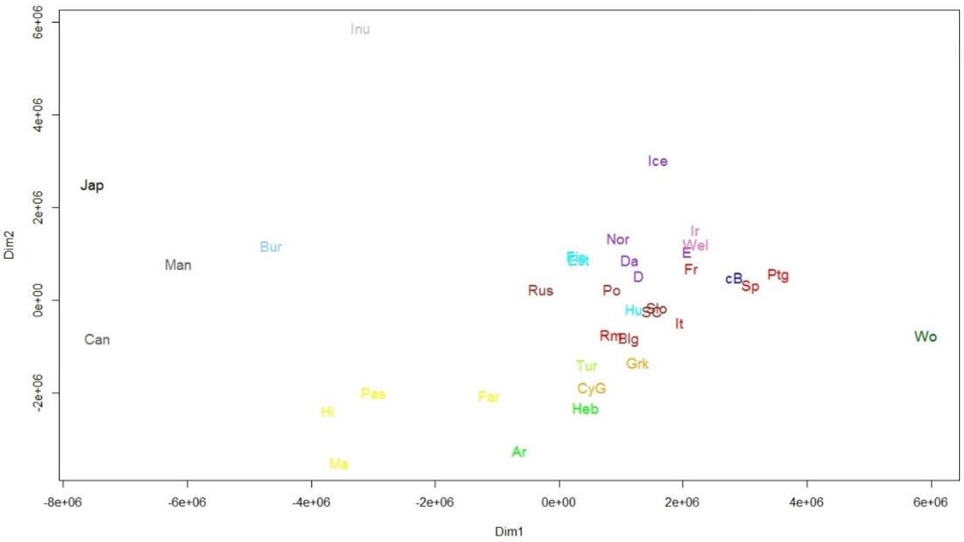

Y-chr and mtDNA diversity in the context of Eurasian language diversity

Daniel Filipe da Silva Pereira

Masters in Biodiversity, Genetics and Evolution

Department of Biology

2015

Supervisor

Guido Barbujani, Professor, University of Ferrara, Italy

Co-supervisor

Roberta Susca, Doctorate student, University of Ferrara, Italy

Co-supervisor

Jorge Rocha, Professor, University of Porto, Portugal

U.PORTO
FC FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

All the corrections determined by the jury, and only those, were made.

The President of the Jury,

Oporto, ______/______/_________

“Nature did not make it easy for those who want to classify species still in their making”

Jean-Jacques Hublin

Acknowledgments

This work would not be possible without the help of a vast group of people, and I would like to leave here my sincere acknowledgements for their help.

To Professor Guido Barbujani, without whom this thesis would not be possible to begin with, for accepting me in is laboratory and to be my supervisor during my thesis, and for all the help he provided within and beyond my thesis.

To Roberta Susca, my co-supervisor, for accepting me to work with her and teach me the “how to” of many things and always have patience to answer my many questions.

To the laboratory of Population Genetics, Conservation and Evolution in the University of Ferrara for welcoming me in their laboratory, series and movie discussions, and off course, the pizza lunches.

To Professor Jorge Rocha for being my supervisor and connection to the Faculty of Sciences of the University of Porto.

To the ESN – Ferrara for welcoming me to Ferrara, presenting me to so many incredible people and make my stay abroad easier.

Last but not least, would like to thank my family, friends (old and new ones) and girlfriend Sabrina for the support they gave me throughout my stay aboard, and especially to my parents for making this journey possible.

To all, my deepest and sincere gratitude!

A todos, a minha mais profunda e sincera gratidão!

A tutti, la mia più profonda e sincera gratitudine!

Abstract/Resumo

Taking in mind the hypothesis first stated by Darwin over the possibility of genomic and linguistic phylogenetic trees sharing the trace of human demographic history, the LanGeLin project is focused on uncover how much of this is true. Making use of new approaches in linguistic classifications, such as the use of syntax and the PCM methodology, and the vast availability of genomic information for our species, such has HVR 1 sequence data, SNPs and STRs markers, this study focused on uncovering the relationship between genomic and linguistic diversity for paternal and maternal lineages within the context of Eurasian languages.

The results of the analysis performed here showed statistically significant correlation between genomic and linguistic diversity with the geographical distance, demonstrating that a clear geographical structure is evident and both genomes and languages tend to be closer to their geographical neighbours. A clear correlation between both genomic and linguistic traits of populations was observed for both parental lineages, in special for the paternal line. The results also showed a higher diversity on the male line than on the female, which is indicative of a higher dispersal rate on females. Both of these results were visible at different time scales, even though a comparison of strength between both time-related analyses could not be performed do to constrain on the genomic dataset. These results are in agreement with previous results, even when using different data for the linguistic or genomic variables.

Tendo em mente a hipótese primeiramente apontada por Darwin sobre a possibilidade de as árvores filogenéticas genómicas e linguísticas partilharem o traçado da história demográfica humana, o foco do projeto LanGeLin é em descobrir quanto desta possibilidade é verdade. Fazendo uso de novas técnicas na classificação linguística, como o uso de sintaxe e da metodologia PCM, e da grande disponibilidade de informação genómica para a nossa espécie, como sequências da região HVR 1, e marcadores SNPs e STRs, este estudo focou-se em descobrir a relação entre a diversidade genómica e linguística para linhagem paterna e materna dentro do contexto das línguas Euroasiáticas.

Os resultados das análises aqui efetuadas revelam uma correlação estatisticamente significante entre as diversidades genómica e linguística com a distância geográfica, demonstrando a existência de uma evidente estruturação geográfica e que tanto genomas como línguas tendem a ser mais semelhantes aos seus vizinhos geográficos. Uma clara correlação entre os traços genómicos e linguísticos das várias populações foi observada em ambas as linhagens parentais, em especial na linha paterna. Os resultados também mostram uma maior diversidade na linha masculina comparativamente com a feminina, o que é indicativo de uma maior taxa de dispersão na linha feminina. Ambos os resultados foram visíveis a diferentes escalas temporais, mesmo que uma comparação da força entre as análises de ambas as escalas temporais não possa ter sido efetuada devido a restrições nos dados genómicos. Estes resultados estão em acordo com resultados prévios, mesmo quando usando diferentes variáveis genómicas e linguísticas.

Keywords/Palavras-chave

Demographic history; Genomics; Linguistics; Genomic and linguistic correlation, Eurasia; LanGeLin Project; Population structure

História demográfica; Genómica; Linguística; Correlação genómica e linguística; Eurásia; Projeto LanGeLin; Estrutura Populacional

Table of Contents

Tables and Figures Index

Abbreviations

AMH	Ancient/ Anatomically Modern Human
AMOVA	Analysis of Molecular Variance
Ar.	Ardipithecus
Au.	Australopithecus
BLAST	Basic Local Alignment Search Tool
CRAN	Comprehensive R Archive Network
DNA	Deoxyribonucleic Acid
EMBL-EBI	European Molecular Biology Laboratory – European Bioinformatics Institute
Geo	Geographic/Geography
H.	Homo
HVR 1	Hypervariable Region 1
IE	Indo-European
ky	Thousand Years
LanCode	Language Code
Ling	Linguistic
LanGeLin	Language and Gene Lineages
mtDNA	mitochondrial Deoxyribonucleic Acid
MDS	Multidimensional Scaling
MUSCLE	Multiple Sequence Comparison by Log-Expectation

MY	Million Years
NCBI	National Center for Biotechnology Information
PCM	Parametric Comparison Method
rCRS	revised Cambridge Reference Sequence
SNPs	Single Nucleotide Polymorphisms
STRs	Short Tandem Repeats
U.K.	United Kingdom of Great Britain and Northern Ireland
U.S.A.	United States of America
Y-chr	Y chromosome

Introduction

1. Humans are different

Humans differ from other species in a great number of ways. Some differences are subtle others are more extreme. We differ from others not just in a genetic level, but also in anatomical and behavioural levels. When taking into account our closest living relatives, the primates, the genetic differences between us may seem not so great (but the results of those differences are plainly to see), for instance, our species differs at the genomic level an estimated 1.5% when compared to chimpanzees (Marques-Bonet, et al., 2009). Although at the phenotypical level we are plainly distinguishable, and even though we inhabit a great variety of environments around the globe, with populations presenting multiple different characteristics (e.g. skin colour, height, cultural practices and languages), we still present a lower level of within-species diversity than our close relatives, with our species presenting an *Fst* equal to 0.15 (Barbujani & Colonna, 2010) while the *Gorilla gorilla* presents a *Fst* equal to 0.38 (Thalmann, Fischer, Lankester, Pääbo, & Vigilant, 2007) and between Western and Eastern chimpanzee a *Fst* equal to 0.32 (The Chimpanzee Sequencing and Analysis Consortium, 2005) is observed.

On top of all the genomic and anatomical differences that can be observed on our species, be it within or in comparison to other species, humans also present some behavioural and cultural differences of great importance. A clear example of these behavioural traits is our linguistic capabilities. Currently there are listed more than 7000 living languages around the globe in the *Ethnologue* database (Lewis, Simons, & Fennig, 2015), which goes to show our disparate linguistic capabilities in relation to other species. A look at the history records will show the viewer a myriad of cultural practices that exist and existed along the time all over the human range.

All this differences made us a species like no other, even though ruled by the same principles as others, and thus a very curious case for study. A species where not just the natural laws apply to shape it, but where its own behaviour and cultures helped shape its genetics and vice-versa (Laland, Odling-Smee, & Myles, 2010).

1.1. How and why

Even though the pursue for understanding the differences between us and our close relatives, and thus to understand what makes us different is relatively new in human history (Varki, Geschwind, & Eichler, 2008), already some data exists on this subject, for example at the Matrix of Comparative Anthropogeny of the Center for Academic Research and Training in Anthropogeny (CARTA, 2015). It is possible to look for comparative information between humans and our close relatives (chimpanzees, bonobos, gorillas and orangutans) at all different levels, from behavioural, anatomical, metabolic and genomic level.

