab	150
Yunusbayev et al., 2013	Caucasus	Kara Nogays	Kara_Nog	130
Yunusbayev et al., 2013	Caucasus	Karachays	Karach	106
Yunusbayev et al., 2013	Caucasus	Kuban Nogays	Kub	131
Yunusbayev et al., 2013	Caucasus	N. Ossetians	N_Oss	138
Yunusbayev et al., 2013, Schonberg et al., 2011	Caucasus	Georgians	Geo	104
Zimmermann et al., 2009	South_Eastern Asia	Thailand	Tha	190
TOTAL	**9**	**113**		**22897**

To perform different types of comparative analysis between Armenians and aforementioned datasets based on haplogroup frequencies, we have fitted the haplogroup resolution of all samples to 40 major mtDNA lineages – A, B, C, D, F, G, H, HV, HV0&V, HV1, HV2, I, J, K, L, M, M1, N, N1a, N1b, N1c, N1d, R, R0a, T1, T*(xT1), U, U1, U2, U3, U4, U5, U6, U7, U8, U9, W, X, Y and Z.

2.2 Data analysis and bioinformatics

The haplogroups for the complete mtDNA genomes were defined using Haplofind online software [Vianello et al., 2013], which assigns the haplogroups based on the last version Phylotree [van Oven, 2015] using the Reconstructed Sapience Reference Sequence [Behar et al., 2012] as a reference sequence.

To assess haplogroups for the samples genotyped for HVRI region only, we have converted mtDNA haplotypes obtained by RFLP into mtDNA HVRI sequences from 16024-16400 b.p. positions by Haplosearch – the online tool for mitochondrial haplotype-sequence two-ways transformation using population genetics nomenclature [Fregel et al., 2011]. To assign haplogroups further we have processed obtained HVRI sequences with several software – mitotool [Fan et al., 2013], mitohap (https://dna.jameslick.com/mthap/), haplogrep [Kloss-Brandstätter et al., 2011] and EMMA. Samples with ambiguously defined haplogroups were discarded from haplogroup frequency based analyses.

The frequency analysis of haplogroup distribution of studied populations was made via MS Excel. The data of mtDNA haplogroup frequencies was further used for the construction of distance matrixes based on different genetic distances (GD).

F_{ST} pairwise GDs, given by (8)

$$F_{ST} = \frac{\theta_S^{\,2}}{\theta_T^{\,2}} \qquad (8),$$

where θ_S^2 is the variance in the frequency of alleles in different subpopulations and θ_T^2 is the variance of allele frequencies in the total population, were estimated using Arlequin v.3.5 software [Excoffier et al., 2010].

For measuring the statistical significance of GDs between studied populations we used Exact test [Raymond et al., 1995] of population differentiation with 1000 steps of permutations. The values p<0,05 were considered as statistically significant. Other GDs (Nei's D and Reinolds's D^2) were calculated by PHYLIP package [Felsenstein, 1989] and were given by formulas 9 and 10, respectively:

$$D = -\ln\left[\frac{\sum_m \sum_i p_1 m_i \, p_2 m_i}{\left|\sum_m \sum_i p_1 m_i\right| \left|\sum_m \sum_i p_2 m_i\right|}\right] \qquad (9)$$

$$D^2 = \frac{\sum_m \sum_i \left(p_1 m_i - p_2 m_i\right)^2}{2\sum_m \left(1 - \sum_i p_1 m_i \, p_2 m_i\right)} \qquad (10),$$

where m is summed over loci, i over alleles at the m-th locus, and where $p_1 m_i$ is the frequency of the i-th allele at the m-th locus in the populations 1 and 2.

Gene-E software was used for the visualization of obtained distance matrices [http://www.broadinstitute.org/cancer/software/GENE-E].

Principal Coordinate Analysis (PCA) was performed on similarity matrices calculated as one minus genetic distance using Genstat software [Trust, 1995]. The variables used for analysis of general Armenian population were the frequency values of 40 major mtDNA haplogroups, mentioned above. Values along the main diagonal, representing the similarity of each population sample to itself, were calculated from the estimated GD between two copies of the same population.

In order to assess the contribution of each haplogroup on the pattern obtained by the PCA, we have also performed Correspondence analysis (CorA) using SPSS ver. 19 software package [Brosius, 2011].

Plots generated by PCA and CorA were visualized using in-home R [R core team, 2013] script based on ggplot2 package [Wickham, 2009], available through our github page https://github.com/GrantHov/My_R_codes.

The mtDNA complete and partial sequences were aligned against rCRS by multiple sequence alignment tool MAFFT using standard parameters [Katoh et al., 2013].

The reconstruction of most-parsimonious trees of the complete mtDNA sequences, which is a key method for investigating the phylogeography of mitogenomic variability of human populations, was made by mtPhyl software [Eltsov et al., 2011, http://eltsov.org]. To construct multiple input files for mtPhyl we have used bash scripting [Free Software Foundation (2007). Bash (3.2.48) [Unix shell program]].

To assess the dynamics of demographic changes that took place during the history of the Armenian populations, we have applied the Bayesian Skyline plot (BSP) method [Drummond et al., 2005] implemented in BEAST [Drummond et al., 2007], which shows the dynamics of effective population size (Ne) changes of the studied population throughout the time.

For performing this type of analysis we have partitioned the mitogenomic alignments into several regions depending on the mutation rates, which are shown to be different for different regions of human mtDNA [Rieux et al., 2014]. This partitioning was done by in-home script written in Python programming language (available through github https://github.com/GrantHov/My_Python_codes). The partitioning scheme was in accordance with recently published study of improved human mitochondrial DNA clock calibration [Rieux et al., 2014] – the first and second nucleotides in codons of protein coding genes (PC1+PC2), the third nucleotides in codons of protein coding genes (PC3), rRNAs+rRNAs, HVRI+HVRI.

For generating input file for BEAST v. 1.8.1 we have used Beauti software [Drummond et al., 2012]. For all four partitions we have linked generating trees, unlinked clock models and substitution models as soon as those two parameters were different for each partition. In this analysis we have used our ancient mitogenomes, which allow much more accurate coalescence of generated trees, thus highly improving the likelihood of obtained models of population dynamics. All substitution models were set to HKY with Invariant Sites for site heterogeneity model and strict molecular clock. Every partition was assigned with a particular mutation rate according to Rieux et al., 2014.

We have run MCMC chains for 100000000 steps and sampling each 10000 steps. As the BEAST coalescence simulations are computationally very intensive, we have run this analysis using CIPRES – online open-access server for phylogenetics [Miller et al., 2010].

To ensure the high quality of models and good mixing of MCMC chains, only the simulations with effective sample sizes greater than 200 were considered for further analysis. BSPs were inspected in visualized by Tracer v. 1.4 [Rambaut et al., 2007].

To calculate basic parameters of genetic diversity of mitogenomic alignments of studied populations we have used the DnaSP v. 5 software [Librado et al, 2009]. These parameters were: number of haplotypes, number of polymorphic sites, haplotype diversity (HD), nucleotide diversity (Pi), average number of nucleotide differences (k) and Tajima's *D* value. Haplotype (genetic) diversity, one of the main parameters of population heterogeneity, was calculated as described in Nei [1987] (11),

$$H = \frac{n}{n-1}\left(1-\sum_{i=1}^{k} p_i^{\,2}\right) \qquad (11),$$

where n – the number of samples, k – the number of haplotypes, p_i – the relative frequency of the i-th haplotype.

In order to assess the aforementioned gene diversity parameters for mtDNA genes and RNA's in all studied populations we have partitioned the mitogenomic alignments into 13 protein coding genes, 22 tRNA's and 2 rRNA's, discarding all indels, and have calculated genetic parameters for each partition.

The partitioning was made by in-home Python script, which is available in our github page https://github.com/GrantHov/My_Python_codes .

The maps displaying spatial distribution of different haplogroup frequencies and genetic diversity were constructed by mapping Surfer v. 11 (Golden Software) package by the gridding method. Geographic coordinates of latitudes and longitude for all populations were based on the sampling centers.

2.3 Database construction

To facilitate the organization, management and further analysis of large human mitogenomic datasets we have designed and implemented a new database for complete human mitogenomes which we have called mtMart. For creating a database itself we have used MySQL open-source relational database management system. The functional characteristics of mtMart, which will be discussed in details in the Results chapter, are implemented in PHP and JavaScript programming languages. The information on complete human mitogenomes is retrieved from NCBI Genbank database using NCBI Application Programming Interface (API).

2.4 Simulation modeling

Today simulation modeling of genetic data plays increasingly important role in addressing different questions of population demographic history. In our project, we have applied the methods of coalescence simulation modeling in order to shed light on the demographic history of the Armenian population, particularly to address the highly debated issue about the origin of the Armenians. To simulate the mitogenomic data we have first constructed alternative models describing different versions of the origin of Armenian population (will be discussed in details in Results chapter).

Simulation of DNA data was performed by Fastsimcoal v. 2.5 software (*fsc*) [Excoffier et al., 2013]. For each model corresponding template (*.tpl*) and estimation (*.est*) files were created. For all alternative models 10000000 simulations were made. *Fsc* outputs simulated data into Arlequin *.arp* format file. Thus for calculating summary statistics of simulated and observed data for further comparative analysis we have used *Arlsumstat* package of Arlequin. Then in order to perform approximate Bayesian computations for obtained summary statistics to assess the most likely model we have used *abc* package [Csillery et al., 2012] of R statistical language.

All scripts in written in Python and R programming languages within this project are available through our Github page https://github.com/GrantHov .

CHAPTER III. RESULTS AND DISCUSSION

3.1 MtDNA genomic structure of Armenians

To elucidate the composition of mtDNA genetic pool of Armenians, we first performed the frequency analysis of mtDNA haplogroup distribution in all Armenian groups based on both complete mitogenomes (Table 6) and HVR1 sequences (Table 7). Sample sets that had multiple ambiguities of haplogroup definition were excluded from further analysis.

Table 6. Haplogroup frequencies based on complete mitogenomes

HG	aDNA	AV	CA	GP	KAR	WA	WA_FTDNA
C	-	-	-	-	-	-	1.4
D	-	-	-	-	1.4	-	-
F	-	-	-	0.7	-	-	-
H	1.9	-	4.1	2.2	4.1	11.0	1.4
H1	-	-	-	1.5	-	2.7	1.4
H13	1.9	1.6	2.7	2.2	2.7	2.7	5.4
H14	1.9	3.3	1.4	0.7	1.4	4.1	4.1
H15	5.8	3.3	2.7	1.5	2.7	1.4	-
H2	3.8	4.9	1.4	2.2	2.7	1.4	-
H20	-	-	1.4	-	6.8	-	-
H26	-	-	-	0.7	-	-	-
H28	-	3.3	2.7	1.5	-	-	-
H29	-	-	-	-	1.4	1.4	-
H3	-	-	-	-	-	1.4	2.7
H33	-	-	1.4	0.7	-	1.4	-
H4	-	-	1.4	0.7	-	1.4	-
H46	-	-	-	-	-	1.4	-
H47	-	-	-	-	-	-	4.1

HG	aDNA	AV	CA	GP	KAR	WA	WA_FTDNA
H5	-	1.6	2.7	0.7	2.7	5.5	6.8
H50	-	4.9	-	-	1.4	-	-
H51	-	3.3	1.4	-	-	-	-
H6	1.9	-	-	0.7	2.7	1.4	1.4
H7	-	-	-	1.5	-	-	2.7
H8	3.8	-	2.7	1.5	-	-	2.7
H87	-	-	-	-	1.4	-	-
H94	-	-	-	-	-	1.4	-
HV	1.9	3.3	-	-	-	-	-
HV0	-	-	-	-	-	-	2.7
HV1	3.8	-	-	6.7	2.7	9.6	4.1
HV12	1.9	1.6	-	2.2	-	-	1.4
HV13	-	-	-	0.7	-	-	-
HV14	-	-	-	-	1.4	-	-
HV18	-	-	1.4	-	1.4	-	-
HV2	-	1.6	-	0.7	1.4	-	-
HV21	-	3.3	-	-	-	-	1.4
HV4	-	1.6	1.4	-	-	-	-
HV8	-	-	-	0.7	1.4	-	-
HV9	-	-	-	2.2	-	-	-
I	-	-	1.4	-	-	-	-
I1	1.9	-	-	2.2	1.4	-	1.4
I2	-	1.6	-	-	-	-	1.4
I4	3.8	-	-	0.7	1.4	-	2.7
I5	1.9	-	4.1	-	-	1.4	1.4
I7	-	-	-	0.7	-	-	-
J1	7.7	6.6	5.5	15.7	10.8	5.5	6.8
J2	-	3.3	-	0.7	2.7	6.8	-

HG	aDNA	AV	CA	GP	KAR	WA	WA_FTDNA
K1	7.7	4.9	8.2	8.2	6.8	6.8	4.1
K3	3.8	-	-	-	-	-	-
L0	-	1.6	-	-	-	-	-
N1	-	4.9	2.7	4.5	6.8	4.1	-
N2	-	-	-	1.5	-	-	-
R0	-	-	1.4	-	-	2.7	2.7
R1	5.8	-	-	1.5	-	-	-
R2	-	1.6	-	-	-	-	-
T	1.9	-	-	-	-	-	-
T1	5.8	13.1	6.8	1.5	1.4	5.5	6.8
T2	3.8	3.3	12.3	4.5	5.4	4.1	8.1
U1	1.9	4.9	4.1	3.7	2.7	6.8	2.7
U2	3.8	1.6	4.1	-	1.4	1.4	-
U3	9.6	4.9	5.5	3.7	4.1	1.4	8.1
U4	3.8	-	-	1.5	1.4	-	2.7
U5	1.9	-	1.4	1.5	1.4	-	1.4
U6	-	-	-	0.7	-	-	-
U7	-	3.3	-	6.7	4.1	1.4	-
U8	3.8	1.6	4.1	0.7	-	-	-
V15	-	-	-	0.7	-	-	-
W	-	1.6	-	-	-	-	-
W3	1.9	1.6	-	-	2.7	-	-
W6	-	-	2.7	1.5	1.4	-	2.7
W7	-	-	-	1.5	-	-	-
W9	-	-	-	-	-	1.4	-
X2	-	1.6	5.5	2.2	5.4	2.7	1.4
X4	-	-	1.4	0.7	-	-	2.7
N	**52**	**61**	**73**	**134**	**74**	**73**	**74**

Table 6 illustrated that overall complexity of Armenian mitochondrial gene pool is very high, as soon as it is represented by multiple large haplogroups families, which in turn are divided into numerous subclades. For example, haplogroup H, which is considered a European lineage (based on its high frequencies in Western Europeans) is represented by more than 20 subclades in Armenians. The same complexity is observed for other large lineages such as HV and U.

Table 7 describes mitochondrial gene pool composition of HVR1 sequences in different Armenian geographic groups. As it was observed for complete mitogenomes, haplogroup composition in this case is also very diverse and complex. However, in this case the resolution of HVR1 region sometimes does not allow unambiguously assigning haplogroups to some samples, which are shown on the Table 7 as "Ambiguous" or CRS (resembling HVR1 region of Cambridge reference sequence, which cannot be assigned to a concrete haplogroup). In some populations, the proportion of ambiguities is quite high, for instance in CA_I and A_UNK groups it reaches 24% and 21%, respectively. Considering that these high proportions of equivocal haplogroups might introduce biases for further comparative analysis, we have used these samples only for calculating haplotype diversity parameters in those groups. Nevertheless, these data are still valuable for other analyses which do not require exact attribution of a given mtDNA sequence to a distinct haplogroup.