During our ancestors' history, from our earlier ancestors to more recent ones, different mechanisms helped shape our evolution and history. These mechanisms acted not just in our genetic composition but also, in a parallel way, in our behaviour and thus, cultural traits, and both our genetic and cultural traits evolution, up to some level, shaped and left marks in each other. The mechanisms thus involved in the shaping of the human species are then natural selection and neutral-factors such has mutations, genetic drift and gene flow.

1.2. Natural selection

Natural selection is a factor that acts on the adaptability of species and populations and leads to an increase in the frequency of alleles responsible for traits that confer adaptive advantage in a given environment (Barreiro, Laval, Quach, Patin, & Quintana-Murci, 2008). An example of the prevalence of a phenotype in certain populations as a response to a given environmental condition is the skin colour, which has been shown to have a strong positive correlation with high ultraviolet radiation intensity (Norton, et al., 2007). Other examples of traits under natural selection are the insulin regulation and immune systems (Barreiro, Laval, Quach, Patin, & Quintana-Murci, 2008).

1.3. Neutral factors: demographic history

The changes due to neutral factors, in oppose to natural selection changes, are not influenced by adaptive pressure, that is, do not affect a given trait making it more or less common in response to an environmental change for example, thus, this changes can be beneficial, neutral or detrimental for the population. Example of neutral factors are mutations, genetic drift and gene flow.

1.3.1. Mutation

Deletions, Insertions, duplications, substitutions and segmental changes are the ultimate source of genetic variation and genetic diversity within a species and can lead to significant differences between isolated populations within a species or between close species (Varki, Geschwind, & Eichler, 2008).

1.3.2. Genetic drift

Genetic drift is the caused by random sampling of the genetic variation from one generation to another. Each generation the occurring sampling will be independent and may produce different level of allele frequency in the population. Therefore, genetic drift alone can lead to differentiation of two isolated populations and to evolution of a given population when compared to its ancestral state (Masel, 2011).

1.3.3. Gene flow

In opposition to genetic drift which leads to population differentiation, gene flow between populations leads to population convergence, by allowing for exchange of genetic material between them, which may lead to the creation of softer allele frequency gradients between ever further distant populations (Barbujani & Colonna, 2010; Gravel, 2012; Patterson, et al., 2012).

The mechanisms leading to population evolution don’t always have a clear single reason, but are rather the product of interaction of the above points. Two populations living under the same conditions may diverge from each other by natural selection acting upon different alleles that may appear in either population due to random mutation or random sampling (genetic drift) from an ancestor population gene pool or even by gene flow with a third population. In the particular case of humans, another factor may enter in play, that is, culture (Varki, Geschwind, & Eichler, 2008; Allentoft, et al., 2015). Two populations with the same allele frequency may differ in it in a few generations due to cultural influences working has a barrier, like linguistic differences (Barbujani & Sokal, 1990). Thus, comprehending the demographic history of humans' populations is focal to understand how the current observed diversity came to be.

2. Human Evolution

The history of human evolution continues to this date to be a "hot topic". Topics over which hominid specimens can be awarded with the *Homo* genus or how many species there was, when and how modern humans transitioned from archaic humans, how the path and timings of spread of ancient/anatomically modern humans, how different populations interacted with each other, and also, which paper behaviour and culture had in all these events, are still subject to investigation and debate (Klein, 2000; Relethford, 2001; Templeton, 2002; Serre, et al., 2004; Relethford, 2008; Wang, Farina, & Li, 2013; Schwartz & Tattersall, 2015).

2.1. Human tree

Studies point for a divergence of the human and chimpanzee (our closest relatives) branches at 8 – 5 MY (million years) ago in east Africa (Klein, 2000; Relethford, 2008) with the birth of our ancestors, known has australopithecines, being dated at 4.4 MY ago (Klein, 2000) with the *Australopithecus anamensis* being the first of the line (Relethford, 2008) with individuals being smaller when compared to current humans, and presenting still skull, trunks, arms and high sexual dimorphism resembling those of apes.

At 3 – 2.5 MY ago two divergent lineages appear, the *Paranthropus* (presenting a modest increase in brain case size and a large increase in cheek teeth size) which became extinct around 1.2 MY ago, and the first species of the genus *Homo* (presenting a significant increase in brain case size and reduction of the size of the cheek teeth), even though it is still controversial if the first species to be assigned should be to the *Homo affarensis habilis,* the *Homo rudolfensis* or a third and more ancient species, the *garhi*, whose genus is still discussed between assignment to the *Australopithecus* or the *Homo* genus (Klein, 2000; Hublin, 2014; Gibbons, 2015).

At 1.8 MY ago a new species emerges, the *H. erectus* (some authors reserve this name for the species present later in Southeastern Asia, while giving to this species the *H.*

ergaster designation), the first ancestor species whose body form indicates, with no doubt, a full commitment to ground living, and the first were sexual dimorphism is reduced to the level of modern humans. This is the first ancestor to leave Africa on an expansion wave at 1.7 MY, which leads him to colonize Eurasia and the Far East (Klein, 2000; Templeton, 2002; Relethford, 2008; Hublin, 2014).

From the initial expansion of *H. erectus* from Africa at 1.7 MY ago, these populations diverged from each other through genetic drift and natural selection, until 500-400 ky (thousand years) ago when its consider to exist at least three lineages: the *H. sapiens* in Africa, the *H. neanderthalensis* in Europe and the *H. erectus* in the Far East (Klein, 2000; Klein, 2008; Wang, Farina, & Li, 2013).

At around 130 ky ago, the *H. neanderthalensis* is considered to have its classical features, and presents adaptations to cooler environment such as massive trunks and short limbs (Klein, 2000). In Africa, around the same period the *H. sapiens* is also considered to present near modern characteristics (Klein, 2000; Templeton, 2002; Relethford, 2008). The most problematic lineage is the Far East one, due to sketchy records, where some believe that rather than one, two lineages exist: one in the southeastern Asia with a continuity within Indonesian *H. erectus* (from before 500 ky ago until 50 ky ago), another in China evolving from the classical *H. erectus* (before 500 ky ago) to the populations that, by 100 ky ago, retained few of its characteristics and approximate more to the *H. sapiens* in brain case size (Klein, 2000; Relethford, 2008).

At around 100 ky ago another expansion event occurs, with the *H. sapiens* expanding out of Africa, reaching Australia by 60 – 40 ky ago and Europe by 40 – 30 ky ago, and the recent Austronesian linguistic family expanding only at 6 ky ago. In general its consider that by 50 ky ago the *H. sapiens* started to replace the existing populations of Eurasia and Far East (Barbujani & Bertorelle, 2001; Cann, 2001; Templeton, 2002; Relethford, 2008; Hoffecker, 2009; Benazzi, et al., 2015).

After this point, multiple events of population movements occurred throughout with populations and cultures moving from their original points. In the Near East, in the Levant region, by 10 ky ago agriculture is born and after spreads through Europe (Barbujani & Bertorelle, 2001). Later demographic events occurred throughout Eurasian leaving both cultural and genetic evidences.

On the human tree construction, it is to note that the time of each event (and distinction between each ancestor species) is always theme of debate, due to lack of complete fossil record, behavioural and cultural material left, and a lack of clear definition of species (in general and human in particular). Moreover, the lack of a clear and widely accepted definition of the boundaries for the genus *Homo* and its species. Thus, any reconstruction of a human ancestral tree is subject to in the future be revised has new materials and new analytical techniques arise (Klein, 2000; Cann, 2001; Relethford, 2008; Hoffecker, 2009; Hublin, 2014; Gibbons, 2015; Schwartz & Tattersall, 2015). An example of this constant actualization of the human tree is the recent publication of the discovery of a new species of the genus *Homo*, the *H. naledi* (Berger, et al., 2015; Dirks, et al., 2015), which, with an as-yet unknown date, may push the genus *Homo* further back on time or be another indicator of the coexistence of multiple species of the genus *Homo* in Africa in the distant past; and the recent results of Meyer et al, reported at this years' meeting of the *European Society for the study of Human Evolution*, on the sequencing and taxonomic classification of fossils of Sima de los Huesos, Spain, pushes the separation time of the *H. neanderthalensis* and *H. denisovan* and these from the *H. sapiens* further back in time, with their predictions for split between Neanderthal and Denisovan predating 430 ky ago (Meyer, et al., 2015). An example of a human tree can be seen in Fig. 1.

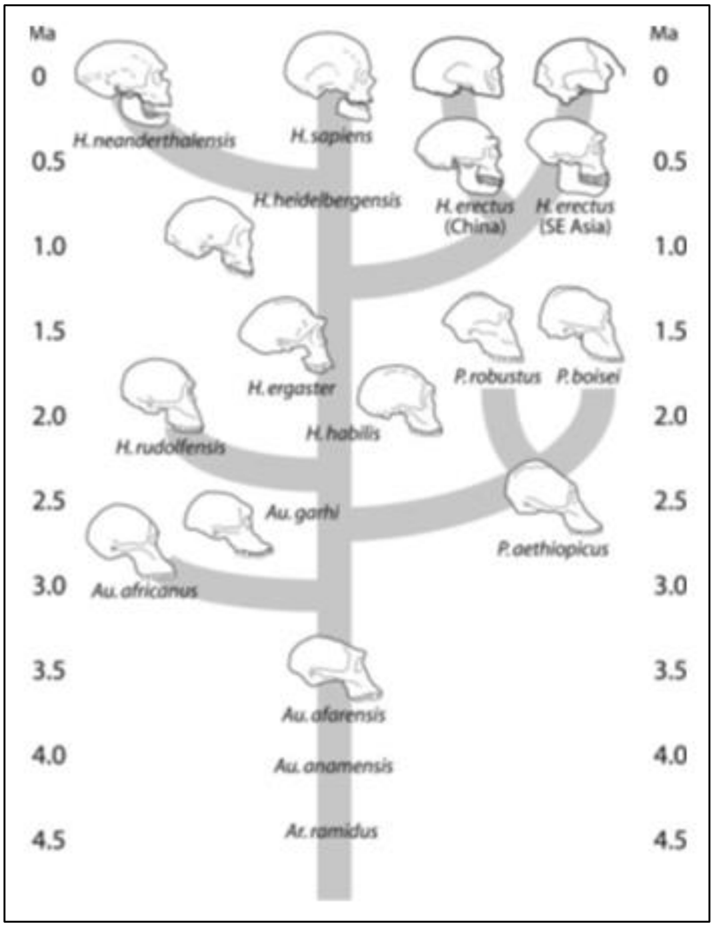

Figure 1 - A working phylogeny of the hominins after 4.5 MY ago. As in Klein 2008

2.2. Modern Behaviour

The fully modern behaviour appearance in ancient modern humans is a complex and highly debated theme, with some postulating its gradual appearance from 120 ky ago to 50 ky ago and others postulating its abrupt appearance at 50 ky ago (Klein, 2000; Klein, 2008). Up until this change in behaviour, both *H. sapiens* and *H. neanderthalensis* remained similar in behaviour, proof of that is the similarity between artifacts of both species at 250 ky ago and even after (Klein, 2000). The later difference may be one of the keys to the replacement of the *H. neanderthalensis* by ancient modern humans without leaving much of genomic or cultural trace, that is, the genomic and behavioural differences may have been too big.

The advanced human behavioural traits appearing after 50 ky ago imply the existence of a fully modern capacity for innovation which underlies culture, with the most significant behavioural novelty being the appearance in cultural artifacts of unequivocal art and personal ornaments, which indicate the capacity for abstract or symbolic thought (Klein, 2008).

One problem to describe the history of the behavioural evolution in humans is on to what may have caused this change, with some postulating for a mutation driving this abrupt change in the evolutionary rate of behaviour, while others advocate for a change due to population growth and social reorganization (Klein, 2008). Another problem is the fact that the artifacts registry not always gives a full picture due to its low numbers and sparsity across time and landscape (Hoffecker, 2009).

2.3. Out of Africa

In the previous point a simplistic approach was used to describe the emergence of the *H. sapiens* through its appearance in Africa and later expansion out of it and replacement of the other hominid populations living in Europe and Asia at that period. This model is described has the Out-of-Africa model, which advocates that ancient modern humans originated in Africa and expanding out of it came to replace the *H. neanderthalensis* and *H. erectus*. Another model, opposing this one, is the multiregional model, which advocates that the populations in Africa, Eurasia and Far East where just one lineage (rather than three distinct lineages) which evolved all together with a constant gene flow, either through isolate trait appearance and coalescent evolution of the three populations or appearance of the traits in Africa and diffusion through gene flow to the others (Relethford, 2001; Relethford, 2008).

Molecular data, such has the African higher genetic diversity and rare alleles presence, a genetic continuity between ancient modern humans and present humans and a discontinuity between the last and the *H. neanderthalensis*, points an African origin of our species and to the Out-of-Africa model (Harpending, et al., 1998; Barbujani & Bertorelle, 2001; Caramelli, et al., 2003; Serre, et al., 2004; Ramachandran, et al., 2005; Relethford, 2008; Wang, Farina, & Li, 2013), with a multidispersal hypothesis, where a southern route through the horn of Africa is included has a possible path, being pointed has the more accurate model to explain the genomic differentiation observed in the East Asian populations (Armitage, et al., 2011; Ghirotto, Penso-Dolfin, & Barbujani, 2011; Reyes-Centeno, et al., 2014), while some interbreeding between ancestral modern humans and *H. neanderthalensis* may have occurred in Europe (Wang, Farina, & Li, 2013; Seguin-Orlando, et al., 2014).

The supporters of the multiregional hypothesis indicate that the genetic evidence between the discontinuity between the *H. neanderthalensis* and the *H. sapiens* can be explained by an extinction of the *Neanderthal* line due to simple genetic drift (Relethford, 2001).

2.4. Peopling of Eurasia

Just like the origin of the *H. sapiens*, also the origin of the extant genetic structure observe in Eurasia and in particular Europe is cause for debate.

When observing the extant genetic structure of Europe, a visible cline from southeast to northwest is present (Cavalli-Sforza, Piazza, Menozzi, & Mountain, 1988; Chikhi, Destro-Bisol, Bertorelle, Pascali, & Barbujani, 1998; González, et al., 2003; Lao, et al., 2008; Novembre, et al., 2008; Barbujani, 2013). This cline is pointed as either being a result of genetic structure caused by the expansion observed in the Paleolithic to Europe or caused later by a demic-diffusion during the Neolithic expansion of populations and agriculture from the Levant at 10ky ago. The last model implies that culture spread with people and not by cultural diffusion alone. Results of molecular analysis throughout the years have not helped to clarify the problem, with teams reporting results pointing for one or the other hypothesis (Chikhi, Destro-Bisol, Bertorelle, Pascali, & Barbujani, 1998; Barbujani & Bertorelle, 2001; González, et al., 2003; Skoglund, et al., 2012; Barbujani, 2013; Seguin-Orlando, et al., 2014; Allentoft, et al., 2015).

The same problematic is observed when trying to decipher the point of origin for the Indo-European linguistic family, were the two main hypothesis try to explain its origin and expansion. A dubbed Anatolian hypothesis points for the origin of the Indo-European language in the Anatolia region and that its spread to Europe occurred at around 8 ky ago (Bouckaert, et al., 2012; Balter & Gibbons, 2015). The other main hypothesis, generally known as the Steppe hypothesis, posits for an origin of the Indo-European root on the steppes north of the Caspian and Black Seas, and its spreading to Europe at around 5 ky ago (Brandt, et al., 2013; Allentoft, et al., 2015; Haak, et al., 2015). Once again, different studies including both genetic and linguistic data support either the Anatolian or the Steppe hypothesis with an answer for the origin of the Indo-European linguistic family still unclear (Gray & Atkinson, 2003; Haak, et al., 2010; Myres, et al., 2011; Bouckaert, et al., 2012; Brandt, et al., 2013; Allentoft, et al., 2015; Haak, et al., 2015).

Both questions remain still open and largely debated and more data, both from biology, archaeology and sociology, along with more realistic models will be needed in order to solve the question (Reich, Thangaraj, Patterson, Price, & Singh, 2009; Bouckaert, et al., 2012; Barbujani, 2013). But either way, it is already clear (through molecular and cultural analysis) that multiple populations and cultural movements occurred throughout Eurasia history and both left their mark, even if small, and thus helped to outline what would become the current genetic, linguistic and cultural background of Eurasian populations (Barbujani & Bertorelle, 2001; Myres, et al., 2011; Brandt, et al., 2013; Lazaridis, et al., 2014; Allentoft, et al., 2015; Haak, et al., 2015).

3. Anthropology

Anthropology in simple words is the study of the human species. It ranges from the study of its social and cultural and linguistic traits, to its anatomy and genomic traits. It tries to uncover the history of human kind and comprehend how our species came to be what it is today and how it functions. Two of its branches are the molecular and linguistic anthropology.

3.1. Molecular anthropology

Molecular anthropology is the branch of anthropology that focuses on studying the human history through molecular analysis, that is, through the study of genetic traits of both extant and extinct populations. Some of the tools used to grasp insight into the evolution and genetic composition of a population are what is called molecular markers, that is, genomic regions whose location is known and whose genetic information (when compared across multiple individuals of a species) can help in uncovering and understanding the evolutionary history of a species/population, its structure and its demographic history. They can further be used to make clear the time of occurrence of specific demographic events or to discriminate sex differentiated demographic patterns (Barbujani & Bertorelle, 2001; Pereira, Dupanloup, Rosser, Jobling, & Barbujani, 2001; Wen, et al., 2004; Relethford, 2008).

For instance, by making use of markers with different mutation rates, that is, creating datasets of genomic information using slow- and fast-evolving markers, whereas slow-evolving makers are markers with a slower rate of mutation, thus expressing patterns of demographic events that occurred further in the past, and fast-evolving markers are markers with a higher mutation rate and thus expressing patterns of demographic events that occurred more recently, it is possible to explore differences in evolutionary history and observe how different historical demographic patterns affected a given population at different time scales (Relethford, 2008).

Another type of markers that can be used in demographic studies are the uniparental markers, that is, markers that are only transmitted by one sex, which are useful for studies which aim to discern the possibility of differential demographic patterns for each sex. For studies of this nature, markers of molecular diversity on the mitochondrial DNA (mtDNA) and the non-recombined portion of the Y chromosome (Y-chr) are used, for they present differential inheritance and complete, or nearly so, absence of recombination, thus the patterns revealed by each will only reflect demographic history of the transmitting sex of that particular marker. For instance, by using mtDNA and Y-chr markers, and comparing the diversity and geographical structuring of each, was possible to observe that in fact both parental lineages present differences in their demographic history (Pereira, Dupanloup, Rosser, Jobling, & Barbujani, 2001; Prugnolle & Meeus, 2002; Pakendorf & Stoneking, 2005).

An example of these sex-biased molecular markers that can be used on the study of human demographic history are the mitochondrial DNA Hypervariable Region 1 (mtDNA HVR 1), the Y chromosome Short Tandem Repeats (Y-chr STRs) and Single Nucleotide Polymorphisms (SNPs) on both uniparental genomic regions (Miller & Kwok, 2001; Stumpf & Goldstein, 2001; Pakendorf & Stoneking, 2005). On a study trying to uncover differential patterns of demographic history for each sex and at different times, a dataset comprising of data from mtDNA HVR 1 and Y-chr STRs markers can be used as fast-evolving dataset while data from SNPs can be used to create a slow-evolving dataset for each sex.