It should be noted that GP_I population does not have any ambivalent haplogroups since this dataset was also typed for the HVRII region or coding region SNPs to avoid ambiguities.

Table 7. Haplogroup distribution in HVR1 sequences.

HG	A_UNK	CA	GP_I	Iran	Salmast	KAR	Syunik	WA
A	-	-	-	-	-	0.4	-	-
Ambiguous	14.8	19.0	-	4.3	6.0	11.1	10.0	12.3
B2c2b	-	-	-	-	0.5	-	-	-

HG	A_UNK	CA	GP_I	Iran	Salmast	KAR	Syunik	WA
C	-	-	-	-	1.0	-	-	-
CRS	6.3	5.2	-	4.3	10.0	10.2	13.6	12.3
D	-	-	-	-	0.5	-	-	1.5
F	-	-	0.3	-	1.0	-	-	-
G	-	-	0.3	-	-	-	-	-
H	6.3	8.6	6.8	10.9	10.0	8.0	1.4	4.6
H1	1.7	1.7	0.5	-	-	1.3	2.1	-
H11	-	3.4	-	-	-	-	-	-
H13	0.6	-	3.3	-	-	0.4	-	-
H14	0.6	-	2.0	2.2	0.5	2.7	1.4	1.5
H15	0.6	1.7	2.5	-	1.0	-	0.7	4.6
H2	2.8	1.7	0.3	4.3	3.5	5.8	0.7	1.5
H20	0.6	1.7	0.8	-	1.0	0.4	-	1.5
H27	-	-	-	-	-	-	-	1.5
H29	-	-	0.3	-	-	-	-	-
H3	-	-	0.3	-	-	0.4	-	-
H33	-	-	0.5	-	-	-	-	-
H4	0.6	-	2.8	-	0.5	-	-	-
H42	0.6	-	-	-	-	-	-	-
H5	1.7	-	2.3	4.3	2.0	0.9	0.7	-
H6	1.7	-	0.8	-	-	0.4	-	-
H7	-	-	0.3	-	-	0.9	-	-
H8	-	-	0.5	-	-	-	-	-
HV	-	-	3.3	-	-	0.4	-	-
HV0	0.6	-	-	-	-	-	1.4	-
HV1	4.0	1.7	2.8	2.2	3.5	0.4	0.7	1.5
HV12	-	-	0.3	-	2.0	-	0.7	-
HV17	0.6	-	-	-	-	-	-	-

HG	A_UNK	CA	GP_I	Iran	Salmast	KAR	Syunik	WA
HV2	0.6	-	-	-	0.5	0.4	0.7	-
HV4	0.6	-	1.3	-	1.5	1.3	-	-
HV6	-	-	0.3	-	0.5	-	-	-
HV9	0.6	-	-	-	-	-	-	-
I	-	-	0.5	-	-	0.4	-	1.5
I1	0.6	-	0.3	-	1.0	2.2	-	1.5
I4	-	-	0.3	-	-	-	-	-
I5	-	-	0.5	-	-	-	-	-
J	1.7	1.7	-	4.3	2.5	2.2	2.1	1.5
J1	6.3	8.6	12.6	-	11.0	6.6	10.0	6.2
J2	1.1	-	2.0	2.2	0.5	0.4	0.7	-
K	-	-	-	-	3.5	-	-	-
K1	5.1	3.4	7.3	2.2	0.5	2.2	2.1	1.5
L0	0.6	-	0.3	-	-	-	-	-
L2	-	-	0.3	-	-	-	-	1.5
L3	-	-	0.3	-	-	-	-	-
M1	0.6	-	0.3	4.3	-	0.4	-	-
M10	-	-	0.3	-	-	-	-	-
N	-	-	-	-	1.0	-	-	-
N1	2.8	8.6	4.0	2.2	3.0	4.0	-	9.2
N2	0.6	-	0.3	-	-	-	-	-
N9	0.6	-	-	-	-	-	-	-
R0a	2.8	1.7	1.5	-	0.5	-	3.6	-
R1	-	-	0.5	-	-	-	-	-
R2	-	-	0.5	-	-	-	-	-
R9	-	1.7	-	-	-	-	-	-
T	0.6	-	-	-	-	1.3	1.4	1.5
T1	4.5	1.7	4.3	6.5	4.0	5.8	12.9	1.5

HG	A_UNK	CA	GP_I	Iran	Salmast	KAR	Syunik	WA
T2	6.3	13.8	7.6	13.0	5.5	7.1	7.9	4.6
U1	4.0	-	4.3	4.3	3.0	4.0	2.9	4.6
U2	3.4	1.7	1.5	6.5	1.0	1.3	2.1	4.6
U3	5.1	5.2	6.3	6.5	3.0	2.7	5.0	6.2
U4	1.1	1.7	0.8	-	2.0	-	2.1	-
U5	2.3	1.7	2.0	2.2	2.5	5.3	5.0	7.7
U6	-	-	0.3	-	-	0.9	-	-
U7	-	1.7	2.0	-	2.0	0.4	-	1.5
U8	-	1.7	0.8	-	0.5	0.9	-	-
V6	-	-	-	-	0.5	-	-	-
W	2.3	-	1.8	-	3.5	0.4	1.4	-
X	0.6	-	3.5	13.0	2.5	3.1	3.6	1.5
X1	0.6	-	-	-	-	0.9	0.7	-
X2	1.1	-	1.3	-	0.5	1.8	2.1	-
X4	-	-	-	-	0.5	-	-	-
N	**176**	**58**	**396**	**46**	**200**	**226**	**140**	**65**

We have also visulaised the haplogroup content of all the Armenian regional populations by reducing the haplogroups depth and constructing pie charts for every group (Fig. 12).

The Figure 12 shows that in a majority the Armenian matrilineal gene pool is composed of several main haplogroups in different proportions – H, U, J, T, HV, etc. with H haplogroup being the modal one in all populations, except aDNA. According to their geographic distribution, these lineages are considered as European haplogroups as they are found at very high proportions in the West European ethnic groups. However, the region of their origin is controversial and still actively debated.

The Armenian modal haplogroup H, which supposed to be originated in West Asia ca. 20-25 kya, is the most common lineage in Europe [Achilli et al., 2004], reaching 55-60% of frequency in western European populations with a negative

south-eastward cline. Thus, the distribution of this haplogroup in the Armenian population is consistent with its geographic location.

Haplogroup U, the second modal haplogroup in Armenians, is found at ca. 11-13% in western European groups [Helgason et al., 2001], and considered as one of the oldest (43-65 kya) matrilineal lineages of this region [Soares et al., 2009].

mtDNA haplogroup J originated ca. 32 kya ago [Soares et al., 2009], and today its frequency reaches 10-14 % in the western European populations [Helgason et al, 2001].

The haplogroup T, originated ~26 kya ago, and currently found at 6-10% in native Europeans, in AV and CA groups, is the second modal haplogroup, reaching 17% and 15% of frequency, respectively.

The haplogroup HV, which presumably originated ca. 27 kya, is also one of the most frequently encountered lineages in Armenians.

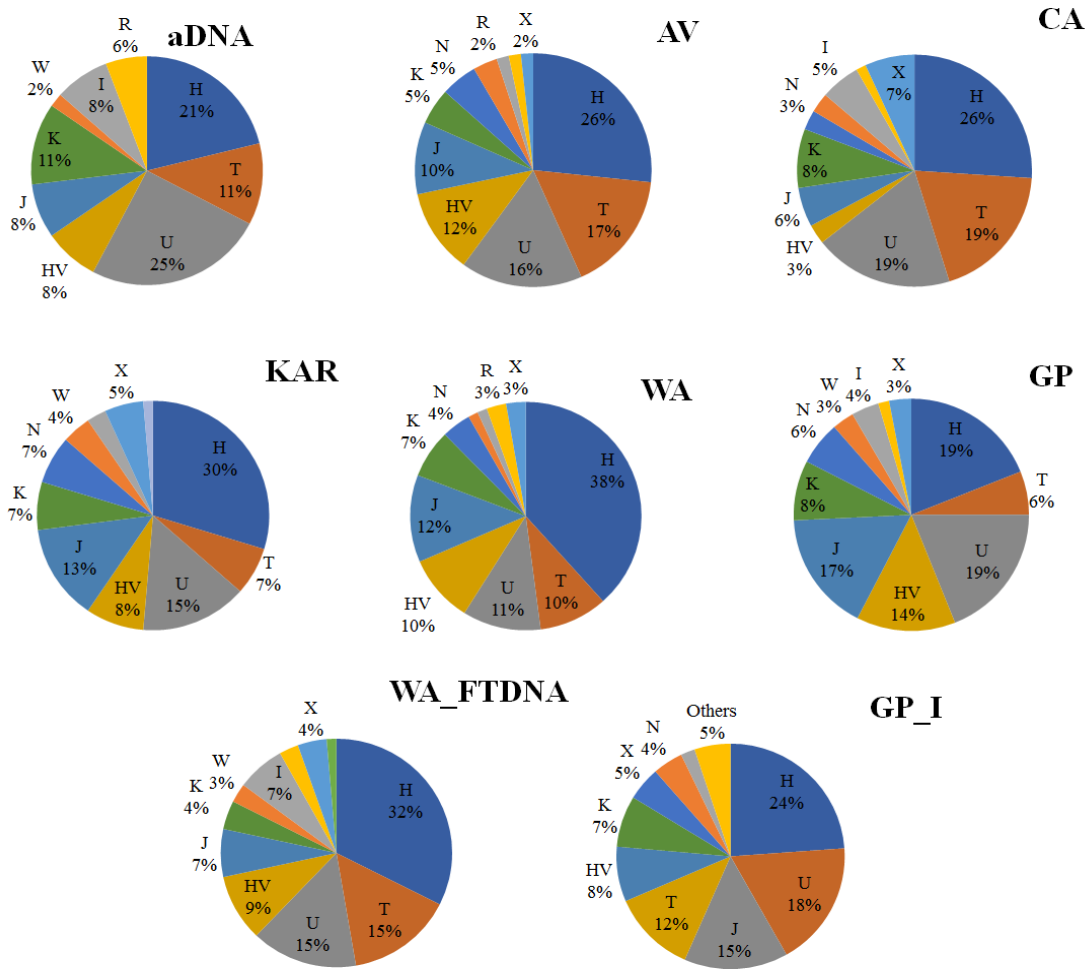


Figure 12. Haplogroup distribution in the Armenian groups studied.

The haplogroup frequency analysis has also shown that all the Armenian groups do not have any mtDNA lineages of Central or East Asian origin, namely A, B, C, D and G, which are the most prevalent haplogups in most of the Asian populations, comprising up to 70-80% of the matrelineal gene pool in some groups. In order to depict the pattern of geographic distribution of these haplogroups we have constructed the map (Fig. 13) using Surfer, showing the aggregated frequencies of A, B, C, D and G haplogroups in numerous populations of Central and East Asia, the Middle East, the Caucasus, Europe and northern Africa.

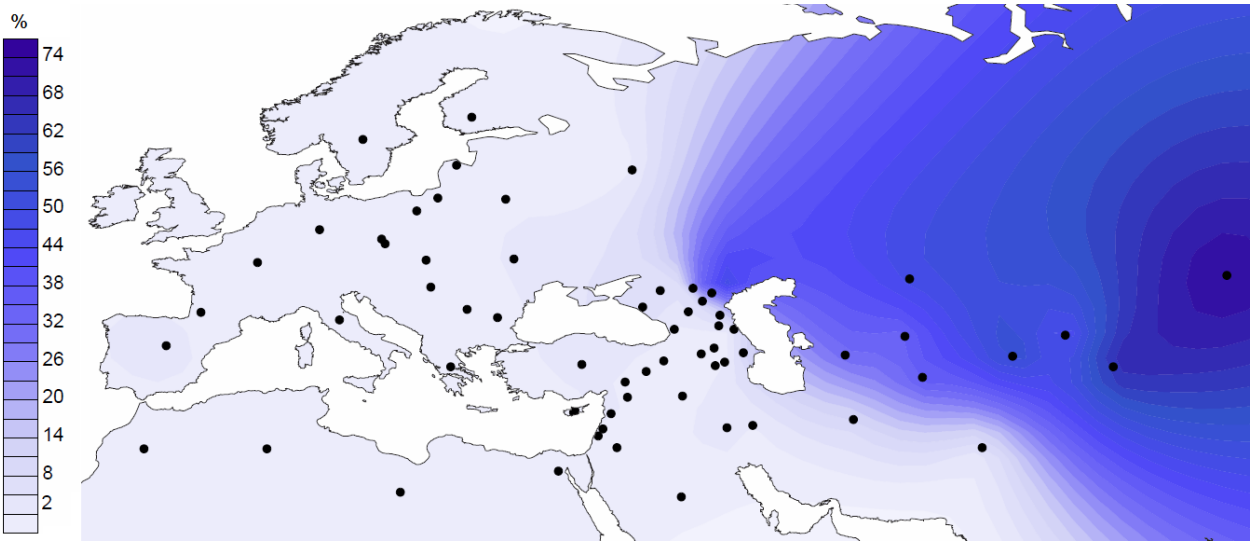


Figure 13. Summary distribution of 6 Asian haplogroups revealing the absence of Asian genetic influence on the Armenian gene pool.

The map clearly illustrates that frequencies of those haplogroups are declining westwards from Central and East Asia and are almost completely absent in gene pools starting from the Middle Eastern region. The exception from this pattern is represented by some Caucasian populations, namely Nagai and Tatar people, which are tracing their roots directly from Asian populations, and Azerbaijanis, in whose the haplogroups A, D and F in aggregate comprise 13% of mtDNA genepool.

As for the Armenian groups, it is notable that as the Aian haplogroups are almost completely absent from all populations (except KAR, WA_FTDNA and GP groups where aggregated percentage of the haplogroups comprise negligible 1.4, 1.4 and 0.7 %, respectively) represented by complete mitogenomes, as well as from HVRI sequences, except GP_I, Iran (Salmast), KAR and WA with 0.6, 3, 0.4, 1.5 %,

respectively, which again comprises insignificant portion of the gene pool. This overall pattern of Asian-specific haplogroups distribution in Armenian groups points out to the fact that despite numerous migrations and invading of central and east Asian Turkic-speaking populations through the Middle East and, subsequently, to Europe, the Armenian matrelineal gene pool does not show any significant signals of these contacts. Moreover, the absence of these haplogroups from aDNA samples and quite similar lineage composition of aDNA samples and modern Armenian populations assumes that at least during last 4 millennia Armenian matrelineal genetic pool was not affected by Asian influences.

In order to assess the diversity of the Armenian matrilineal genetic pool, we first have aligned all the mitogenomes to rCRS using multiple sequence alignment software MAFFT, then manually excluded insertions and deletions using MEGA software sequence editing tool. Afterwards we have extracted the HVR1 region from all the complete Armenian mitogenomes using our in-home python script. To be able to calculate haplotype diversity parameters for HVRI data, which was represented by mtDNA haplotypes, we have converted the latter into sequences in the positions 16024-16400 b.p., which corresponds to HVRI using Haplosearch. Diversity parameters were calculated for using DNASP5 software, and the results are shown in Table 8.

To visualize the pattern of haplotype diversity *h* in Armenians and neighboring populations, we have constructed a geographic map of *h* distribution of populations, representing the territory of Europe, the Caucasus, the Middle East and the North Africa (Fig. 14).The map shows that the North African and the Middle Eastern populations as expected have relatively higher diversity of their matrilineal gene pool, reflecting higher age of these groups than those of Western European populations.

Strongly negative Tajima's D values observed in the all Armenian groups might suggest the recent population expansion or the presence of purifying selection of some variable sites in Armenians.

Table 8. Gene diversity parameter of the Armenian populations. Abbreviations: N-number of individuals, S – number of segregating sites, Eta – number of polymorphisms', h – haplotype diversity, pi – numer of mutations per site, SigD – statistical significance of Tajimas's D value, * - < 0.05, ** - < 0.01.

Population	N	S	Eta	Hap	*h*	Pi	AvNumDif	G+Ctot	TajimaD	SigD
SYN	140	81	86	77	0.973	0.015	5.634	0.460	-2.025	*
AV	61	65	66	40	0.975	0.016	5.915	0.460	-1.982	*
WA	65	69	72	51	0.982	0.014	5.378	0.459	-2.196	**
KAR_I	226	102	109	128	0.984	0.016	5.846	0.460	-2.088	*
WA_FTDNA	74	62	62	47	0.985	0.015	5.581	0.459	-1.867	*
Iran_Salmast	200	107	112	131	0.986	0.015	5.651	0.458	-2.189	**
GP_I	396	125	138	241	0.989	0.015	5.527	0.460	-2.201	**
WA	73	65	66	55	0.99	0.015	5.543	0.460	-1.978	*
CA	73	73	75	57	0.99	0.018	6.610	0.460	-1.922	*
A_UNK	176	99	102	117	0.991	0.015	5.720	0.460	-2.122	*
Iran	46	54	55	38	0.991	0.016	5.856	0.462	-1.867	*
CA	58	69	71	49	0.993	0.017	6.299	0.458	-2.031	*
KAR	74	76	79	60	0.993	0.016	6.023	0.458	-2.115	*
GP	134	99	104	107	0.996	0.017	6.259	0.458	-2.151	**

Interestingly, the high values of gene diversity alongside with strongly negative Tajima's D might assume that the processes of genetic diversification of the Armenians ended before the population expansion [Hellenthal et al., 2014].

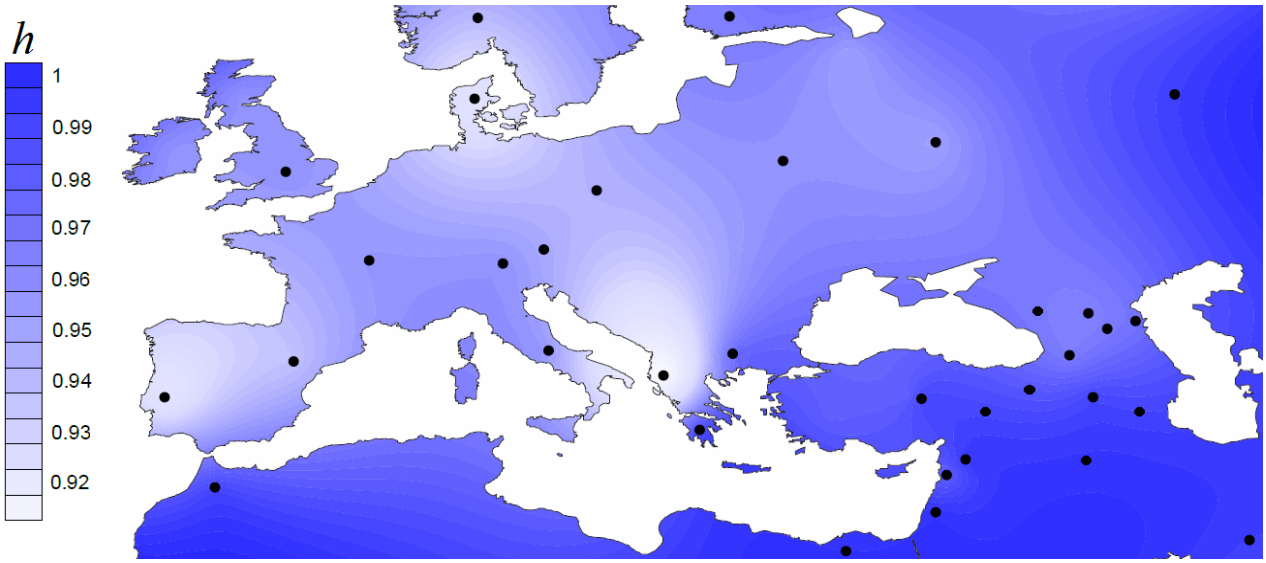


Figure 14. The spatial distribution of gene diversity (*h*) values in different populations.

Additionally, we have calculated gene diversity parameters for every gene of mtDNA in all the Armenia complete mitochondrial genomes. For this aim we have used our indels-free alignment and splitted whole mitochondrial genome into 13 genes, and for each gene for each Armenian group we have assessed gene diversity parameters using DNASP5.Within this dissertation we have not used this data to make any inferences about Armenian mitochondrial gene pool structure, as soon as coding regions might introduce biases because of possible selective pressures acting on different mutations influencing fitness of the organism. However, this data will be involved in other projects related to investigation of distribution of pathological mutations in Armenian populations to reveal possible associations of different mtDNA variants with pathological phenotypes.

In order to characterize the relationships between Armenian groups and other populations, we have compiled a large dataset of comparative datasets, representing more than a hundred populations worldwide. We have assessed the position of the Armenians among this datasets by using several methods. First, to see if our comparative sample set was collected appropriately, we have calculated Fst genetic distances between all the populations based on 40 haplogroup frequencies, and based on Fst matrix constructed a heat-map and performed hierarchical clustering (Fig. 15) with the aid of Gene-e software.

As it was expected, the populations grouped according to their geographic origin, show once again that mtDNA genetic distances are well correlated with geographic distances. Heat map clearly demonstrates that there are 5 main clusters of populations – a very large joint cluster of the Middle East, the Caucasus and Europe (notable that European groups clustered tightly with each other, which show their genetic homogeneity), Eastern Asia, Siberia, Central Asia, and Africa.

As soon as the heat map is hardly readable on the scale of individual populations because of their large number and figure resolution, it should be mentioned that Armenian groups are located within a large joint cluster, thus indicating genetic proximity to those groups.

Considering that the heat map shows only a rough pattern of genetic affinities of the studied groups, here we have also applied more statistically sophisticated multivariate approach – correspondence analysis.

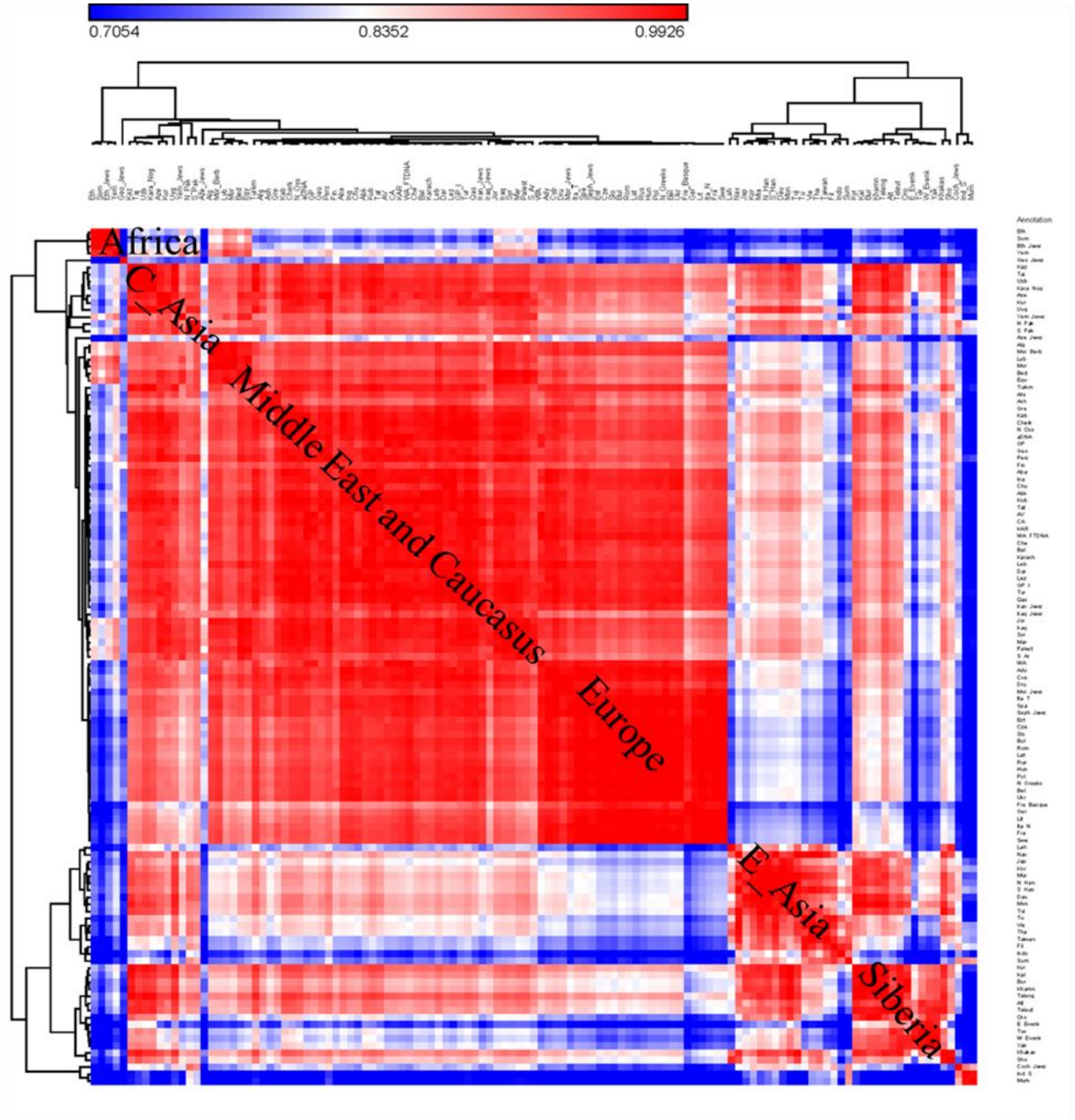


Figure 15. The heat map of all studied groups based on the Fst genetic distance. Blues shades indicate higher distance, red – lower.

It was based on haplogroup frequencies of the studied populations, which not only reduces the dimension of multidimensional haplogroup frequency table, giving

an easily readable and informative two-dimensional representation of relative locations of the studied groups, but also establishes a correspondence between the pattern of distribution of populations and the impact of different variables, i.e. haplogroups, that shaped the given localization of populations on the plot. The last feature can be also used as a characteristic of the appropriateness of sampling of the comparative datasets, as soon as some mtDNA haplogroups are considered to be region-specific, as is shown in Figure 13. The correspondence analysis was performed by SPSS statistical package and results were visualized using ggplot2 module of R programming language.

The results of correspondence analysis are presented in Figure 16 and 17, showing the pattern of distribution of analyzed populations and the location of haplogroups. As a rule, these two types of information are depicted on the same plot, but considering that the number of populations is very big and most of the groups are located close to each other, we have plotted the results on populations and corresponding haplogroups on two different plots to make the overall result more readable.

The Figure 16 shows that the studied groups are forming similar clusters observed on heat-map. The Middle Eastern, Caucasus and European groups appear together, while other groups from Asian regions tend to form distinct clusters. This pattern of distribution is well explained by the Figure 17, showing clear separation of haplogroups, which push the Asian populations to the right side of the plot. Indeed, these haplogroups, such as A, B, C, etc. comprise the main portion of Asian maternal gene pools and thus are considered in literature as Asian specific mtDNA haplogroups. On the other hand, such haplogroups as H and subclades of U, J, T, etc. are widespread in Europe and the Middle East, thus placing corresponding populations close to each other and forming a large cluster of genetically similar populations.

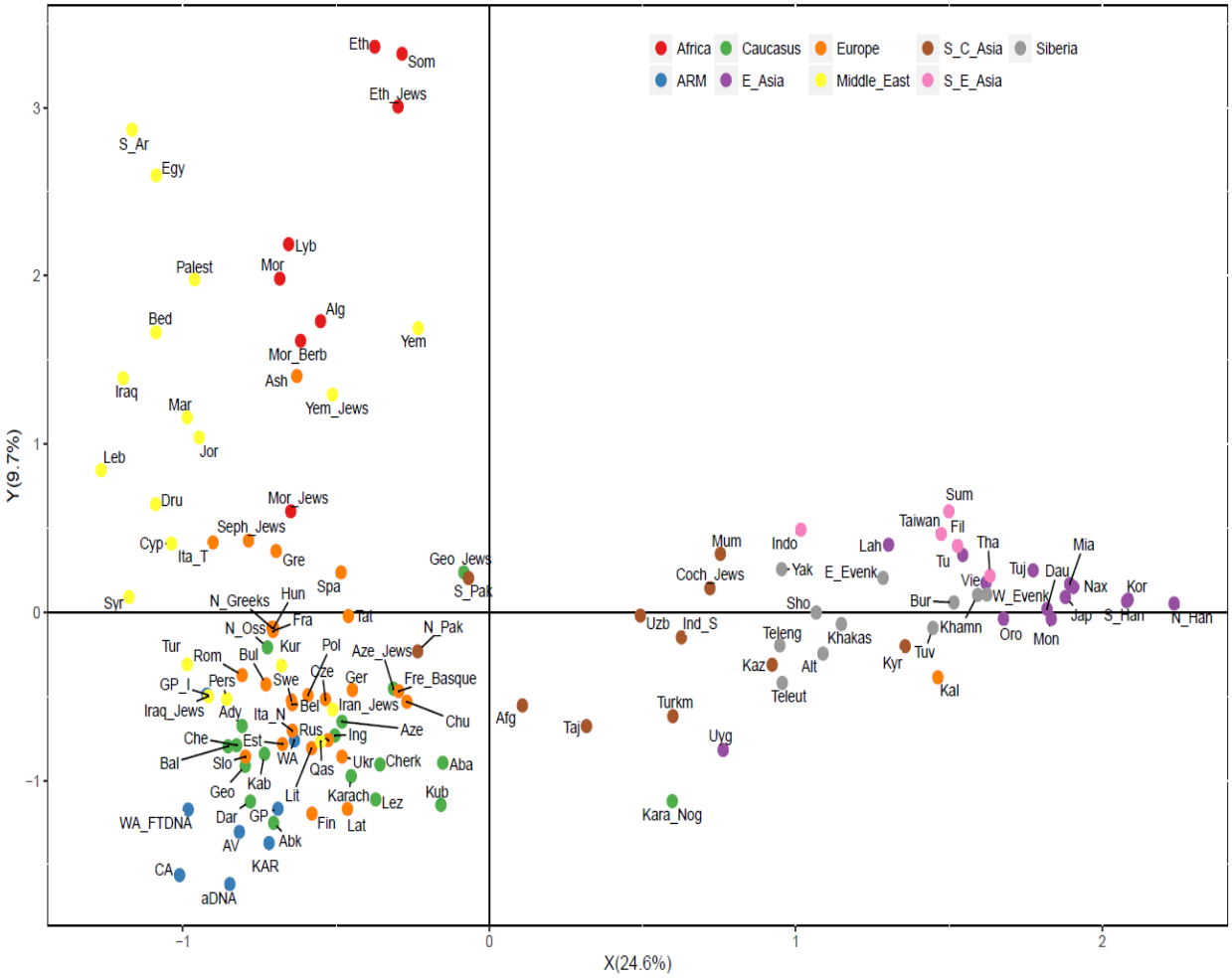


Figure 16. Correspondence analysis of studies popualtions based on haplogroup frequencies.