Obviously, not all is straight has presented above and some precautions are always necessary for the results interpretation. For instance, the rate mutations of each Y-chr STR may vary largely from one to the other, and a dataset comprised of data from this marker may present a mean rate mutation different from a dataset of mtDNA HVR 1 data. Other factors, such as effective population size, may also skew the results, and an obvious thing to take in account is that any uniparental marker only represents the demographic history of the sex in which its inheritance occurs, and when analysing results from these markers, this must be taken in consideration (Miller & Kwok, 2001; Prugnolle & Meeus, 2002; Gusmão, et al., 2005; Pakendorf & Stoneking, 2005; Barbujani, 2013).

3.2. Linguistic anthropology

Within the linguistic branch of anthropology, scientists try to understand the process that shaped human communication, be it verbal and non-verbal, variation in languages across space and time, its social uses and the relationship between language and culture.

In particular, through the retracing of a language or linguistic group history, anthropologists are able not just to identify demographic patterns but also to, through study of the relationship between people and cultures during these events, to understand how these events shaped the history of humankind or a particular population or linguistic group. This can be made because (just like for genetic traits) linguistic traits also suffer processes of evolution, that is, they suffer change over generations on a process called descent with modification. Linguistic traits can also diverge due to mutations or drift between diverging populations and even through horizontal transfer of traits from contact with neighbours (Atkinson & Gray, 2005; Dunn, Terrill, Reesink, Foley, & Levinson, 2005; Gray, Atkinson, & Greenhill, 2011; Atkinson, 2013; Bouchard-Côté, Hall, Griffiths, & Klein, 2013; Pakendorf B. , 2014).

Studies whose emphasis is the resolution of linguistic families, history and their relationship, use linguistic information either based on the lexical or syntactical characteristics of the languages in study. These two types of linguistic information regard two different aspect of a language, where lexical characteristics are based on the structure of phonemes and the morphology (that is, how the words are composed and their relation, meaning and root), while the syntactical information takes in account not the individual words but the grammatical rules of the language, that is, the rules for the structuring of phrases thus to make meaning of an agglomeration of words (Guardiano & Longobardi, 2005; Longobardi & Guardiano, 2009).

The primary tool still in use are the lexical characteristics of languages, mainly due to the existence of large databases for lexical characteristics. It has been used to try to retrace the history of languages and linguistic groups such as the Indo-European and Austronesian linguistic groups (Gray, Drummond, & Greenhill, 2009; Longobardi & Guardiano, 2009; Gray, Atkinson, & Greenhill, 2011; Chang, Cathcart, Hall, & Garret, 2015).

The use of lexical characters presents some problems for the study of linguistics, such as, the value assigned to a given difference between languages tends to be arbitrary rather than exact value, word/morphemes sound and meaning between languages tends to dissolve over time and the appearance of random similarities. Because of this problems, inferring the relatedness between two languages pertaining to distant linguistic groups may present itself a problem, and studies incorporating ever so distant languages may see their attempts in discover their relation be hinder (Dunn, Terrill, Reesink, Foley, & Levinson, 2005; Guardiano & Longobardi, 2005; Longobardi & Guardiano, 2009; Colonna, et al., 2010; Pakendorf B. , 2014).

As an alternative, the syntactic characteristics of a language present themselves as very useful for the classification of the differences between two languages, for a given characteristic is expressed in an exact value and the syntactic characteristics tend to have a slower rate of evolution and not suffer so much with horizontal transfer which only occurs under prolonged and intense contact between languages (Cann, 2001; Dunn, Terrill, Reesink, Foley, & Levinson, 2005; Colonna, et al., 2010). Recently a new methodology has been developed and suggested in Longobardi (2003) and Guardiano and Longobardi (2005) to make use of the syntactic approach for describing a language. This methodology called Parametric Comparison Method (PCM) (Longobardi, 2003; Guardiano & Longobardi, 2005) makes use of a binary code composed of the values assigned for each syntactic characteristic of each language (much like a genetic string code of nucleotides), and calculates the exact linguistic distance between all the pairs. The power from this method comes from the universality of the syntactic characteristics used to describe each language, thus bypassing the problems of loss of in-common characteristics sometimes present when comparing long distance languages using lexicon and from the binary coding of these traits, allowing for a more biological approach (Longobardi, 2003; Guardiano & Longobardi, 2005; Longobardi & Guardiano, 2009).

4. Human Demographic history: an Interdisciplinary approach

4.1. Darwins' hypothesis

In 1859 Darwin (Darwin, 1859) firstly noted on his *Origin of Species* "If we possessed a perfect pedigree of mankind, a genealogical arrangement of the races of man would afford the best classification of the various languages now spoken throughout the world; and if all extinct languages, and all intermediate and slowly changing dialects, were to be included, such an arrangement would be the only possible one.". After, in his *The Descent of Man* (Darwin, 1871), he noted that "The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously parallel".

With these statements Darwin acknowledge that both genes and languages, even though through different mechanisms, suffered descent with modification, that is, both evolved over generations through similar evolutionistic mechanisms and that by retracing the demographic and evolutionary history of each, we would be able to retrace the demographic history of mankind.

4.2. History

The study of human demographic history using insights of both genetic and linguistic data gave its first steps with the publishing by Sokal et al. (1988) and Cavalli-Sforza et al. (1988) of two papers where they show the existence of a correlation between the genetic structure of their samples and the linguistic structure for those samples. Subsequent studies on the genetic and linguistic structure observed in European populations demonstrated that there was a clear sign of correlation between these two variables (Barbujani & Sokal, 1990; Sokal, et al., 1990).

After these initial results, more scientists started to adhere to the use of both genetic and linguistic data to try to answer questions that either one or the other couldn't alone resolve. In Europe, one of the main questions concerns the origin of the Indo-European languages and their time of dispersal. In East Asia one questions revolves around the origin and path of dispersal of the Austronesian linguistic family. Currently these questions are being resolved with use of both variables, based on the assumption that the demographic events that may have occurred let not just a mark on the genetic pool of those populations but also on their languages (Cann, 2001; Pakendorf B. , 2014).

4.3. Difficulties

Since the start of these interdisciplinary studies to infer the demographic history of populations, doubts and "red flags" have been issued, mainly by classical linguistics who pointed problems on the linguistics part of the analysis, due to their discredit on the use of statistical models drawn from genetics to linguistic analysis (Pakendorf B. , 2014; Hunley, 2015). Problems were noticed when inferences were made for languages with greater distance between them, with the methods used failing to infer correctly or at all their relationship (Colonna, et al., 2010).

A point as also been noted on the overall analysis by some, pointing out that not always the demographic history of a population will leave the same effects on both genetic and linguistic characters, and thus approaches relying on this assumption may faultily infer the history of a given population.

In the genomic side of the approach some old problems have been noticed too, such has the problematic of what genomic data to use, since autosomal chromosomes suffer from recombination, thus potentially shuffling the history of a population, and that both mitochondrial and non-recombinant Y chromosome portion just give their sex-specific lineage story, and they may, as has been proven, present different demographic patterns.

Ultimately, in order to solve this problems, new methodologies were created to try to overcome them, in the linguistic side, the arise of the syntax use has a way to overcome the problematic of infer the relationship over distant languages, and the PCM methodology to infer said relationship, while on the genomic side studies are using both sex-specific markers to try to infer the populations demographic pattern, and with the aid of archaeology inferences on both data are being made taking into account precise demographic events dates, thus enhancing the precision of the analysis, thus enhancing the capacity to uncover the demographic history of a population.

4.4. Results

The results of this interdisciplinary approach are already far, even though not all coming to the same conclusions. For example Bosch et al. (2006) found that even though cultural and linguistic barriers exist in the Balkans region, the populations show genetic homogeneity between them (except for a population isolate, the Aromuns), which points has either these barriers not being strong enough to make an effect on the genetic structure of the population or that they might have been introduced later after the genetic composition was already in place only through cultural diffusion. But overall studies have been showing a clear pattern of correlation between genetic and linguistic (Tishkoff, et al., 2009; Colonna, et al., 2010; Pakendorf B. , 2014; Creanza, et al., 2015).

In Europe, and up to some extent in the rest of the Indo-European linguistic range, several studies have been focused on discerning the origin of the Indo-European linguistic families and how its expansion occurred and shaped Europe genetic pool, both by using genetic and archaeological data to infer on the linguistic history or the reverse (See Human Evolution: 2.4 Peopling of Eurasia), with results showing a clear pattern of correlation between linguistic and genetic structure, indicating that both variables, even if slightly, are correlated and coevolved (Barbujani & Sokal, 1990; Barbujani & Pilastro, 1993; Gray & Atkinson, 2003; Allentoft, et al., 2015).

In other points of the world the results are also prosperous, with, e.g., Lansing et al. (2007) also finding evidence for a correlation between genomics and linguistics in Sumba (eastern Indonesia), even when controlling for geographical distances, and Tishkoff (2009) also finding a correlation between linguistics and genomic in Africa

4.5. Implications of results

Darwins' idea that a complete tree of all the languages and its variants of the world would perfectly match the genetic tree of all human populations seems now somewhat far from true. Even though a certain parallelism exists between linguistic and genetic evolution and that both are shaped by demographic events, today we now that the evolutionary process and the effect that demographic events have on both traits differ and in some cases may lead to different histories for both, thus confounding our inferences on one using the other (Barbujani, 1997).

Taking all these problematics in to counsel, the implications of these studies is great for they allow us to ever more gain knowledge on each populations' history, knowledge that can be then used in other fields, such has, for example, in personalized medicine, were the roots of a given individuals and population are taken in account.

Objectives

1. The LanGeLin Project

The LanGeLin (Language and Gene Lineages) (2015) is an international and interdisciplinary project founded by the ERC research project "Meeting Darwins' last challenge: toward a global tree of human languages and genes".

The project aim is to access at a world scale the hypothesis pointed by Darwin in his *The Origin of Species* (1859), that there is a correlation between the transmission of cultural traits, namely, the language, and genetic traits, and so, that both traits are affected by population demographic events, such has divergence and isolation. The question will be accessed through: the evaluation of the level of gene-language correlation for selected populations, to understand if observed genetic structure of said populations is mirrored by similar pattern of linguistic structure; the evaluation of the level of gene-language parallelism when it comes to relationships with neighbours, to understand if like in genes languages also suffer horizontal transmission; and the evaluation of the effect of each sex in the transmission of both traits to the descendants, to understand what part each sex plays on the transmission of genetic and linguistic traits.