The plot shows that Armenian regional groups tend to cluster with each other with the only exception of WA dataset, which falls closer to the European and Caucasus groups. Nevertheless, this fluctuation does not necessarily imply significant difference between the WA group and other Armenians, but rather points to the genetic proximity of overall Armenians and neighboring populations based on mtDNA. It should be noted that ancient DNA sample set is also involved in the Armenian cluster and is the closest to Karabakh and AV samples, which indicates that these groups might have close relationships with each other. And indeed, considering that the bone samples were collected from nowadays Republic of Armenian and Karabakh, this assumption is in a good congruence with the sampling strategy and obtained results.

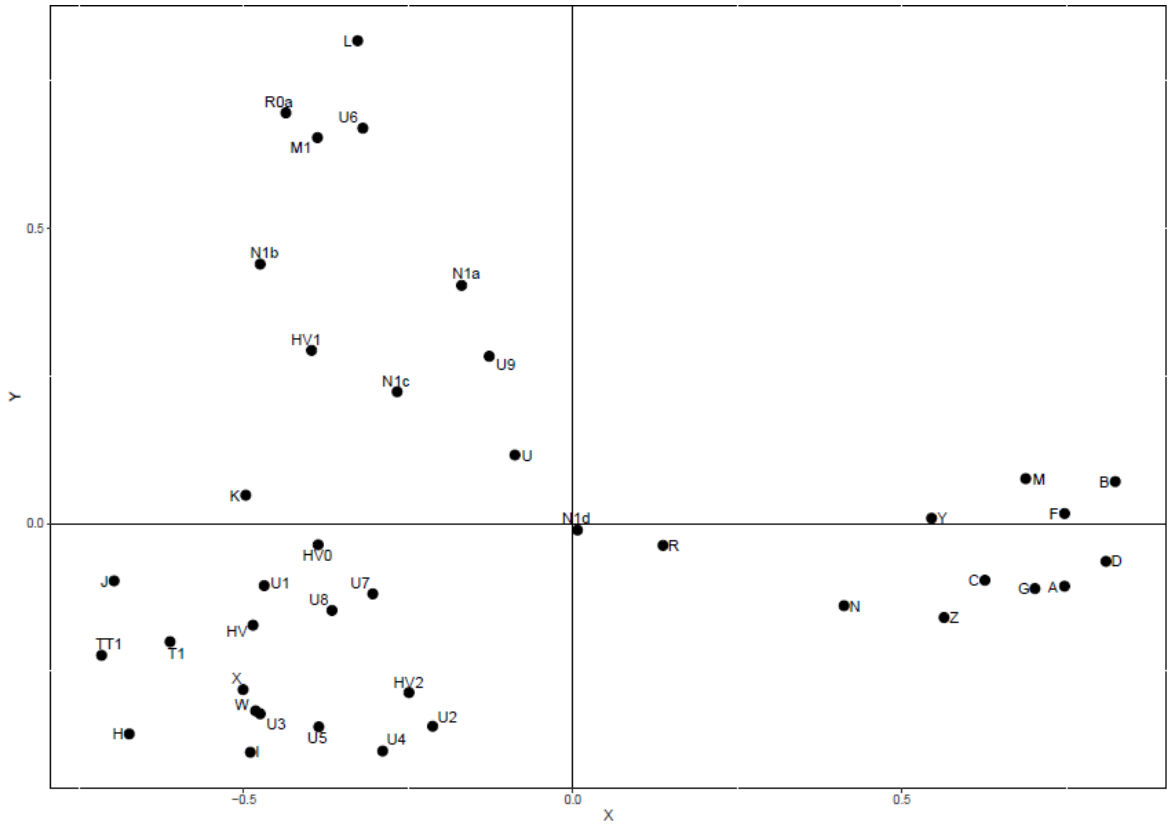


Figure 17. Correspondence analysis demonstrating the contribution of haplogroups in the distribution of populations indicated in Fig. 16.

It also should be mentioned that in spite of misleading dispersed pattern of Levant populations on the plot, the Armenians are also in tight relationships with these groups, i.e. Syrians, Lebanese, Palestinians, etc. These counterintuitive interpretations are explained by the proportion of variation that each axis of the plot describes: X axis describes 24.6% of total variation, while Y axis – only 9.7%, which means that visual distance on horizontal axes is almost 2.5 times longer than on the Y axis, and thus if the lengths of axes would be transformed to be proportional to the described variance, populations located on the top of the plot will be visually displaced towards Europeans, Caucasians and Armenians.

To further elucidate the relationships of the Armenian groups with each other we have estimated the statistical significance of Fst genetic distances between them. Table 9 shows the portion of the p-values matrix that corresponds to the Armenian groups. Minus signs indicate that Fst genetic distance between pair of populations is not statistically significant ($p<0.05$).

Table 9. The values of the test for population differentiation of Armenian group based Fst genetic distances calculated form haplogroup frequencies. In lower triangular matrix "–" indicates the absence of significant difference (p>0.05), "+" – the presence, upper triangular matrix – p-values.

Populations	aDNA	AV	CA	GP	KAR	WA	WA_FTDNA	GP_I
aDNA	*	0.3889	0.4130	0.0401	0.0823	0.0175	0.3768	0.0001
AV	-	*	0.2975	0.3096	0.6716	0.8148	0.2607	0.0113
CA	-	-	*	0.0016	0.2626	0.1351	0.5451	0.0033
GP	+	-	+	*	0.9552	0.1770	0.0040	0.0000
KAR	-	-	-	-	*	0.7618	0.2057	0.0005
WA	+	-	-	-	-	*	0.2122	0.0008
WA_FTDNA	-	-	-	+	-	-	*	0.0000
GP_I	+	+	+	+	+	+	+	*

As it is visible from the table, most of the Armenian groups according to genetic distances based on haplogroup frequencies are not significantly different, thus implying that Armenian matrilineal gene pools from different geographic regions are homogeneous. Noteworthy that the ancient DNA dataset also follows this tendency, which again supports the idea that the ancient sample set resembles genetic composition of modern Armenian groups. The only exception are the datasets representing both the General population, which might be explained that those two comprise not only the geographic groups considered in this study, but also Armenians form other geographic region, who might potentially be differentiated from our datasets.

To test this result on a higher and ultimate resolution of mtDNA data, we have conducted the same analysis using complete mitogenomes of only Armenian groups, and here we have also included Armenian mitogenomes reported in Schoenberg et al., 2011. The results, which are depicted on Table 10 show that even using the data on complete genomes, the Armenian populations do not significantly differ from each other, which confirms the previous result based on haplogroups frequencies, suggesting that the maternal portion of population's gene pool is highly diverse, but nevertheless it comprises a single ethnocultural group. This striking result is in contrast with previous studies [Weale et al., 2001, Yepiskoposyan et al., 2001]

performed on Armenians using the Y-chromosomal markers. The most plausible explanation of this discrepancy might be the social practice of patrilocality in the Armenian communities. Patrilocality, or a higher rate of women's spatial dispersal, is a widely spread cultural tradition for human societies. This practice is being used to explain the following patterns detected in ethnically different populations: relatively low Y chromosome and high mtDNA variability within populations, large between-group genetic distances for the male chromosome and small distances for the mtDNA [Seielstad et al., 1998; Perez-Lezaun et al., 1999; Jorde et al., 2000; Wilson et al., 2001]. It is usually explained by higher rate of female versus male migration which takes place when women move to their husbands' residence after marriage. Traditionally, the Armenian society is patrilocal, and the deviation from this cultural practice is very rare and only due to specific circumstances [Redgate, 2000]. And indeed, our study demonstrates that the practice of patrilocality had a strong effect in shaping the genetic composition of the Armenians.

Notable, that aDNA sample also follows the previously shown pattern of genetic homogeneity with other Armenian groups, strengthening the confidence that this ancient population has direct roots with modern Armenian groups, which however will be addressed further using more sophisticated bioinformatics techniques.

It also should be mentioned, that dataset taken from the paper of Schoenberg and coauthors is falling out from the overall picture, showing a difference with all the other groups, except GP. And this is also a good indication that the sample size and, more importantly, sampling strategy has a crucial role when studying Armenian population.

Table 10. The values of the test for population differentiation of Armenian groups based Fst genetic distances calculated form complete mitogenomes. In lower triangular matrix "–" indicates the absence of significant difference ($p>0.05$), "+" – the presence, upper triangular matrix – p-values.

Populations	aDNA	AV	CA	KAR	WA	WA_FTDNA	GP	ARM_D
aDNA	*	0.186	0.712	0.338	0.030	0.319	0.120	0.002
AV	-	*	0.373	0.361	0.229	0.246	0.038	0.004

Populations	aDNA	AV	CA	KAR	WA	WA_FTDNA	GP	ARM_D
CA	-	-	*	0.155	0.026	0.413	0.007	0.003
KAR	-	-	-	*	0.294	0.209	0.685	0.004
WA	+	-	+	-	*	0.281	0.024	0.000
WA_FTDNA	-	-	-	-	-	*	0.009	0.001
GP	-	+	+	-	+	+	*	0.105
ARM_D	+	+	+	+	+	+	-	*

Taking the advantage of having ancient mitogenomes from Karabakh and Republic of Armenia, which are spanning from the Bronze Age to Medieval, we aimed to reconstruct the demographic history of Armenian population using these ancient samples. Since the calibrating time points to coalsecence simulation modeling alogrythms, there were implemetned in different phylogenetic software. Here, we have used BEAST v. 1.8.3 to generate Bayesian skyline plots showing the changes of Effective population size Ne throughout the time.

We have generated BSP plots for each population and for aggregated Armenian population comprising aDNA+AV+CA+KAR (Fig. 18), which are the samples that do not significantly differentiated form each other as was shown in previous analysis.

The patterns of demographic changes in all groups are quite similar – there are four main demographic events for all groups, which are (1) population expansion ca 40 kya with a relatively constant size until 20 kya, when a small population size decline occurs, (2) after this we see an extensive population expansion around 15-18 kya (3) that might explain negative valuse of Tajima's D parameter, which coincides with Last Glacial Maximum and can be interpreted as the fact that Armenian Higland was a refugial zone for human populations at that period of time; after that for in all populations, except KAR, we observed that ca 7-10 kya population size starts declining slowly (4) and reaches to current day female effective population size. Notable that when we have performed the BSP analysis for the aggreagated population we have obtained a similar result indicating that on the whole the territory of the Armenian Higland the demograhic histories of populations were very similar and one can argue that indeed, as was shown in other analyses above, these

similarities point to the fact that the analysed populations might belong to a single, genetically isolated from other populations, group.

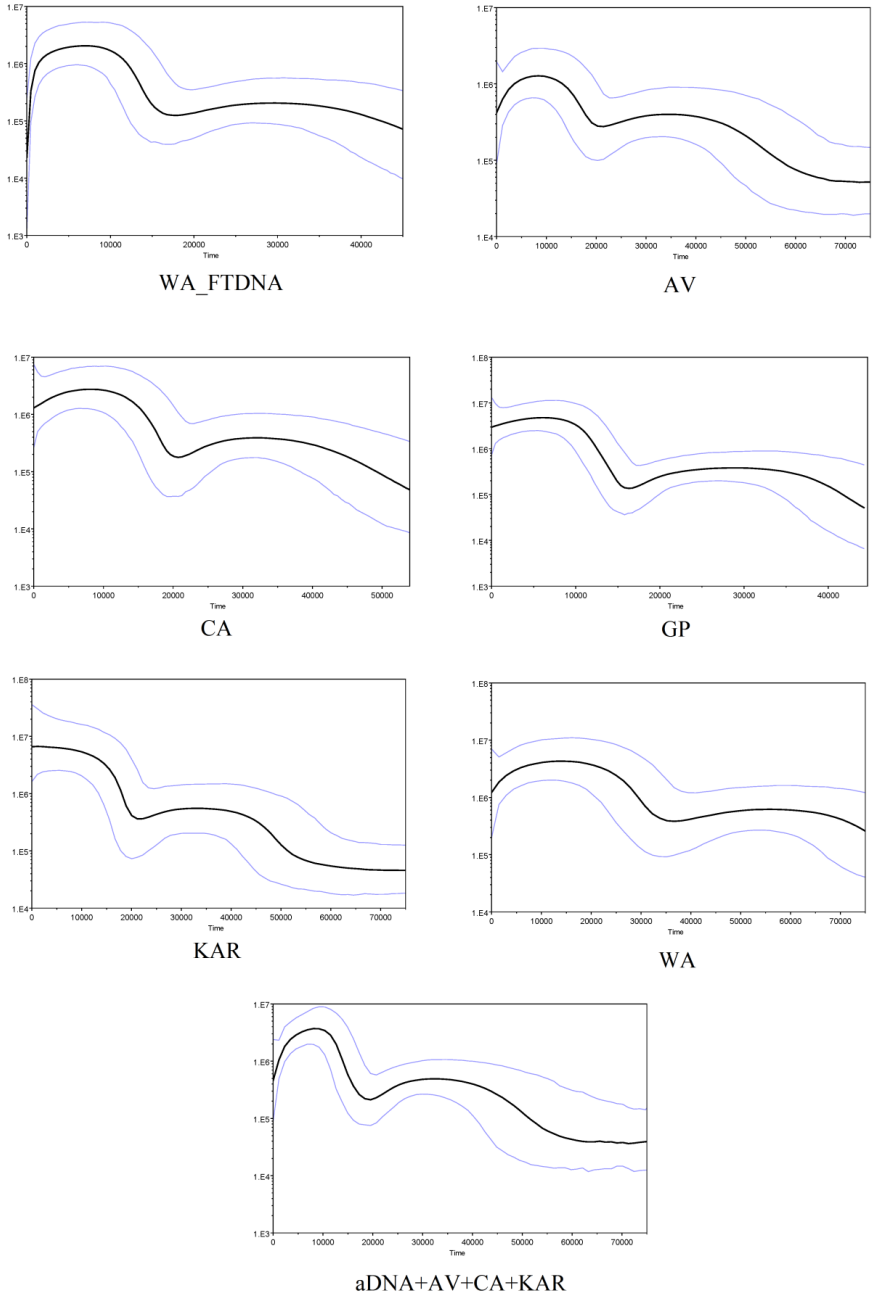


Figure 18. Bayesian skyline plots reflecting demographic fluctuations of studies population. Black line shows mean Effective population size, blue lines indicate 95% confidence intervals.

3.2. Simulation Modeling

Using different population genetics and bioinformatics methods in our study, we have shown above that in most of the cases ancient DNA samples show high similarity with modern day Armenian populations. Particularly, analysis of haplogroup

distribution and test for the population differentiation have shown that these two types of data do not significantly differ from each other. Additionally, BSP plots have also shown a very similar pattern of demographic changes that occurred throughout the history of several Armenian territorial groups separately and in aggregate by including aDNA as calibration points.

Considering all above mentioned, one can argue that there is a marked genetic continuity between ancient and modern data, indicating that at least for ca 4 thousand years Armenians did not substantially change their genetic composition. However, in order to make this striking conclusion, it is necessary to perform more sophisticated and deep comparison between modern and ancient data. Thus, in our work we have conducted recently developed, but widely accepted powerful simulation modeling of DNA data under different demographic scenarios and comparison of obtained results using the method of Approximate Bayesian computations.

To simulate DNA sequences under 5 different demographic scenarios we have used *fastsimcoal2* method. The best model and its parameters were inferred via Approximate Bayesian Computation (ABC) using the "ABC" package for R programming language. For all models, we have used generation time 25 years and a mutation rate $5.5x10^{-7}$/site/generation. The summary statistics of simulated and observed data were calculated with *arlsumstat*, a modified version of Arlequin software. To compare the models using ABC, we have considered four different types of parameters as summary statistics: Fst genetic distances between the modern and ancient Armenian groups, number of polymorphic sites (S), mean number of pairwise differences (π) and Tajima's D (D) calculated for each group separately.

The first 4 models start with a bottleneck around 800 generations ago, e.a at the end of the Last Glacial Maximum (LGM), which is clearly visible from the Bayesian skyline plots. In model 1, this bottlenecked maternal population gives rise to the ancient Armenian group around 50-300 generations ago which in its turn forms the modern Armenian population of size N_M (genetic continuity model). In model 2, the initial small group gives rise to two different populations representing the modern

and ancient groups around 301-799 generations ago, sometime after the LGM and before the sampling of ancient population. In the third model, the initial maternal population splits into two groups right after the LGM without any admixture events between the two populations forming the modern and ancient populations with populations size N_M and N_A, respectively. Model 4 is similar to model 3 with an additional admixture event (ranging from 10 to 90%) at the time of sampling of the ancient population. The last model starts with a bottleneck of non-European maternal population around 2,000 generations ago which gives rise to an ancestral population to the modern and ancient Armenian group and an Asian lineage (ghost population). The demographic relationship of the ancient and modern Armenian groups is similar to the one in model 4, but we also simulate a constant migration event at a rate 0.0001 from the ghost population from Asia to the Armenian maternal gene pool for the past 50 generations.

We used approximate Bayesian computation (ABC) analysis to test 5 possible demographic model scenarios (Figure 19) each of which were simulated 1,000,000 times with *fastsimcoal2*. For the modern Armenian group, we used the combined Armenian population (AV+CA+KAR).

The final analysis has shown that the most favored model is the model1 which assumes genetic continuity between the ancient and the modern Armenian groups.

As illustrated in the posterior probabilities table, genetic continuity model has the highest probability (0.647) of being accepted as the best model describing the genetic variability of the observed data. The second most probable model is Model 4 (posterior probability = 0.159), which is assuming a different origin of modern and ancient samples but with an extensive gene flow from ancient to modern samples ca 200 generations ago. And models 2, 3 and 5 have very low probabilities, and thus might be discarded as least possible evolutionary scenarios of Armenian ethnogenesis. The posterior model probabilities and Bayes factors are presented in Table 11 and 12, respectively.

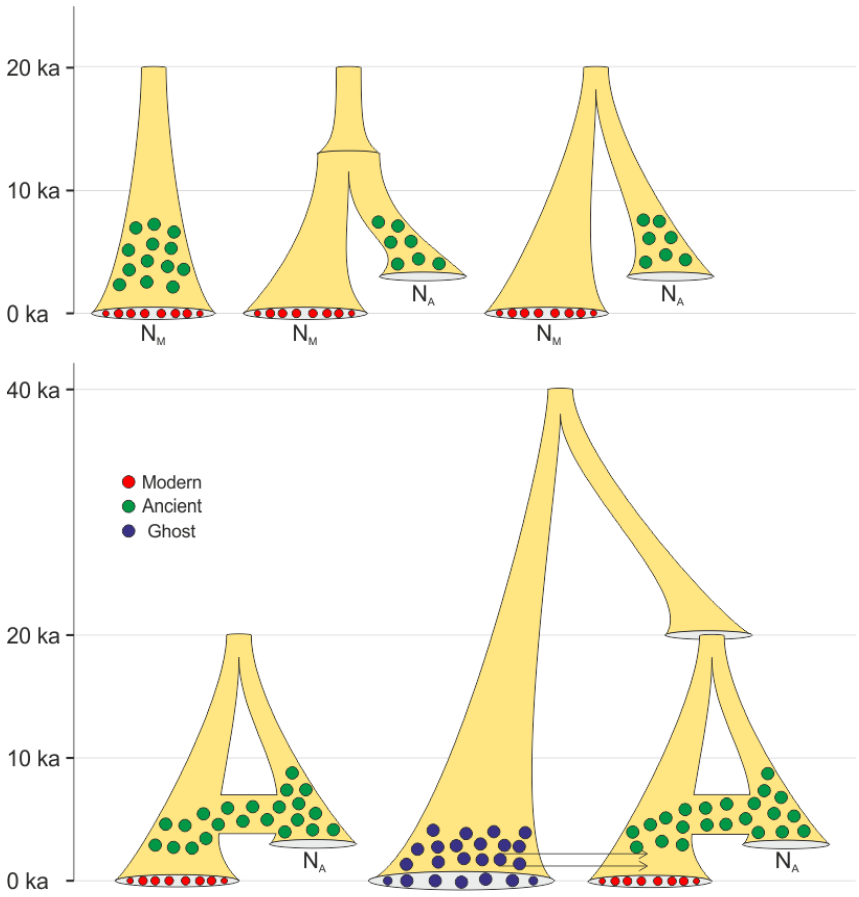


Figure 19. Schematic representation of the 5 demographic models used for coalescent simulations. N_M-population size of modern Armenians (red), N_A-population size of ancient Armenian group (green), N_G population size of the "ghost" Asian population (purple).

Table 11. The posterior model probabilities for the simulated demographic models.

model-1	model-2	model-3	model-4	model-5
0.6471	0.0983	0.042	0.1596	0.053

Bayes factors, which reflect the degree of confidence about the acceptance or rejection of null and alternative hypothesis, demonstrate that the model 1 can be considered as the most likely scenario of the Armenian ethnogenesis (BF>9 in average, if model 1 considered as alternative, and DF<0.5 vice versa) as it is indicated in the Table 12 and 13. The most favorable model is model-1 suggesting a genetic continuity between modern day Armenians and ancient samples.

Table 12. Bayes factors for the simulated demographic models.

Model	model-1	model-2	model-3	model-4	model-5
model-1	1	6.5804	15.4242	4.0539	12.2104
model-2	0.152	1	2.344	0.6161	1.8556
model-3	0.0648	0.4266	1	0.2628	0.7916
model-4	0.2467	1.6232	3.8047	1	3.012
model-5	0.0819	0.5389	1.2632	0.332	1

3.3. Database

Today, the advances of novel DNA sequencing methods have made the mitochondrial genome a versatile tool for phylogenetics, population genetics, forensic science and other disciplines [Brandon et al., 2006, Behar et al., 2012]. Even though human mitochondrial DNA is a tiny molecule ca. 16,570 b.p. [Anderson et al., 1981], continuously accumulating large-scale mtDNA data make the handling, analyzing and comparing the mtDNA gene pools of different human populations hard. Recently, several attempts were made to design and create publicly available human mitochondrial DNA databases; however, some of them are no longer updated and maintained, while the rest do not provide convenient functionality for effective data management and further analysis. For instance, HvrBase++ [Kohl et al., 2006] and mtDB [Ingman et al., 2006] databases were launched in 2006, but they have not been updated since 2007, while the number of new mtDNA partial and complete sequences has increased significantly since then. Another quite functionally rich database, hmtDB that has numerous options for data searching, mtDNA haplogorup assignment, etc., is rarely updated. On the other hand, the Phylotree [van Oven et al, 2009] and Mitomap [Kogelnik et al., 1995] human mitochondrial DNA databases are updated regularly; however, the absence of functional tools that might help to deal with large datasets of mtDNA molecules of different characteristics restricts the efficacy of data handling.

Here, we introduce the mtMart (by analogy with BioMart search engine of Ensemble genome browser [Yates et al., 2016]), a new database for human complete mitochondrial DNA data, which is designed to fill the gaps of the above described databases and implement our own ideas in order to significantly facilitate the effective treatment of largescale human mitochondrial genome information.

The main principles of mtMart workflow are described in Fig. 20.

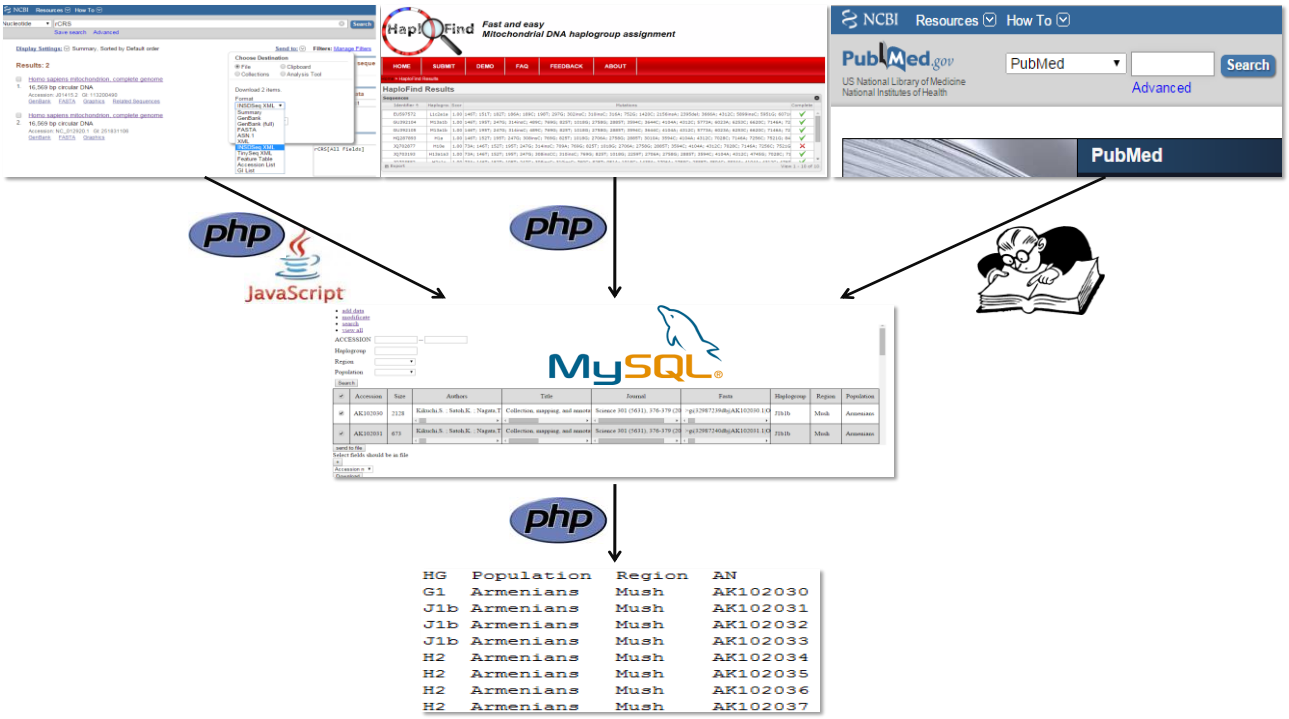


Figure 20. General workflow of mtMART.

The database is designed with MySQL open-source relational database management system. Its functional characteristics are implemented in PHP and JavaScript (jQuery) programming languages. mtMart retrieves the information on human mitochondrial genomes (however, in principle it can be used for retrieving any query) directly from the National Center of Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/), which provides the API (Application Programming Interface) using an arbitrary set or range of accession numbers of interest. Using in-home PHP script to process the INSDSeq XML file, which is generated from

GenBank (http://www.ncbi.nlm.nih.gov/genbank) and contains all the information about the query, mtMart stores obtained data in internal memory.

One of the main features of mtMart is the possibility to semi-automatically add the information about mtDNA haplogroup, mutation data compared to revised Cambridge Reference Sequence, ethnic group and geographic region of the population to any record of the database. For assigning the haplogroup and defining mutated positions, mtMart is synchronized with Haplofind, a fast and reliable web application for high-throughput human mitochondrial genome haplogroup assignment based on the most recent Phylotree built. Besides haplogroup assignation, Haplofind also defines the mitochondrial DNA mutations that were reported to be associated with the particular disorders, and this data is also added to mtMart. The data on population and geographic region are appended to mtMart manually. In order to avoid some undesired manipulations of users, the functions of addition and removal of the data are restricted for all users, except the Administrator. Another important feature of mtMart is the possibility to search, sort and download data in a very customizable way. One can sort it either by haplogroup (at any resolution of phylogenetic tree), population and geographic region or combine all these options in order to obtain the necessary result. After being obtained, the information can be retrieved in several output formats. For complete sequences and mutation data the mtMART outputs the information into FASTA and FASTA-like formats respectively, since some widely used software for mitochondrial DNA data analysis, such as MITOTOOL use FASTA-like format composed of mutated sites only, instead of DNA sequences. On the other hand, for downloading the data from sortable columns (haplogroup, population, region, etc.) and accession number the database allows to output the result in commonly used text formats (tab, comma, colon, etc.-separated files) with the arbitrary order of columns. Moreover, for haplogroup and population or geographic data we have implemented the algorithm allowing to automatically generate the .arp Arlequin input file with relative haplogroup frequency values. In our project, we have developed a new user-friendly database for complete human

mitochondrial genomes, mtMART which might significantly support largescale human mitogenomic studies. mtMART is built taking into account the conveniences of previously designed mtDNA databases and filling the gaps of the last. Our database will be continuously developing allowing researchers to use new functionally convenient features for fast and effective data management. For now, the development of mtMart is still under way; however, it is publicly available for usage though http://mtmart.org/.

CONCLUSION

In this dissertation, for the first time we have studied Armenian matrilineal genetic structure by analyzing massive amount of high quality mitogenomic data collected from different regions covering the whole area of Historical Armenian by our team throughout last decade and already available but unpublished data on Armenian mtDNA. Moreover, here, we have also studied the first so far available data on ancient mitogenomes representing different historical epochs, which were collected for the different archaeological sites of the Republic of Armenia and Artsakh. This overall dataset of the Armenian mitogenomic data comprised ca 1900 samples, making our population one of the best-characterized ethnic groups worldwide based on mitochondrial DNA.

To elucidate the genetic and demographic history of the Armenian using these data, we have applied sophisticated bioinformatics and data analysis methods that were already available, as well as developed our own scripts for efficient data parsing and further analysis. Moreover, in frames of our project, we have also designed and developed a new publicly available database human mitogenomic studies, which will enhance the effectiveness of large human mitogenomic studies, thus supporting mtDNA related scientific community.

Based on these highly informative datasets, we demonstrated that the Armenian groups from different geographic regions have a very complex haplogroup composition, represented by multiple large haplogroups families, which in turn are divided into numerous subclades. The majority of the mtDNA gene-pool is represented by several large lineages, as H, U, J, T, etc., which are considered as European haplogroups based on their high frequencies in Central and Western Europe. Haplogroup H is the most widespread in all the datasets, except aDNA samples. The haplogroup frequency analysis has also demonstrated that Armenian groups do not have any mtDNA haplogroups of Central or East Asian origin, namely A, B, C, D and G, which are the most prevalent haplogroups in many Asian populations. This fact clearly demonstrates that despite numerous migration on the

Armenian Highland that took place during the past millennium, the Armenian gene pool was not affected by the Asian influences.

Despite the genetic complexity of the Armenians, our analysis have shown than when comparing the Armenian groups with each other, in most of the cases there is no marked genetic differentiation between them. Indeed, the test of statistical significance of genetic distances Fst in all pairs of groups showed that p values are more than 0.05. This phenomenon was shown either when using complete mitogenomes or partial mitosequences, indicating that the signal of genetic homogeneity of the Armenian gene pools in geographically diverse groups is quite strong. The most likely explanation of this fact might a cultural practise of patrilocality, the ethno-cultural phenomenon when after a marriage a female migrates to the residence of the husband, thus on the population level making females more mobile between geographic locations, which consequently brings to the homogenization of the matrilineal gene pools in different geographic regions.