The LanGeLin (2015) project will rely on the expertise of different research groups, from linguistics, population genetics and molecular anthropology. It is coordinated by Prof. Giuseppe Longobardi from University of York (U.K.) in partnership with the Università Degli Studi di Ferrara (Italy) coordinated by Prof. Guido Barbujani and the Alma Mater Studiorum – Università di Bologna (Italy) coordinated by Prof. Davide Pettener.

These groups will make use of the new comparative method for linguistic classification, the PCM (Guardiano & Longobardi, 2005; Longobardi & Guardiano, 2009), based on the syntactic characteristics of each language rather than using the lexical traits as in other works, along with the extensive genetic information that has been stored for the last 15 years in public databases, thus increasing the reliability of the approach for calculating distances between distant languages, and the power of the genetic approach due to increased amount of genetic data used compared to other population genetics studies.

2. Thesis objective

The aim of this thesis is, through the analysis of genomic, geographic and linguistic data for populations within the Eurasian linguistic diversity to access the existence of a correlation between the genomic, geographic and linguistic data, and if these correlations are observed for both parental lineages and at different time scales. Thus, in resume three question are intended to be resolved in this work:

- Is there a correlation between both genomic (both parental lineages) and linguistic diversity with the geographic diversity of the studied populations, and is this observed independently of the time scale used?
- Is there a correlation between the genomic diversity (for both parental lineages) and the linguistic diversity on the populations studied, independently of the time scale used?
- Is there a difference in the demographic history of each parental lineage, independently of the time scale used?

Materials

In this project an interdisciplinary approach was taken and therefore data stemming from different fields was used. In order to make clearer the kind of materials used for resolution of the objectives proposed the materials will be presented in three sections: data, statistical tools, bioinformatics tools. A brief summary of the importance of each material selected is presented after each indication.

1. Data

1.1. Genomic data

In order to study the possibility of different demographic patterns in the female and male lineages and to study historical demographic patterns at different time periods a combination of uniparental markers (mitochondrial DNA and Y chromosome) and fast- and slow-evolving markers (HVR 1 and STRs and SNPs, respectively) was used, respectively. Our genetic data can be then divide in four datasets, mitochondrial DNA Hypervariable Region 1, mitochondrial DNA Single Nucleotide Polymorphisms, Y chromosome Short Tandem Repeats and Y chromosome Single Nucleotide Polymorphisms.

For the purpose of treatment of mtDNA data, the revised Cambridge Reference Sequence (Andrews, et al., 1999) was used.

A comprehensive explanation of the characteristics of each marker can be found in Introduction: 3.1 Molecular anthropology.

1.2. Geographic data

The geographical dataset for this project is composed of a collection of the latitudinal and longitudinal geographical coordinates for each of the language population locations selected for the study.

1.3. Linguistic data

For this project the data used to create the linguistic dataset is composed of the syntactic characteristics of each language, described using the Parametric Comparison Method (Longobardi, 2003; Guardiano & Longobardi, 2005).

A comprehensive explanation of the characteristics of this data can be found in Introduction: 3.2 Linguistic anthropology.

2. Statistical tools

2.1. AMOVA

The Analysis of Molecular Variance or AMOVA (Excoffier, Smouse, & Quattro, 1992) is a statistical tool used to compute various diversity indices intra and inter populations. Different *a priori* information can be given to the program about our data, the computations we want the program to perform and the different sets we want it to use. Thus, for this project, this tool was used to access the inter-populations diversity.

2.2. Fst

Also known as *Fixation index*, is an index first formulated by Wright (Wright, 1951) as a tool to calculate the genetic diversity between populations, or by other words, this index indicates the degree of genetic differentiation between two pairs of population based on their allele (or haplotypic) frequencies. Being a dissimilarity index, the more close the populations are, the lower the value assigned for the pair, being 0 assigned for two pairs of population that present no genetic differences, and 1 for a pair of populations that are completely divergent. This index was used to calculate the genetic differentiation between populations on all our datasets with exception for the mtDNA HVR1 dataset.

2.3. Great Circle distances

This index indicates the smallest distance between two given geographical locations taking in consideration that both are represented in a spherical surface. A Great Circle is a section of a sphere whose center coincides with center of the sphere and it gives only a unique path between any two points on the sphere surface, with exception for antipodal locations for which gives an infinite number of paths. From the path given by the Great Circle, two arcs are obtained, being the smaller one an orthodrome, also known as Great Circle distance (Weisstein, 2015) .

2.4. Jaccard distances

This is an index used to measure the dissimilarity between two finite sets of data. It was created by Paul Jaccard (1908) and is complementary to his other measurement, the Jaccard coefficient which measures the similarity between data sets. Within the context of this project, the Jaccard distance index is used to calculate the distance between the languages, that is, it is used to create a dissimilarity matrix for the linguistic data.

2.5. Mantel tests

Devised by Nathan Mantel (Mantel, 1967), these tests, the Mantel and the partial Mantel tests, are used to test the correlation between two or three (partial Mantel test) data sets, that is, it gives us information on the degree and direction of the correlation between datasets. The degree and directionality of the correlation is given by the *mantel statistic r*, an index whose value varies between -1 and 1, where the correlation is stronger to the extremes and lower to the mid-values, being the directionality given by the sign of the value, that is, positive values indicate a positive correlation and vice-versa. The significance of the test is accessed by a permutation test (see: Materials: 2.8 Permutation test). In the Mantel test we test the correlation between two matrices directly, whereas in the partial Mantel test, a third matrix is used to serve as a control, that is, we access the correlation of two matrices controlling for the effect that might be caused by the third matrix data, as an example, the partial Mantel test can be used to access the correlation between genomic and linguistic distances while controlling for the geographical distances, thus the correlation that might be observed between the first two matrices is not due to effects of the third.

2.6. Nonmetric Multidimensional Scaling

The Nonmetric Multidimensional Scaling (hereafter Multidimensional Scaling or MDS) is a method designed to construct a graphic revealing the underlying structure of the input data creating thus a “map” of it in any *n* dimensions desired (Wickelmaier, 2003; Jaworska & Chupetlovska-Anastosova, 2009). This statistical approach was first devised by Shepard (1962) and Kruskal (1964a; 1964b) and differs from the classical Multidimensional Scaling by the computation relying in the ordinal information of the data and not the distance values given. The Nonmetric MDS aims to map the objects in a configuration whose distance between each object-pair is in the same rank order as in the input data. With this approach based on the rank order of distances, a problem of fitness, i.e., the resemblance between the data and the final configuration, appears. In order to correct this, the programs calculates different configurations and a *stress value* for each, until it reaches a minimum value, thus the best configuration. The *stress value* thus serves to indicate the goodness-of-fit between the configuration obtained and the input data, where the smaller the value the best relation. Even though the criteria to determine the goodness of each value tends to be arbitrary and decided by each investigator, here its followed the criteria defined by Kruskal (1964a):

Table 1 - Stress and Goodness of fit

Stress value (%)	Goodness-of-fit
> 20	Poor
10	Fair
5	Good
2.5	Excellent
0	Perfect

2.7. Parametric Comparison Method

The Parametric Comparison Method or PCM, by Longobardi (2003) and Guardiano and Longobardi (2005), is a methodology devised to extract syntactic information from languages in order to create a binary description of each language assessed.

More information on this statistical tool can be read in Introduction: 3.2 Linguistic anthropology.

2.8. Permutation test

This is a type of statistical significance test devised to access the significance level for a correlation between matrices. The significance level is accessed by calculating the correlation for the matrices as given, thus obtaining a test statistic T_0 (e.g., the Mantel *r* statistic). A test is calculated by permutation of the rows/columns *n* number of times of each matrix and calculate the correlation, obtaining a new test statistics, T_n. This results in a distribution of the possible test statistics T_n under the null hypothesis. The statistical significance (*p-value*) is accessed by calculating the proportion of tests where the T_n is higher than or equal to the T_0 (Welch, 1990).

2.9. Phi-st

Introduced by Excoffier in 1992, the *Φst* or *Phi-st* is a variation of the *Fst*. Like the former it also indicates the degree of differentiation between two populations, only that this index, instead of calculating the dissimilarity between two populations using allele frequencies, it calculates the proportion of nucleotide diversity between them relative to the total. This index was used to calculate the degree of differentiation between populations for our mtDNA HVR1 dataset.

2.10. Scatterplot

Scatterplot is a plotting system where a graphic (plot) is created using information from two matrices. The data is displayed as a set of points, were the values of one matrix are regarded has the abscissa coordinate while the values of the other matrix are regarded as the ordinate coordinate. This kind of plots serve to illustrate the relationship between the two matrices sets.

2.11. Statistical significance

Statistical significance is a tool used to indicate if an experimental value was achieved due to randomness, for instance, due to sampling errors, or if it reflects properly the characteristics of the population sampled (Fisher, 1925). This statistical significance is accessed by comparing the values of the *ρ-value* obtained in the study and the *α* (or significance level) value decided *a priori*. If the *ρ-value*, that is, the probability of occurrence under the *null hypothesis* of a value equal or superior to the one observed, is lower than the *α* value, i.e., the probability of rejecting the *null hypothesis* when it is true, one can conclude that the result is statistically significant, or by other words, it means that the results achieved are not due to randomness or sampling error.

3. Bioinformatics tools

3.1. Arlequin

Arlequin (Excoffier & Lisher, 2010) is a bioinformatics software designed to perform population genetic data analysis. It allows the user to perform several statistical tests, such as Mantel or AMOVA test (Excoffier, Smouse, & Quattro, 1992), on the provided data to extract information pertaining to the diversity and structure of the dataset. It is suited to analyse multiple formats of genetic data (e.g., sequence or haplotypic), and with different degrees of information (e.g., grouping of samples by sub-populations). It is, therefore, a tool with great versatility and of great use in the population genetics, being used in this project to perform AMOVA (Excoffier, Smouse, & Quattro, 1992) analyses on the genomic dataset. The version used is the Arlequin v. 3.5.1.2 (Excoffier & Lisher, 2010).

3.2. BioEdit

BioEdit v.7.2.5 (Hall, 1999) is a genomic sequences editing program that allows users to manipulate, align and inspect sequences in an easy and intuitive way, thus enabling researchers to fit their sequences to their project needs.

3.3. Databases

The data-mining, that is the retrieving of data and related information's useful for a project, can be done whether by looking into papers and books or by consulting databases. Databases are informatics centers were information is stored and organized. Some databases include tools for the direct manipulation of the data directly prior to its acquisition (download). These databases are usually curated to be always in par with current knowledge, and are design to allow one for a simple and efficient way to retrieve all the desired information possible of the data. For this project different databases were used to access, check and complement information of genetic, geographic and linguistic nature. Next follows a list of the databases accessed for this project and their pertinence for it.

Ensembl: a joint project between EMBL - EBI and the Wellcome Trust Sanger Institute. It's a biological database of selected eukaryotic genomes. It works has a place for the annotation of genomic information, but also as a tool to perform some analyses on the data, such as visual display of data or nucleotide comparison between sequences (Cunningham, et al., 2015; Ensembl, 2015).

Ethnologue: a comprehensive language catalog. In it, it is possible to find information pertaining more than 7 thousand languages, from their family tree, origin, status and number of speakers, amongst other information (Lewis, Simons, & Fennig, 2015; Ethnologue, 2015). It is used in this project to scan the geographical delimitations and/or ethnicities corresponding to each of the languages included in this study, information needed for the collection of the geographical and genomic data.

HvrBase++: is an update on the *HvrBase* (Burckhardt, von Haeseler, & Meyer, 1999), a database of genomic information on the HVR1 region of the human mitochondria. The updated database incorporates more information by adding more populations, information on other primates, complete mitochondrial sequences and even sequences from X-chromosome and autosomal loci. Like other genomic databases, it gives useful information about the samples, like the population and the language of the sampled individuals (Kohl, Paulsen, Laubach, Radtke, & von Haeseler, 2005; HvrBase++, 2015). Like the other genomic databases, this one was used for data-mining, in the case, of the mtDNA HVR1 data.

GenBank: a bioinformatics center created to store molecular data (Benson, et al., 2013; GenBank, 2015). The database also incorporates diverse tools that allow the user to visualize the information in different ways and even to perform genomic analyses tasks. Amongst its many utilities, the well-known *BLAST* - Basic Local Alignment Search Tool (Altschul, Gish, Miller, Myers, & D.J., 1990; Blast, 2015) performs sequence alignment between given sequences, accessing their similarity levels. The GenBank database, part of the National Center for Biotechnology Information - NCBI (2015), was thus used for this project for data-mining and for simple alignments.

3.4. *Haplosearch*

Since the genetic information can be retrieved from the many sources in different formats, i.e., in nucleotide sequence or in haplotype format (the list of variant positions detected comparing the genome sampled to a reference sequence), *Haplosearch* is a tool that allows users to transform the data in the desired format in a quick and efficient way. This tool (Fregel & Delgado, 2011), that can be accessed through the *Haplosite* webpage

(2015), was used in the project to convert haplotypic information of the mitochondrial DNA HVR1 data to nucleotide sequences data.

3.5. MUSCLE

MUSCLE, which stands for Multiple Sequence Comparison by Log-Expectation, is a multi-sequence alignment program, both for nucleotide and protein sequences (Edgar, 2004). It was used in the project for the alignment of the mtDNA HVR1 sequences during their preparation.

3.6. R

R is a statistical computing environment on which a user can perform multiple statistical task and to compute graphics. The *R* is not a finite program, in the sense that can be extended by addition of packages. These packages can be added to allow the computation of tasks that are not included in the basic package. Since it is an open source program, new packages can be created by any user and made available in a database for any other users to make use of. The *R* program, along with the available packages and further information and utilities, can be accessed using the *Comprehensive R Archive Network* – CRAN (2015; R Core Team, 2015) . It was used in this project for the computation of multiple statistical tasks. Next follows a list of the packages used and their utility within the scope of the project.

basic package: the *R* comes with various functions in itself, from data manipulation to plotting functions. Functions from this package were used for loading and preparation of data and materials for the execution of the analysis, along with the creation of scatterplots (2015; R Core Team, 2015).

gdata: package with multiple tools for data manipulation, used in this project for preparation of data and support materials for the computations (Warnes, et al., 2015).

MASS: described by the authors has *Functions and datasets to support Venables and Ripley, 'Modern Applied Statistics with S'* (Venables & Ripley, 2002). It's a package with focus on matrix comparisons, being use in this project to compute *Multidimensional Scaling* (Shepard, 1962; Kruskal, 1964a; Kruskal, 1964b) statistics.

vegan: described by the authors has *Community Ecology Package*, it's a package created to perform analyses on populations diversity (Oksanen, et al., 2015). In this project it was used to compute the *Mantel* and *partial Mantel tests* (Mantel, 1967).

Methodology

Being the LanGeLin (2015) project an interdisciplinary effort to understand the demographic history of human populations, different groups were responsible for the collection of different datasets and for their preliminary analysis, being the Ferrara group, were I am inserted, responsible for the mtDNA datasets. The comparative analysis performed over all the datasets was performed by just the Ferrara group. Inside the Ferrara group, on the matters of data collection and preliminary analysis, I was tasked with the collection of the mtDNA HVR1 dataset and its preliminary analysis, for this reason it is only given a small explanation of how the datasets other than mtDNA HVR1 were collected, prepared and analysed prior to the multi-dataset analysis.

If needed, more information relative to the data-mining and preliminary analysis performed by other groups of the LanGeLin (2015) can be requested directly to each group.

1. Data-mining

1.1. Linguistic data-mining

The York group was involved in the collection of the linguistic data. The languages chosen by the group for the data collection are languages representative of the linguistic diversity within Eurasia, plus some outlier languages, e.g. the Wolof (West Africa). The data collection was made using the syntactic characteristics of each language and using the Parametric Comparison Method (Longobardi, 2003; Guardiano & Longobardi, 2005). The languages present in the linguistic dataset will serve as indicators for which populations to sample in the geographic and genetic datasets. In Table 2 it is annotated the languages, respective linguistic families and language codes of each language used in this project. In Figure 2 are represented the geographical locations of each linguistic family and sub-family according to the locations given in Table 3.

Table 2 - Languages used for the analysis. IE – Indo-European, LanCode – Language Code

Family	Language	LanCode
Basque	central Basque	cB
	western Basque	wB
Eskimo-Aleut	Inuit	Inu
IE-Celtic	Irish	Ir
	Welsh	Wel
IE-Germanic	German	D
	Danish	Da
	English	E
	Icelandic	Ice
	Norwegian	Nor
IE-Greek	Cypriot Greek	CyG
	Greek	Gre
IE-Indo-Iranian	Farsi	Far
	Hindi	Hi
	Marathi	Ma
	Pashto	Pas
IE-Romance	Calabrese	Cal
	French	Fr
	Italian	It
	Portuguese	Ptg
	Romanian	Rm
	Salentino	Sal
	Sicilian	Sic
	Spanish	Sp
Japonic	Japanese	Jap
Mongolic	Buriat	Bur
Niger-Congo	Wolof	Wo
Semitic	Arabic	Ar
	Hebrew	Heb
Sino-Tibetan	Cantonese	Can
	Mandarin	Man
Turkic	Turkish	Tur

Family	Language	LanCode
IE-Slavic	Bulgarian	Blg
	Polish	Po
	Russian	Rus
	Serb-Croat	SC
	Slovenian	Slo

Family	Language	LanCode
Uralic	Estonian	Est
	Finnish	Fin
	Hungarian	Hu

1.2. Geographic data-mining

The acquisition of the geographic data, made by the Bologna group, was performed by sampling the latitude and longitude of a chosen city within the distribution of each language in study. In the case of languages with a national distribution, the city chosen was the country capital, in the case of languages whose distribution was not national (languages whose distribution was just confined to a part of the country or a region that extends through more than one country) the city chosen was a city central to the language distribution. To confirm the distribution of each language, the linguistic database *Ethnologue* (Lewis, Simons, & Fennig, 2015; Ethnologue, 2015) was consulted. In Table 3 are listed the cities used for geographic sampling for each of the languages. In Figure 2 are represented the geographical locations of each linguistic family and sub-family according to the classification given in Table 2.