Using methods of multivariate analysis we have shown that Armenian groups display close genetic affinity with each other and the Armenian cluster occupies an intermediate position between the Levant, the Caucasus and European populations, thus indicating an autochthonous origin of the Armenians on the area of the Armenian Plateau.

In order to define the location of Armenians on the genetic landscape of the Near East and World, in our work we have compiled a large comparative dataset comprising more than a hundred populations from different geographic regions and have performed different types of comparative analysis between Armenians and those datasets.

Correspondence analysis based on haplogroup frequencies has demonstrated that populations from the same geographic region, i.e. Asians, Africans, Europeans, Middle Easterners, etc, form different tight clusters, which supports the idea that the rate of genetic affinity based on the mtDNA is well correlated with geographic distances. Armenians on this plot, as expected, form a distinct group, which is located

within a larger cluster of populations from the Middle East, Caucasus and Europe. Thus, our analysis has shown that Armenians are occupying an intermediate position between aforementioned regions, which once again indicate an autochthonous origin of the Armenian populations. Moreover, it should be noted that aDNA samples have also clustered tightly with present day Armenians, showing their high genetic proximity.

Further, we have elucidated the demographic history of all the Armenian populations that were represented by complete mitogenomes, via constructing Bayesian skyline plots which reflect the changes of *effective population size* through time. As soon as in this analysis we have used our ancient mitogenomes as calibration timepoints for coalescence simulations, the resulted plots are highly accurate and robust. Moreover, using the program that was written in-house in Python programming language, we have partitioned mitogenomes into several regions and assigned different mutation rates for each of them, thus strengthening the accuracy and likelihood of the obtained results.

The results show that the patterns of demographic changes are very similar in all the Armenian groups – there are four major demographic events: population expansion ca 40 kya, constant size until 20 kya, intensive population expansion ca 15 kya, slow decline until current time. Obtained results indicate also that Armenian Highland was a refugial zone during the Last Glacial Maximum, and that population expansion ca 15 kya might be also explained by the raise of agriculture. Notable that when performing the BSP analysis for the aggreagated population we have obtained a similar result indicating that on the whole the territory of the Armenian Higland the demograhic histories of populations were very similar and one can argue that indeed, as was shown in other analyses above, these similarities point to the fact that the analysed populations belong to a one reproductively isolated (during many centuries, even millennia) group.

To investigate the genetic and demographic history of the Armenians further, we have applied coalescence simulation modelling techniques paired with

Approximate Bayesian computations to simulate different demographic scenarios for the Armenians and evaluate which one of them characterises current mitochondrial genetic composition best, and thus can be considered as the most likely demographic history of our population. On the other hand, one of the major aims of this analysis was to conclusively characterise relationships between ancient and modern Armenian data we had.

The results of simulation modelling analysis strikingly demonstrate, that among 5 alternative models of Armenian demographic history, the most plausible one consideres the model showing that there is a genetic continuity between modern days Armenians and ancient DNA samples, showing that, indeed, the ancient samples represent direct ancestors of modern Armenian populations, evidently highlighting that Armenian maternal genetic history is at least 3,500-4,000 years old.

Also, in our dissertation we have developed a new database specifically for human mitochondrial DNA, which we have called mtMart (www.mtmart.org), that has advantages of previously published and reported similar databases, however lacks their drawbacks and has many other convenient functional characteristics that will enable researchers to compile, parse and further analyze human large-scale mitogenomic data more efficiently and accurately.

INFERENCES

1. Matrilineal genetic pool of the Armenians from different geographic regions of the Armenian Highland is very complex and diverse. It is represented by numerous large lineages and sublineages that are present in high concentration in European and Middle Eastern populations. On the other hand, Armenian mitochondrial gene pool does not contain Asian-specific haplogroups, which reflects absence of Asian influence of the Armenian genetic structure, despite numerous migrations through the Armenian Highland in the past millennium.

2. All Armenian geographic groups show high level of genetic diversity, which is comparable with those of the Levant populations, thus strengthening the hypothesis of autochthonous origin of the Armenians.

3. Despite high complexity and diversity of the Armenian maternal genepools, Armenians from different geographic regions in majority do not differ significantly from each other, which might be explained by the cultural practice of patrilocality in Armenians.

4. On the genetic landscape of the region Armenians are located in intermediate position between the Middle Eastern and the Caucasus populations, indicating indigenous origin of our population on the territory of the Armenian Highland.

5. Analysis of demographic changes in of all the Armenian groups studied revealed a very similar demographic history for all the groups and demonstrated that Armenian Highland was a refugial zone for human populations during the Last Glacial Maximum.

6. Applying simulation modeling analysis we have reconstructed the most likely model of the Armenian matrilineal population history, which shows the presence

of genetic continuity between ancient and modern Eastern Armenians samples starting from Eneolithic period.

7. In our work, we have designed and developed a new publicly available database for complete human mitogenomes – mtMART, and integrated new bioinformatics tools in it to facilitate and enhance the efficacy of large-scale human mtDNA studies.

REFERENCES

1. Abu-Amero K., González A., Larruga J, Bosley T. and Cabrera V. Eurasian and African mitochondrial DNA influences in the Saudi Arabian population. BMC evolutionary biology. 2007; 7(1), pp.32.

2. Achilli A., Olivieri A., Pala, M. Metspalu, E., Fornarino S., Battaglia V., Accetturo M., Kutuev I., Khusnutdinova E., Pennarun E. and Cerutti N. Mitochondrial DNA variation of modern Tuscans supports the near eastern origin of Etruscans. The American Journal of Human Genetics. 2007; 80(4), pp.759-768.

3. Achilli, A., Rengo, C., Magri, C., Battaglia, V., Olivieri, A., Scozzari, R., Cruciani, F., Zeviani, M., Briem, E., Carelli, V. and Moral, P., 2004. The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. The American Journal of Human Genetics. 2007; 75(5), pp.910-918.

4. Akçar N., Yavuz V., Ivy-Ochs S., Kubik P., Vardar M., & Schlüchter C. Paleoglacial Records from Kavron Valley, NE Turkey: Field and Cosmogenic Exposure Dating Evidence. Quaternary International. 2007; 164: pp. 170-183.

5. Álvarez-Iglesias V., Mosquera-Miguel A., Cerezo M., Quintáns B., Zarrabeitia T., Cuscó I., Lareu M., García Ó., Pérez-Jurado L., Carracedo Á. and Salas A. New population and phylogenetic features of the internal variation within mitochondrial DNA macro-haplogroup R0. PloS one. 2009; 4(4), p.e5112.

6. Al-Zahery N., Pala M., Battaglia V., Grugni V., Hamod M., Kashani B., Olivieri A., Torroni A., Santachiara-Benerecetti A. and Semino O. In search of the genetic footprints of Sumerians: a survey of Y-chromosome and mtDNA variation in the Marsh Arabs of Iraq. BMC evolutionary biology. 2011; 11(1), pp.1.

7. Anderson S., Bankier A., Barrell B., de Bruijn M., Coulson A., Drouin J., Eperon I., Nierlich D., Roe B., Sanger F., Schreier P., Smith A., Staden R.,

Young I.G. Sequence and organization of the human mitochondrial genome. Nature. 1981; 290(5806): pp.457-465.

8. Anderson S., Bankier A., Barrell B., De Bruijn, M., Coulson A., Drouin J., Eperon I., Nierlich D., Roe B., Sanger F. and Schreier P. Sequence and organization of the human mitochondrial genome. Nature. 1981; 457-465.
9. Arslanov Kh., Dolukhanov P. and Gei N. Climate, Black Sea levels and human settlements in Caucasus Littoral 50,000–9000BP. Quaternary international 167. 2007; pp. 121-127.
10. Attimonelli M., Accetturo M., Santamaria M., Lascaro D., Scioscia G., Pappadà G., Russo L., Zanchetta L. & Tommaseo-Ponzetta M. HmtDB, a human mitochondrial genomic resource based on variability studies supporting population genetics and biomedical research. BMC Bioinformatics. 2005; 6(Suppl 4), S4.
11. Avise J., Arnold J., Ball R., Bermingham E., Lamb T., Neigel J., Reeb C. and Saunders N. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. Annual review of ecology and systematics. 1987; 18(1), pp.489-522.
12. Avise J., Neigel J. and Arnold J. Demographic influences on mitochondrial DNA lineage survivorship in animal populations. Journal of Molecular Evolution. 1984; 20(2), pp.99-105.
13. Bandelt H., Richards M., Macaulay V. Human Mitochondrial DNA and the Evolution of Homo Sapiens. 2006.
14. Barnard H., Dooley A., Areshian G., Gasparyan B. and Faull K. Chemical evidence for wine production around 4000 BCE in the Late Chalcolithic Near Eastern highlands. Journal of Archaeological Science. 2011; 38(5), pp.977-984.
15. Beaumont M., Cornuet J., Marin J. and Robert C.P. Adaptive approximate Bayesian computation. Biometrika. 2009; 96(4), pp.983-990.

16. Beaumont M., Zhang W. and Balding D. Approximate Bayesian computation in population genetics. Genetics. 2002; 162(4), pp.2025-2035.
17. Behar D., Metspalu E., Kivisild T., Achilli A., Hadid Y., Tzur S., Pereira L., Amorim A., Quintana-Murci L., Majamaa K. and Herrnstadt, C. The matrilineal ancestry of Ashkenazi Jewry: portrait of a recent founder event. The American Journal of Human Genetics. 2006; 78(3), pp.487-497.
18. Behar D., Metspalu E., Kivisild T., Rosset S., Tzur S., Hadid Y., Yudkovsky G., Rosengarten D., Pereira L., Amorim A. and Kutuev I. Counting the founders: the matrilineal genetic ancestry of the Jewish Diaspora. Plos one. 2008; 3(4), p.e2062.
19. Behar D., van Oven M., Rosset S., Metspalu M., Loogväli E., Silva N., Kivisild T., Torroni A., &Villems R. A "Copernican" reassessment of the human mitochondrial DNA tree from its root. American Journal of Human Genetics. 2012; 90(4), 675-684.
20. Behar D., Yunusbayev B., Metspalu M., Metspalu E., Rosset S., Parik J., Rootsi S., Chaubey G., Kutuev I., Yudkovsky G. and Khusnutdinova E. The genome-wide structure of the Jewish people. Nature. 2010; 466(7303), pp.238-242.
21. Bekada A., Fregel R., Cabrera V., Larruga J., Pestano J., Benhamamouch S. and González A. Introducing the Algerian mitochondrial DNA and Y-chromosome profiles into the North African landscape. PLoS One. 2013; 8(2), p.e56775.
22. Brandon M., Baldi P., &Wallace D. Mitochondrial mutations in cancer. Oncogene. 2006; 25(34), 4647-4662.
23. Brandstätter A., Niederstätter H., & Parson W. Monitoring the inheritance of heteroplasmy by computer-assisted detection of mixed basecalls in the entire human mitochondrial DNA control region. International Journal of Legal Medicine. 2004;118(1), 47-54.

24. Brandstätter A., Peterson C., Irwin J., Mpoke S., Koech D., Parson W. and Parsons T. Mitochondrial DNA control region sequences from Nairobi (Kenya): inferring phylogenetic parameters for the establishment of a forensic database. International journal of legal medicine. 2004; 118, no. 5 : 294-306.
25. Brosius F., 2011. SPSS 19. MITP-Verlags GmbH & Co. KG.
26. Brown W., George M. and Wilson A. Rapid evolution of animal mitochondrial DNA. Proceedings of the National Academy of Sciences. 1979; 76(4), pp.1967-1971.
27. Brown W., George M., & Wilson A. Rapid evolution of animal mitochondrial DNA. Proceedings of the National Academy of Sciences. 1979; 76(4), 1967-1971.
28. Burckhardt F., von Haeseler A., & Meyer S. HvrBase: compilation of mtDNA control region sequences from primates. Nucleic Acids Research. 1999; 27(1), 138-142.
29. Cann R., and Wilson A. Length mutations in human mitochondrial DNA. Genetics. 1983; 104, no. 4: 699-711.
30. Cann R., Stoneking M., Wilson A. Mitochondrial DNA and human evolution. Nature. 1987; 325:31–36.
31. Cavalli-Sforza L. Luca, and Feldman M. The application of molecular genetics approaches to the study of human evolution. Nature genetics. 2003; 33: 266-275.
32. Cavalli-Sforza L. The DNA revolution in population genetics. Trends in Genetics. 1998; 14(2), pp.60-65.
33. Csillery K., François O. and Blum M. abc: an R package for approximate Bayesian computation (ABC). Methods in Ecology and Evolution. 2012; 3(3), pp.475-479.
34. Delfin F., Ko A., Li M., Gunnarsdóttir E., Tabbada K., Salvador J., Calacal G., Sagum M., Datar F., Padilla S. and De Ungria M. Complete mtDNA genomes of Filipino ethnolinguistic groups: a melting pot of recent and ancient lineages

in the Asia-Pacific region.European Journal of Human Genetics. 2014; 22(2), pp.228-237.

35. Derenko M., Malyarchuk B., Bahmanimehr A., Denisova G., Perkova M., Farjadian S., & Yepiskoposyan L. Complete mitochondrial DNA diversity in Iranians. PLoSOne. 2013. 14;8(11):e8067
36. Derenko M., Malyarchuk B., Denisova G., Perkova M., Litvinov A., Grzybowski T., Dambueva I., Skonieczna K., Rogalla U., Tsybovsky I. & Zakharov, I. Western Eurasian ancestry in modern Siberians based on mitogenomic data. BMC Evolutionary Biology. 2014; 14(1), 217.
37. Derenko M., Malyarchuk B., Grzybowski T., Denisova G., Dambueva I., Perkova M., Dorzhu C., Luzina F., Lee H., Vanecek T. and Villems R. Phylogeographic analysis of mitochondrial DNA in northern Asian populations. The American Journal of Human Genetics. 2007; 81(5), pp.1025-1041.
38. Dolukhanov P., Aslanyan S., Kolpakov E., and Belyayeva E. Prehistoric sites in northern Armenia. Antiquity. 2004. http://antiquity.ac.uk./projgall/dolukhanov/index.html.
39. Drummond A. and Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC evolutionary biology. 2007; 7(1), p.1.
40. Drummond A., Rambaut A., Shapiro B. & Pybus O. Bayesian coalescent inference of past population dynamics from molecular sequences. Molecular biology and evolution. 2005; 22(5), pp.1185-1192.
41. Drummond A., Suchard M., Xie D. & Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular biology and evolution. 2012; 29(8), pp.1969-1973.
42. Duggan A., Evans B., Friedlaender F., Friedlaender J., Koki G., Merriwether D., Kayser M. & Stoneking M. Maternal history of Oceania from complete mtDNA genomes: contrasting ancient diversity with recent homogenization

due to the Austronesian expansion. The American Journal of Human Genetics. 2014; 94(5), 721-733.

43. Dunning J., Stewart D., Danielson B., Noon B., Root T., Lamberson R. and Stevens E. Spatially explicit population models: current forms and future uses. Ecological Applications. 1995; 5(1), pp.3-11.
44. Dupanloup I., Schneider S. and Excoffier L. A simulated annealing approach to define the genetic structure of populations. Molecular Ecology. 2002; 11(12), pp.2571-2581.
45. Eltsov N., Volodko N., Starikovskaya E. & Sukernik R. New tool (mtPHYL) proposed for phylogenetic analysis of human complete mitochondrial genomes. 2011.
46. Estoup A., Lombaert E., Marin J., Guillemaud T., Pudlo P., Robert C.P. and Cornuet J. Estimation of demo-genetic model probabilities with Approximate Bayesian Computation using linear discriminant analysis on summary statistics. Molecular ecology resources. 2012; 12(5), pp.846-855.
47. Evanno G., Regnaut S. and Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular ecology. 2005; 14(8), pp.2611-2620.
48. Excoffier L. and Lischer H. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows.Molecular ecology resources. 2010; 10(3), pp.564-567.
49. Excoffier L., Dupanloup I., Huerta-Sánchez E., Sousa V. and Foll M. Robust demographic inference from genomic and SNP data. PLoS Genet. 2013; 9(10), p.e1003905.
50. Fadhlaoui-Zid K., Rodríguez-Botigué L., Naoui N., Benammar-Elgaaied A., Calafell F. & Comas D. Mitochondrial DNA structure in North Africa reveals a genetic discontinuity in the Nile Valley. American journal of physical anthropology. 2011; 145(1), pp.107-117.

51. Fan L. and Yao Y. An update to MitoTool: using a new scoring system for faster mtDNA haplogroup determination. Mitochondrion. 2013; 13(4), pp.360-363.
52. Felsenstein J. PHYLIP-phylogeny inference package (version 3.2). Cladistics. 1989;5:163-6.
53. Finilla S., Lehtonen M. and Majamaa K. Phylogenetic network for European mtDNA. The American Journal of Human Genetics. 2001; 68(6), pp.1475-1484.
54. Fregel R. and Delgado S. HaploSearch: a tool for haplotype-sequence two-way transformation. Mitochondrion. 2011; 11(2), pp.366-367.
55. Gandilyan P. Archaeobotanical evidence for evolution of cultivated wheat and barley in Armenia. Proceedings of the Harlan Symposium "The Origins of Agriculture and the Domestication of Crop Plants in the Near East". 10-14 May, 1997, Aleppo, Syria. ICARDA. 1998; 280-285.
56. Guillemaud T., Beaumont M., Ciosi M., Cornuet J. and Estoup A. Inferring introduction routes of invasive species using approximate Bayesian computation on microsatellite data. Heredity. 2010; 104(1), pp.88-99.
57. Gunnarsdóttir E., Li M., Bauchet M., Finstermeier K., & Stoneking, M. High-throughput sequencing of complete human mtDNA genomes from the Philippines. Genome research. 2011; 21(1), 1-11.
58. Gunnarsdóttir E., Nandineni M., Li M., Myles S., Gil D., Pakendorf B. and Stoneking M. Larger mitochondrial DNA than Y-chromosome differences between matrilocal and patrilocal groups from Sumatra. Nature communications. 2011; 2, p.228.
59. Hagström E., Freyer C., Battersby B., Stewart J., & Larsson N. No recombination of mtDNA after heteroplasmy for 50 generations in the mouse maternal germline. Nucleic acids research. 2014; 42(2), 1111-1116.

60. Hartl D., Clark A. and Clark A. Principles of population genetics. 1997; (Vol. 116). Sunderland: Sinauer associates.
61. Harutyunyan A., Khudoyan A., & Yepiskoposyan L. Patrilocality and Recent Migrations Have Little Impact on Shaping Patterns of Genetic Structure of the Armenian Population. Russian Journal of Genetics. 2009; 45(8): 987-993.
62. Hauswirth W. and Clayton D. Length heterogeneity of a conserved displacement-loop sequence in human mitochondrial DNA. Nucleic acids research. 1985; 13, no. 22: 8093-8104.
63. Herrera K., Lowery R., Hadden L., Calderon S., Chiou C., Yepiskoposyan L., Regueiro M., Underhill P. & Herrera R. Neolithic Patrilineal Signals Indicate that the Armenian Plateau was Repopulated by Agriculturalists. European Journal of Human Genetics. 2012; 20: 313-320.
64. Hoban S., Bertorelle G., & Gaggiotti O. Computer simulations: tools for population and evolutionary genetics. Nature Reviews Genetics. 2012;13(2), 110-122.
65. Hovhannisyan A., Khachatryan Z., Haber M., Hrechdakian P., Karafet T., Zalloua P. & Yepiskoposyan L. Different waves and directions of Neolithic migrations in the Armenian Highland. Investigative genetics. 2014; 5(1), 15.
66. Hovsepyan R. & Willcox G. The Earliest Finds of Cultivated Plants in Armenia: Evidence from Charred Remains and Crop Processing Residues in Pisé from the Neolithic Settlements of Aratashen and Aknashen. Vegetation History and Archaeobotany. 2008; 17(1): 63-71.
67. Hudson R. Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics. 2002; 18(2), 337-338.
68. Hudson R., & Kaplan N. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics. 1985; 111(1), 147-164.
69. Hutchison C., Newbold J., Potter S. & Edgell M. Maternal inheritance of mammalian mitochondrial DNA. Nature. 1974; 251(5475):536-8.

70. Ingman M., & Gyllensten U. mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences. Nucleic Acids Research. 2006; 34(suppl 1), D749-D751.
71. Irwin J., Egyed B., Saunier J., Szamosi G., O'Callaghan J., Padar Z. and Parsons T. Hungarian mtDNA population databases from Budapest and the Baranya county Roma. International journal of legal medicine. 2007; 121(5), pp.377-383.
72. Irwin J., Ikramov A., Saunier J., Bodner M., Amory S., Röck A., O'Callaghan J., Nuritdinov A., Atakhodjaev S., Mukhamedov R. and Parson W. The mtDNA composition of Uzbekistan: a microcosm of Central Asian patterns. International journal of legal medicine. 2010; 124(3), pp.195-204.
73. Irwin J., Saunier J., Strouss K., Diegoli T., Sturk K., O'Callaghan J., Paintner C., Hohoff C., Brinkmann B. and Parsons T. Mitochondrial control region sequences from a Vietnamese population sample. International journal of legal medicine. 2008. 122(3), pp.257-259.
74. Karachanak S., Carossa V., Nesheva D., Olivieri A., Pala M., Kashani B., Grugni V., Battaglia V., Achilli A., Yordanov Y. and Galabov A. Bulgarians vs the other European populations: a mitochondrial DNA perspective. International journal of legal medicine. 2012; 126(4), pp.497-503.
75. Kartal M. Anatolian epi-paleolithic period assemblages: problems, suggestions, evaluations and various approaches. Anatolia 24. 2003: 45-62.
76. Kasperaviciute D., Kučinskas V. and Stoneking M. Y chromosome and mitochondrial DNA variation in Lithuanians. Annals of human genetics. 2004;68(5), pp.438-452.
77. Katoh K. and Standley D. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution. 2013; 30(4), pp.772-780.

78. King T, Fernandez-Jalvo Y, Moloney N, Andrews P, Melkonyan A, Ditchfield P, Yepiskoposyan L, Karapetyan S. Exploration and Survey of Pleistocene Hominid Sites in Armenia and Karabagh. Antiquity. 2003; 77(295).

79. Kingman J. Origins of the coalescent: 1974-1982. Genetics. 2000; 156 (4), 1461-1463.

80. Kingman J. The coalescent. Stochastic processes and their applications. 1982; 13(3), 235-248.

81. Kivisild T. Maternal ancestry and population history from whole mitochondrial genomes. Investigative genetics. 2015; 6(1), p.1.

82. Kloss-Brandstätter A., Pacher D., Schönherr S., Weissensteiner H., Binna R., Specht G. and Kronenberg F. HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. Human mutation. 2011; 32(1), pp.25-32.

83. Kohl J., Paulsen I., Laubach T., Radtke A. & von Haeseler, A. HvrBase++: a phylogenetic database for primate species. Nucleic acids research. 2006; 34(suppl 1), D700-D704.

84. Kong Q., Yao Y., Liu M., Shen S., Chen C., Zhu C., Palanichamy M. and Zhang Y. Mitochondrial DNA sequence polymorphisms of five ethnic populations from northern China. Human genetics. 2003; 113(5), pp.391-405.

85. Kuhner M., Yamato J. and Felsenstein J. Maximum likelihood estimation of population growth rates based on the coalescent. Genetics. 1998; 149(1), pp.429-434.

86. Kushnareva K. The Southern Caucasus in Prehistory: Stages of Cultural and Socioeconomic Development from the Eighth to the Second Millennium BC. University of Pennsylvania Museum. 1997.

87. Leuenberger C. and Wegmann D. Bayesian computation and model selection without likelihoods. Genetics. 2010; 184(1), pp.243-252.

88. Librado P. and Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009; 25(11), pp.1451-1452.

89. Macaulay V., Richards M., Hickey E., Vega E., Cruciani F., Guida V., Scozzari R., Bonné-Tamir B., Sykes B. & Torroni A. The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. The American Journal of Human Genetics. 1999; 64(1), 232-249.

90. Malaspinas A., Westaway M., Muller C., Sousa V., Lao O., Alves I., Bergström A., Athanasiadis G., Cheng J., Crawford J., Heupink T. A genomic history of Aboriginal Australia. Nature. 2016.

91. Mallick S., Li H., Lipson M., Mathieson I., Gymrek M., Racimo F., Zhao M., Chennagiri N., Nordenfelt S., Tandon A., Skoglund P., Lazaridis I., Sankararaman S., Fu Q., Rohland N., Renaud G., Erlich Y., Willems T., Gallo C., Spence J., Song Y., Poletti G., Balloux F., van Driem G., de Knijff P., Romero I., Jha A., Behar D., Bravi C., Capelli C., Hervig T., Moreno-Estrada A., Posukh O., Balanovska E., Balanovsky O., Karachanak-Yankova S., Sahakyan H., Toncheva D., Yepiskoposyan L., Tyler-Smith C., Xue Y., Abdullah M., Ruiz-Linares A., Beall C., Di Rienzo A., Jeong C., Starikovskaya E., Metspalu E., Parik J., Villems R., Henn B., Hodoglugil U., Mahley R., Sajantila A., Stamatoyannopoulos G., Wee J., Khusainova R., Khusnutdinova E., Litvinov S., Ayodo G., Comas D., Hammer M., Kivisild T., Klitz W., Winkler C., Labuda D., Bamshad M., Jorde L., Tishkoff S., Watkins W., Metspalu M., Dryomov S., Sukernik R., Singh L., Thangaraj K., Pääbo S., Kelso J., Patterson N., Reich D. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature. 206; 538(7624), pp.201-206.

92. Malyarchuk B., Derenko M., Denisova G. and Kravtsova O. Mitogenomic diversity in Tatars from the Volga-Ural region of Russia. Molecular biology and evolution. 2010; 27(10), pp.2220-2226.

93. Malyarchuk B., Grzybowski T., Derenko M., Czarny J., Drobnič K. and Miścicka-Śliwka D. Mitochondrial DNA variability in Bosnians and Slovenians. Annals of human genetics. 2003; 67(5), pp.412-425.

94. Malyarchuk B., Grzybowski T., Derenko M., Czarny J., Wozniak M. and Miscicka-Sliwka D. Mitochondrial DNA variability in Poles and Russians. Annals of human genetics. 2002; 66(04), pp.261-283.
95. Malyarchuk B., Grzybowski T., Derenko M., Perkova M., Vanecek T., Lazur J., Gomolcak P. and Tsybovsky I. Mitochondrial DNA phylogeny in eastern and western Slavs. Molecular biology and evolution. 2008;25(8), pp.1651-1658.
96. Marin J., Pudlo P., Robert C. and Ryder R. Approximate Bayesian computational methods. Statistics and Computing. 2012; 22(6), pp.1167-1180.
97. Matevosyan L., Chattopadhyay S., Madelian V., Avagyan S., Nazaretyan M., Hyussian A., Vardapetyan E., Arutunyan R. and Jordan F. HLA-A, HLA-B, and HLA-DRB1 allele distribution in a large Armenian population sample. Tissue Antigens. 2011. 78, no. 1: 21-30.
98. Merriwether D., Clark A., Ballinger S., Schurr T., Soodyall H., Jenkins T., Sherry S. & Wallace D. The structure of human mitochondrial DNA variation. Journal of Molecular Evolution. 1991; 33(6), 543-555.
99. Michaels G., Hauswirth W., & Laipis P. Mitochondrial DNA copy number in bovine oocytes and somatic cells. Developmental biology. 1982; 94(1), 246-251.
100. Mielnik-Sikorska M., Daca P., Malyarchuk B., Derenko M., Skonieczna K., Perkova M., Dobosz T. and Grzybowski T. The history of Slavs inferred from complete mitochondrial genome sequences. PloS one. 2013; 8(1), p.e54360.
101. Mikkelsen M., Fendt L., Röck A.W., Zimmermann B., Rockenbauer E., Hansen A.J., Parson W. and Morling N. Forensic and phylogeographic characterisation of mtDNA lineages from Somalia.International journal of legal medicine. 2012; 126(4), pp.573-579.
102. Miller M.A., Pfeiffer W. and Schwartz T. November. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. InGateway Computing Environments Workshop (GCE), 2010; pp. 1-8. IEEE.

103. Nasidze I, Ling E, Quinque D, Dupanloup I, Cordaux R, Rychkov S, Naumova O, Zhukova O, Sarraf-Zadegan N, Naderi GA, Asgary S, Sardas S, FarhudDD,Sarkisian T, Asadov C, Kerimov A, Stoneking M. Mitochondrial DNA and Y-chromosome variation in the Caucasus. Annals of human genetics. 2004; 68(3), 205-221.
104. Nasidze I., & Stoneking M. Mitochondrial DNA variation and language replacements in the Caucasus. Proceedings of the Royal Society of London. Series B: Biological Sciences. 2001; 268(1472), 1197-1206.
105. Nasidze I., Sarkisian T., Kerimov A., & Stoneking M. Testing Hypotheses of Language Replacement in the Caucasus: Evidence from the Y-Chromosome. Human Genetics. 2003; 112(3): 255-261.
106. Nei M. Molecular evolutionary genetics. 1987. Columbia university press.
107. Nunes M. and Balding D. On optimal selection of summary statistics for approximate Bayesian computation. Statistical applications in genetics and molecular biology. 2010; 9(1), p.34.
108. O'Fallon B. & Fehren-Schmitz L. Native Americans experienced a strong population bottleneck coincident with European contact. Proceedings of the National Academy of Sciences. 2011; 108(51), 20444-20448.
109. Pagani L, Lawson D,Jagoda E, Mörseburg A, Eriksson A, Mitt M, Clemente F, Hudjashov G, DeGiorgio M, Saag L, Wall J, Cardona A, Mägi R, Sayres M, Kaewert S, Inchley C, Scheib C, Järve M, Karmin M, Jacobs G, Antao T, Iliescu F, Kushniarevich A, Ayub Q, Tyler-Smith C, Xue Y, Yunusbayev B, Tambets K, Mallick C, Saag L, Pocheshkhova E, Andriadze G, Muller C, Westaway M, Lambert D, Zoraqi G, Turdikulova S, Dalimova D, Sabitov Z, Sultana G, Lachance J, Tishkoff S, Momynaliev K, Isakova J, Damba L, Gubina M, Nymadawa P, Evseeva I, Atramentova L, Utevska O, Ricaut F, Brucato N, Sudoyo H, Letellier T, Cox M, Barashkov N, Škaro V, Mulahasanovic L, Primorac D, Sahakyan H, Mormina M, Eichstaedt C, Lichman D, Abdullah S, Chaubey G, Wee J, Mihailov E, Karunas A, Litvinov

S, Khusainova R, Ekomasova N, Akhmetova V, Khidiyatova I, Marjanović D, Yepiskoposyan L, Behar D, Balanovska E, Metspalu A, Derenko M, Malyarchuk B, Voevoda M, Fedorova S, Osipova L, Lahr M, Gerbault P, Leavesley M, Migliano A, Petraglia M, Balanovsky O, Khusnutdinova E, Metspalu E, Thomas M, Manica A, Nielsen R, Villems R, Willerslev E, Kivisild T, Metspalu M. Genomic analyses inform on migration events during the peopling of Eurasia. Nature. 2016; 538(7624), pp.238-242.