Table 3 - Geographic locations chosen for each language. To note that Khyber Pass is not a city but a region within the geographic distribution of the Pashto language and thus chosen for its geographic sampling

Language	Location
Arabic	Riyadh
Bulgarian	Sofia
Buriat	Ulan-Ude
Calabrese	Catanzaro
Cantonese	Hong Kong
central Basque	Donostia
Cypriot Greek	Nicosia
Danish	Copenhagen

Language	Location
Irish	Dublin
Italian	Rome
Japanese	Tokyo
Mandarin	Beijing
Marathi	Mumbai
Norwegian	Oslo
Pashto	Khyber Pass
Polish	Warsaw

Language	Location
English	London
Estonian	Tallinn
Farsi	Tehran
Finnish	Helsinki
French	Paris
German	Berlin
Greek	Athens
Hebrew	Jerusalem
Hindi	New Delhi
Hungarian	Budapest
Icelandic	Reykjavík
Inuit	Cape Dezhnev
Portuguese	Lisbon
Romanian	Bucharest
Russian	Moscow
Salentino	Lecce
Serb-Croat	Zagreb
Sicilian	Palermo
Slovenian	Ljubljana
Spanish	Madrid
Turkish	Ankara
Welsh	Cardiff
western Basque	Bilbao
Wolof	Dakar

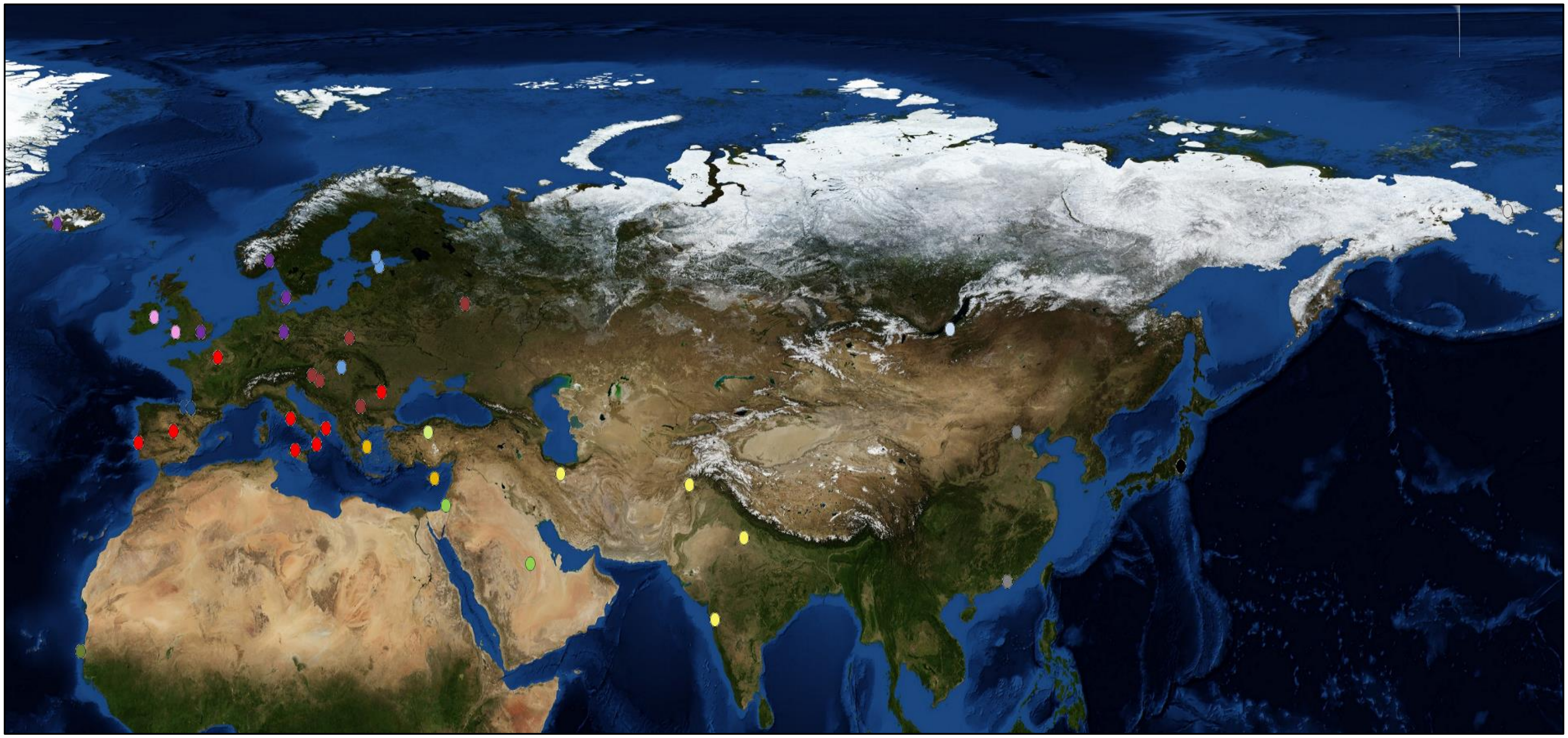

Figure 2 - Geographical representation of each Linguistic family and subfamily. The colour scheme represents each family. Dark Blue – Basque; Light Grey – Eskimo-Aleut; Pink, IE-Celtic; Purple – IE-Germanic; Dark Yellow – IE-Greek; Light Yellow – IE-Indo Iranian; Bordaux– IE-Slavic; Red – IE-Romance; Black – Japonic; Sky Blue – Mongolic; Dark Green – Niger Congo; Light Green – Semitic; Dark Grey – Sino-Tibetan; Chartreuse – Turkic; Light Blue – Uralic. (NASA, 2015)

1.3. Genomic data-mining

The genomic data collection followed some *a priori* rules. The genomic data was thus collected from individuals whose mother language was one of the selected by the linguistic group (Table 2), as well as having well establish genealogical roots within the geographic region of his language distribution. The sampling process of the authors was confirmed to not be biased towards any given character and the individuals sampled were independent, thus to obtain the most realistic representativeness of the populations genetic diversity.

This dataset was created by data-mining genetic databases, published papers and information gathered by personal communications. The same data was sometimes found in the genetic databases and in papers, in those cases the authorship of the samples was annotated but the information was retrieved directly from the database in order to retain the accession numbers. In the Table 4 is indicated the number of samples obtained by language for each of the molecular markers, and in Table 15 in Annex: 1 Data-mining is annotated for each language the author, paper and number of samples retrieved for the mtDNA HVR 1 dataset.

Some languages, e.g. French or Welsh, have a clear and recognizable location, and if not indicated otherwise, samples identified as French or Welsh individuals are considered for the study. In contrast with the clearly located French languages, some languages are restricted to a geographic region within a given country, e.g. central Basque or Marathi. In the case of languages whose linguistic region doesn't correspond to the geographic region of a given country but rather a small part of it or a continuous part expanding by multiple countries, a database, *Ethnologue* (Lewis, Simons, & Fennig, 2015; Ethnologue, 2015), was used to obtain information of the geographic delimitations were samples could be retrieved.

Since the creation of the genetic databases depends on the information available to public, not all the languages could be data-mined for the two kinds of *time-biased* markers. In the Table 4 can be observed in which languages was possible to obtain genetic information for the fast- and slow-evolving markers datasets (hereafter called fast- and slow-evolving datasets). Since this is a project still in course it's hoped that not just the language dataset can cover more languages but also the genetic datasets increase in the level of information.

Table 4 - Sample number per population per marker. “—” indicate populations for which genetic information could not be retrieved.

Language	Fast-evolving markers		Slow-evolving markers	
	mtDNA HVR1	Y-chr STRs	mtDNA SNPs	Y-chr SNPs
Arabic	1892	1980	1879	1859
Bulgarian	883	96	996	908
Buriat	473	215	—	—
Calabrese	—	—	145	94
Cantonese	205	510	—	—
central Basque	407	215	446	217
Cypriot Greek	91	418	—	—
German	667	2063	1331	349
Danish	683	290	449	106
English	172	488	—	—
Estonian	48	124	—	—
Farsi	465	79	—	—
Finnish	587	416	951	536
French	824	556	890	558
Greek	453	579	616	92
Hebrew	233	38	—	—
Hindi	220	196	144	744
Hungarian	435	632	20	407
Icelandic	433	100	—	—
Inuit	142	287	—	—
Irish	319	949	20	796
Italian	1515	2231	1976	975
Japanese	277	1643	—	—
Marathi	215	162	—	—
Mandarin	231	597	—	—
Norwegian	343	72	659	72
Pashto	230	611	273	297
Polish	817	2200	436	415
Portuguese	1088	1445	534	795
Romanian	105	406	92	216

Language	Fast-evolving markers		Slow-evolving markers	
	mtDNA HVR1	Y-chr STRs	mtDNA SNPs	Y-chr SNPs
Russian	522	944	568	1437
Salentino	—	—	91	39
Serbo-Croat	154	2058	594	302
Sicilian	—	—	915	508
Slovenian	233	102	104	75
Spanish	603	1142	781	507
Turkish	46	60	29	523
western Basque	—	—	189	79
Welsh	92	118	—	—
Wolof	59	74	—	—
Total	16162	24096	15128	13574

2. Data processing

In this section it will only be explained in detail the processing of the mtDNA HVR1 database which I was tasked to assemble and prepare. A small explanation of the preparation of the SNPs databases is also given, even though not in full detail do to both being handle by other groups.

2.1. Genetic

Genetic information can be retrieved in different forms. Information on a population's genomic characteristics can be found as nucleotide sequences or haplotype (point mutations relative to a reference sequence) for each individual, or as haplogroup frequencies for the all population. The information can also be given for the all chromosome or just partial. On top of these work-flow problems, problems on the sequences themselves can arise, such as errors being introduced during sequencing, due to the imperfection of the current sequencing systems to retrieve the correct data from the DNA molecules. These errors can be simple false *indels*, substitutions, nucleotide positioning errors and even expansion of short tandem sequences, errors which, depending on their number, can lead to results not representative of the true genomic diversity on a population, thus injuring population genetic studies, but can also have a negative impact in medical studies, by misrepresentation of the true genome of a give individual (Bandelt, Lahermo, Richards, & Macaulay, 2001; Bandelt, Quintana-Murci, Salas, & Macaulay, 2002; Yao, Bravi, & Bandelt, 2004; Brandstätter, et al., 2005; Bandelt & Kivisild, 2006; Yao, Salas, Logan, & Bandelt, 2009).