110. Pakendorf B., Wiebe V., Tarskaia L.A., Spitsyn V.A., Soodyall, H., Rodewald A. and Stoneking M. Mitochondrial DNA evidence for admixed origins of central Siberian populations. American journal of physical anthropology. 2003; 120(3), pp.211-224.
111. Pala M., Olivieri A., Achilli A., Accetturo M., Metspalu E., Reidla M., Tamm E., Karmin M., Reisberg T., Kashani B. and Perego U. Mitochondrial DNA signals of late glacial recolonization of Europe from near eastern refugia. The American journal of human genetics. 2012; 90(5), pp.915-924.
112. Parson W. & Dür A. EMPOP—a forensic mtDNA database.Forensic Science International: Genetics. 2007; 1(2), 88-92.
113. Pikó L., & Matsumoto L. Number of mitochondria and some properties of mitochondrial DNA in the mouse egg. Developmental biology. 1976: 49(1), 1-10.
114. Pinhasi R., Gasparian B., Areshian G., Zardaryan D., Smith A., Bar-Oz G. and Higham T. First direct evidence of chalcolithic footwear from the near eastern highlands. PloS one. 2010; 5, no. 6: e10984.
115. Pinhasi R., Gasparian D., Wilkinson K., Bailey R., Bar-Oz G., Bruch A., Chataigner C. Hovk 1 and the Middle and Upper Paleolithic of Armenia: a preliminary framework. Journal of Human Evolution. 2008; 55, no. 5: 803-816.
116. Pliss L., Tambets K., Loogväli E., Pronina N., Lazdins M., Krumina A., Baumanis V. and Villems R. Mitochondrial DNA Portrait of Latvians:

Towards the Understanding of the Genetic Structure of Baltic-Speaking Populations. Annals of human genetics. 2006; 70(4), pp.439-458.

117. Poetsch M., Wittig H., Krause D. and Lignitz E. Mitochondrial diversity of a northeast German population sample. Forensic science international. 2003; 137(2), pp.125-132.
118. Qian Y., Chu Z., Dai Q., Wei C., Chu J., Tajima A. and Horai S. Mitochondrial DNA polymorphisms in Yunnan nationalities in China. Journal of human genetics. 2001; 46(4), pp.211-220.
119. Quintana-Murci L., Chaix R., Wells R., Behar D., Sayar H., Scozzari R., Rengo C., Al-Zahery N., Semino O., Santachiara-Benerecetti A. and Coppa A. Where west meets east: the complex mtDNA landscape of the southwest and Central Asian corridor. The American Journal of Human Genetics. 2004; 74(5), pp.827-845.
120. R core team. 2013. R: A language and environment for statistical computing.
121. Rakha A., Shin K., Yoon J., Kim N., Siddique M., Yang I., Yang W. and Lee H. Forensic and genetic characterization of mtDNA from Pathans of Pakistan. International journal of legal medicine. 2011;125(6), pp.841-848.
122. Rambaut A. and Drummond A. Tracer v1. 4. 2007.
123. Raymond M. and Rousset F. An exact test for population differentiation. Evolution. 1995; 49(6), pp.1280-1283.
124. Redgate A. The Armenians. Blackwell Publishers Ltd, Oxford, 2000.
125. Richard C., Richard C., Pennarun E., Kivisild T., Tambets K., Tolk H., Metspalu E., Reidla M., Chevalier S., Giraudet S. and Lauc L. An mtDNA perspective of French genetic variation. Annals of human biology. 2007;34(1), pp.68-79.
126. Richards M., Macaulay V., Hickey E., Vega E., Sykes B., Guida V., Rengo C., Sellitto D., Cruciani F., Kivisild T., Villems R., Thomas M., Rychkov S., Rychkov O., Rychkov Y., Gölge M., Dimitrov D., Hill E., Bradley D., Romano V., Calì F., Vona G., Demaine A., Papiha S., Triantaphyllidis C.,

Stefanescu G., Hatina J., Belledi M., Di Rienzo A., Oppenheim A., Nørby S., Al-Zaheri N., Santachiara-Benerecetti S., Scozzari R., Torroni A., and Bandelt H. Tracing European founder lineages in the Near Eastern mtDNA pool. The American Journal of Human Genetics. 2000; 67(5), 1251-1276.

127. Rieux A., Eriksson A., Li M., Sobkowiak B., Weinert L., Warmuth V., Ruiz-Linares A., Manica A. and Balloux F. Improved calibration of the human mitochondrial clock using ancient genomes. Molecular biology and evolution. 2014; 31(10), pp.2780-2792.
128. Röck A.W., Dür A., van Oven M. and Parson W. Concept for estimating mitochondrial DNA haplogroups using a maximum likelihood approach (EMMA). Forensic Science International: Genetics. 2013; 7(6), pp.601-609.
129. Ronquist F. and Huelsenbeck J. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 2003; 19(12), pp.1572-1574.
130. Rootsi S., Myres N., Lin A., Järve M., King R., Kutuev I., Cabrera V., Khusnutdinova E., Varendi K., Sahakyan H., Behar D., Khusainova R., Balanovsky O., Balanovska E., Rudan P., Yepiskoposyan L., Bahmanimehr A., Farjadian S., Kushniarevich A., Herrera R. J., Grugni V., Battaglia V., Nici C., Crobu F., Karachanak S., Hooshiar-Kashani B., Houshmand M., Sanati M.H., Toncheva D., Lisa A., Semino O., Chiaroni J., Di Cristofaro J., Villems R., Kivisild T. & Underhill P. Distinguishing the Co-Ancestries of Haplogroup G Y-Chromosomes in the Populations of Europe and the Caucasus. European Journal of Human Genetics. 2012; 20(12): 1275-1282.
131. Rubino F., Piredda R., Calabrese F., Simone D., Lang M., Calabrese C., Petruzzella V., Tommaseo-Ponzetta M., Gasparre G. & Attimonelli M. HmtDB, a genomic resource for mitochondrion-based human variability studies. Nucleic Acids Research. 2012; 40(D1), D1150-D1159.
132. Schaffner S., Foo C., Gabriel S., Reich D., Daly M. and Altshuler, D. Calibrating a coalescent simulation of human genome sequence variation. Genome research. 2005;15(11), pp.1576-1583.

133. Scheible M., Kim S., Sturk-Andreaggi K., Coble M. and Irwin J. Mitochondrial control region variation in a Korean population sample. International journal of legal medicine. 2014; 128(5), pp.745-746.
134. Schönberg A., Theunert C., Li M., Stoneking M. and Nasidze I. High-throughput sequencing of complete human mtDNA genomes from the Caucasus and West Asia: high diversity and demographic inferences. European Journal of Human Genetics. 2011; 19(9), pp.988-994.
135. Shlush L., Behar D., Yudkovsky G., Templeton A., Hadid Y., Basis F., Hammer M., Itzkovitz S. and Skorecki K. The Druze: a population genetic refugium of the Near East. PLoS One. 2008;, 3(5), p.e2105.
136. Slatkin M. Simulating genealogies of selected alleles in a population of variable size. Genetical research. 2001; 78(01), 49-57.
137. Soares P., Ermini L., Thomson N., Mormina M., Rito T., Röhl A., Salas A., Oppenheimer S., Macaulay V. and Richards M. Correcting for purifying selection: an improved human mitochondrial molecular clock. The American Journal of Human Genetics. 2009; 84(6), pp.740-759.
138. Soares P., Rito T., Trejaut J., Mormina M., Hill C., Tinkler-Hundal E., Braid M., Clarke D., Loo J., Thomson N. and Denham T. Ancient voyaging and Polynesian origins. The American Journal of Human Genetics. 2011; 88(2), pp.239-247.
139. Soares P., Trejaut J., Loo J., Hill C., Mormina M., Lee C., Chen Y., Hudjashov G., Forster P., Macaulay V. and Bulbeck D. Climate change and postglacial human dispersals in Southeast Asia. Molecular Biology and Evolution. 2008; 25(6), pp.1209-1218.
140. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006; 22(21), pp.2688-2690.

141. Stoljarova M., King J., Takahashi M., Aaspõllu A. and Budowle B. Whole mitochondrial genome genetic diversity in an Estonian population sample. International journal of legal medicine. 2016; 130(1), pp.67-71.
142. Stoneking M., DNA and recent human evolution. Evolutionary Anthropology: Issues, News, and Reviews. 1993; 2(2), pp.60-73.
143. Sunnaker M., Busetto A., Numminen E., Corander J., Foll M. & Dessimoz C. Approximate bayesian computation. PLoS Comput Biol. 2013; 9(1), e1002803.
144. Tambets K., Rootsi S., Kivisild T., Help H., Serk P., Loogväli E., Tolk H., Reidla M., Metspalu E., Pliss L. and Balanovsky, O. The western and eastern roots of the Saami—the story of genetic "outliers" told by mitochondrial DNA and Y chromosomes. The American Journal of Human Genetics, 2004; 74(4), pp.661-682.
145. Tamura K., Stecher G., Peterson D., Filipski A. and Kumar S. MEGA6: molecular evolutionary genetics analysis version 6.0. Molecular biology and evolution. 2013; 30(12), pp.2725-2729.
146. Tanaka M., Cabrera V., González A., Larruga J., Takeyasu T., Fuku N., Guo L., Hirose R., Fujita Y., Kurata M. and Shinoda K.I. Mitochondrial genome variation in eastern Asia and the peopling of Japan. Genome research. 2004; 14(10a), pp.1832-1850.
147. Tanaka M., Francis A., Luciani F. and Sisson S. Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data. Genetics. 2006; 173(3), pp.1511-1520.
148. Tarasov P., Volkova V., Webb T., Guiot J., Andreev A., Bezusko L., Bezusko T., Bykova G., Dorofeyuk N., Kvavadze E., Osipova I., Panova N., Sevastyanov D. Last Glacial Maximum Biomes Reconstructed from Pollen and Plant Macrofossil Data from Northern Eurasia. Journal of Biogeography. 2000; 27(3):609-620.

149. Tillmar A., Coble M., Wallerström T. and Holmlund G. Homogeneity in mitochondrial DNA control region sequences in Swedish subpopulations. International journal of legal medicine. 2010; 124(2), pp.91-98.
150. Toren D., Barzilay T., Tacutu R., Lehmann G., Muradian K. & Fraifeld V. MitoAge: a database for comparative analysis of mitochondrial DNA, with a special focus on animal longevity. Nucleic acids research. 2015; gkv1187.
151. Torroni A., Rengo C., Guida V., Cruciani F., Sellitto D., Coppa A., Calderon F., Simionati B., Valle G., Richards M. and Macaulay V. Do the four clades of the mtDNA haplogroup L2 evolve at different rates?.The American Journal of Human Genetics. 2001; 69(6), pp.1348-1356.
152. Trejaut J., Kivisild T., Loo J., Lee C., He C., Hsu C., Li Z. and Lin M. Traces of archaic mitochondrial lineages persist in Austronesian-speaking Formosan populations. PLoS Biol. 2005; 3(8), p.e247.
153. Trust L. Genstat Software Version 5 Release 3.2 for Windows NT (Software for Statistical Analysis). 1995. Rothamstead Agricultural Station.
154. van Oven M. PhyloTree Build 17: Growing the human mitochondrial DNA tree. Forensic Science International: Genetics Supplement Series. 2015; 5, e392-e394.
155. van Oven M., Kayser M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat. 2009; 30(2):E386-E394.
156. Vasconcelos A., Guimarães A., Castelletti C., Caruso C., Ribeiro C., Yokaichiya F., Armôa G., Pereira G., da Silva I., Schrago C., Fernandes A., da Silveira A., Carneiro A., Carvalho B., Viana C., Gramkow D., Lima F., Corrêa L., Mudado M., Nehab-Hess P., Souza Rd., Corrêa R., Russo C. MamMiBase: a mitochondrial genome database for mammalian phylogenetic studies. Bioinformatics. 2005; 21(10), 2566-2567.

157. Vianello D., Sevini F., Castellani G., Lomartire L., Capri M., & Franceschi C. HAPLOFIND: A New Method for High-Throughput mtDNA Haplogroup Assignment. Human mutation. 2013; 34(9), 1189-1194.
158. Weale M., Yepiskoposyan L., Jager R., Hovhannisyan N., Khudoyan A., Burbage-Hall O., Bradman N., & Thomas M. Armenian Y Chromosome Haplotypes Reveal Strong Regional Structure Within a Single Ethno-National Group. Human Genetics. 2001; 109(6): 659-6.
159. Wen B., Li H., Gao S., Mao X., Gao Y., Li F., Zhang F., He Y., Dong Y., Zhang Y. and Huang W. Genetic structure of Hmong-Mien speaking populations in East Asia as revealed by mtDNA lineages. Molecular Biology and Evolution. 2005; 22(3), pp.725-734.
160. Wen B., Li H., Lu D., Song X., Zhang F., He Y., Li F., Gao Y., Mao X., Zhang L. and Qian J. Genetic evidence supports demic diffusion of Han culture. Nature. 2004; 431(7006), pp.302-305.
161. Wickham H. ggplot2: elegant graphics for data analysis. 2009. Springer Science & Business Media.
162. Yao Y., Kong Q., Wang C., Zhu C. and Zhang Y. Different matrilineal contributions to genetic structure of ethnic groups in the silk road region in china. Molecular Biology and Evolution. 2004; 21(12), pp.2265-2280.
163. Yates A., Akanni W., Amode M., Barrell D., Billis K., Carvalho-Silva D., Cummins C., Clapham P., Fitzgerald S., Gil L. and Girón C. Ensembl 2016. Nucleic Acids Research. 2016; 44(D1), pp.D710-D716.
164. Yuan X., Miller D., Zhang J., Herrington D. & Wang Y. An overview of population genetic data simulation. Journal of Computational Biology. 2012;, 19(1), 42-54.
165. Yunusbayev B., Metspalu M., Järve M., Kutuev I., Rootsi S., Metspalu E., Behar DM., Varendi K., Sahakyan H., Khusainova R., Yepiskoposyan L., Khusnutdinova EK., Underhill PA., Kivisild T., Villems R. The Caucasus as

an asymmetric semipermeable barrier to ancient human migrations. Molecular biology and evolution. 2012; 29(1), 359-365.

166. Zheng H., Yan S., Qin Z. & Jin L. MtDNA analysis of global populations support that major population expansions began before Neolithic Time. Scientific Reports. 2012; 2:745

167. Zimmermann B., Bodner M., Amory S., Fendt L., Röck A., Horst D., Horst B., Sanguansermsri T., Parson W. and Brandstätter A. Forensic and phylogeographic characterization of mtDNA lineages from northern Thailand (Chiang Mai). International journal of legal medicine. 2009; 123(6), pp.495-501.