Due to the constraints pointed-out above, before proceeding to analysis of the data, a preparation was performed to eliminate the possible errors and have the data in the desired format.

mtDNA HVR1 – For this dataset all the information was retrieved in the nucleotide sequence or haplotype format, and with samples presenting different nucleotide lengths. Thus a conversion for the nucleotide sequence and a constraint of the sequence length were performed. Firstly the *Haplosearch* tool (Fregel & Delgado, 2011) provided in the *Haplosite* (Haplosite, 2015) page was used to convert all our samples present in haplotype format to sequence format, conversion which used has a reference template the revised Cambridge Reference Sequence – rCRS (Andrews, et al., 1999), a revised version of the earlier Cambridge Reference Sequence published by Anderson et al. (1981). Secondly, a multi-alignment was performed with all the sequences using the *MUSCLE* (Edgar, 2004) program. After the alignment run, using the sequence editing capability of *BioEdit* (Hall, 1999) the sequences were trimmed to the HVR 1 region using the rCRS (Andrews, et al., 1999) has reference for the positions (from nucleotide 16024 to 16569). Due to the appearance of sequencing *indels* all over the samples, in higher or lesser degree, after the trimming to the HVR1 region, a check-up was performed on the aligned sequences to identify which might be true polymorphisms or sequencing artifacts. Those appearing in less than 10 individuals of all the dataset were considered to be an artifact. For the region of the HVR1 between positions 16180 to 16193, problems with the alignment and nucleotide identification were spread through all dataset. This problem is caused by the appearance of an Adenine stretch followed by a Cytosine stretch which causes the sequencing program to mislabel the true nucleotides and their position, and the alignment software to further misalign the sequences (Brandstätter, et al., 2005; Bandelt & Kivisild, 2006). In order to correct the problem, the polymorphism present at the position 16190 was maintained and the rest of the problematic region was modified to be consider monomorphic in relation to the reference sequence. Thirdly, after the check-up and correction of the sequence a second trim was performed, this due to some samples just comprise not the totality of the HVR1 region, but a smaller part of it, that is, some sequences presented missing values at both ends of the HVR 1 sequence length. The higher the percentage of missing values present the higher the hinder will be on the subsequent analysis of genetic diversity. So, an analysis of the data was made to check the extent of the HVR1 that presented a lower missing values percentage but still preserving the maximum of data possible. The region with the HVR1 selected was from position 16024 to 16383 (has always, the nucleotide positions are in relation to the complete rCRS (Andrews, et al., 1999)). The trimming was performed using the *BioEdit* (Hall, 1999) program. The "N" symbol was assigned to all the missing values of the dataset, serving has a marker for the programs in which the analysis of the sequences would be performed. During the multiple check-ups performed in our data, some problems arose in some samples, such as samples duplicated for a given author, samples presenting high number of missing values within the selected HVR1 region, or samples presenting a general problematic nucleotide assembly for our region of interest. These problems where corrected by the elimination of these samples, thus, the number of samples identified in each of the used papers may not correspond to the final number of samples used by us.

SNPs: to assemble the information for the two SNPs databases, the Y chromosome and mitochondrial DNA, data was retrieved both in haplotype form and as haplogroup frequencies. In order to have all the data in the same format, using the haplotype information of each sample, their haplogroup was obtained and the haplogroup frequencies of each population calculated.

3. Data analysis

All the following analysis, excluding the diversity analysis on all of the raw datasets, excepting the mtDNA HVR1 dataset, leading to the acquisition of the dissimilarity matrices was performed by me, for this reason only a small explanation is dedicated to the analysis of each raw dataset for the acquisition of their dissimilarity matrices were I was not involved. This is made to give a better comprehension on the material being used for the subsequent analysis involving all datasets.

3.1. Diversity analysis

Different data kinds, such has genomic, geographic and linguistic data, are retrieved in different formats, e.g., genomic information can be retrieved in nucleotide sequence while geographic information can be retrieved in latitude/longitude values. This implies that a direct comparison of these data is impossible, thus a medium/format through which different data information can be compared is needed, one such way is using distance or dissimilarity matrices for each data kind. These distance/dissimilarities are thus after comparable, even if using different units, for they allow us to compare the different data by comparison of the patterns of relationship within each. Taking this into account, each dataset was analysed using specific diversity indices, resulting in the acquisition of distance/dissimilarity matrices were the degree of dissimilarity between each population is expressed.

3.1.1. Genomic data

The first analysis performed using the genomic datasets was on its diversity, that is, we analysed the data for their intra and inter-population diversity. To obtain the indices of diversity an AMOVA (Excoffier, Smouse, & Quattro, 1992) was performed on each of the projects genomic datasets. This analysis was performed using the Arlequin v.3.5.1.2 (Excoffier & Lisher, 2010).

Due to the datasets being different in their genomic information, different indices were used to compute the diversity of each data. For the mtDNA HVR1 dataset, being it constituted by nucleotide sequence information, the diversity index used to obtain the dissimilarities between populations was the *Phi-st* (Excoffier, Smouse, & Quattro, 1992). The diversity index used for all the other genomic datasets was the *Fst* index (Wright, 1951), for it calculates the diversity between populations based on the allelic information, which is on what these datasets are based. For both the calculations a 5% missing value was allowed.

With the results of the AMOVA (Excoffier, Smouse, & Quattro, 1992) analysis dissimilarity matrices were created for each of the genomic datasets. These dissimilarity matrices will be therefore the information to be used for the subsequent analysis.

3.1.2. Geographic data

Using the Great Circle distance (Weisstein, 2015) methodology, the distance between each location used for the project was computed. This task was taken by others, responsible for the collection of this data and its preliminary analysis, to obtain a matrix of distance between locations. This matrix of location distances was used to compute the analysis of relation between all the data.

As indicated in Methodology: 1.3 Genomic data-mining and shown in Table 4, the genomic data set, other than being divided by uniparental markers, is divided by markers presenting different rates of evolution. In the same paragraph and table it is possible to observe that for the different marker sets some differences exist regarding the populations in study. In order to adequate our geographic data, our dataset was partitioned in two datasets each containing the information relative to the populations identified for each of the markers in study, and as indicated above being, when needed, indicated as fast- or slow-evolving datasets.

3.1.3. Linguistic data

Using the PCM (Longobardi, 2003; Guardiano & Longobardi, 2005) data, the group responsible for the linguistic data obtained dissimilarity values between each language pair using the Jaccard distance (Jaccard, 1908) method. This dissimilarity matrix was used to compute the analysis of relation between all the data.

For the same reason expressed in the previous point, also for the linguistic dataset, two datasets were devised, each containing the information relative to the populations in study for the fast- and slow-evolving markers of the genomic dataset.

All the following analysis of the data was performed in the statistical environmental *R* (R Core Team, 2015). These analyses were performed using *functions* present at the base package and freely available packages on *CRAN* (CRAN, 2015). In the description of the methodology opted for the analyses performed thereafter, when pertinent, the codes used are indicated.

3.2. Auxiliary materials

To help in the interpretation of the graphical results of the Multidimensional Scaling (Shepard, 1962; Kruskal, 1964a; Kruskal, 1964b) computations, some additional material was created. That material consists of tables and matrices and with label and colour schemes and vector files containing the names of the populations.

For the Multidimensional Scaling (Shepard, 1962; Kruskal, 1964a; Kruskal, 1964b) plots as supplementary material it was used a vector containing the populations LanCode (Table 2) and a table containing the LanCodes and an assigned colour for each population taking in consideration their linguistic family, that is, populations whose language belongs to a same linguistic family were assigned the same colour. This was intent to allow to identify not just the population represented but also its linguistic family in each plot created. It is here important to notice that sub-families within a same linguistic family were assigned a different colour, thus e.g., it is possible to distinguish by colour identification the different sub-families of the Indo-European linguistic family.

For the creation of these auxiliary materials, those created in *R* (R Core Team, 2015), *functions* of the *R base* (R Core Team, 2015) package and *gdata* (Warnes, et al., 2015) package were used.

3.3. Multidimensional Scaling

The Multidimensional Scaling (Shepard, 1962; Kruskal, 1964a; Kruskal, 1964b) computation was performed over each of the datasets using *functions* from the *MASS* (Venables & Ripley, 2002) package and making use of auxiliary material for help in the interpretation of the graphical representation of each result.

Firstly the distance matrices of each dataset and all the auxiliary material were loaded into the working space, and any needed preparation, such has row and column ordering by a given order, were executed. This was followed by the loading of the *MASS* (Venables & Ripley, 2002) package to the work space.

The Multidimensional Scaling (Shepard, 1962; Kruskal, 1964a; Kruskal, 1964b) computations were performed in each of the matrices using the `isoMDS` *function* has follows:

```
MDS.result <- isoMDS (matrix, k=2)
```

Where:

- *`MDS.result`* it's the assigned object for the computation results to be stored;
- *`matrix`* it's the matrix upon which the computation will be performed;
- *`k`* it's the number of dimensions on which we want the multidimensional scaling to be resolved.

The results obtained, which can be visualized by calling the *`MDS.results`* object, are a two dimensional set of coordinates for each population on the matrix, plus the *stress value* of the computation given in percentage. These table with coordinates by themselves are little useful to drawn any information, thus graphical representation of the results were created to enhance the interpretation of the obtained MDS (Shepard, 1962; Kruskal, 1964a; Kruskal, 1964b) computation.

To create the graphical representations of the MDS (Shepard, 1962; Kruskal, 1964a; Kruskal, 1964b) analysis the `plot`, `text` and `legend` *functions* of the *base* (R Core Team, 2015) package were used has follows:

```
plot(MDS.results$points, pch=NA, main="Matrix isoMDS")

text(MDS.results$points, labels=vectornames, col=mfcolor[,2])

legend("topright", MDS.results$stress, title="Stress")
```

Where:

- *MDS.results$points* is the coordinates information;
- *pch* is an argument for the symbol to be used for each coordinate set, in our case being the non-inclusion of a symbol;
- *main* is the argument for the title of the plot;
- *labels* is the argument for the names to be used for the labeling of each coordinate in the plot;
- *col* is an argument used to indicate the color chosen for each label in the plot
- *"topright"* is the argument indicating the position for the legend to appear in the plot screen;
- *MDS.results$stress* is the argument for the information to be shown on the legend;
- *title* is the argument for the legends' name.

In some cases a zoom-in was needed to better discern the distribution of the different populations. In those cases a restriction was created on both dimensions as needed by adding the `xlim` and `ylim` arguments in the `plot` *function* line of the script, as follow:

```
plot(MDS.results$points, … , xlim=c(-0.1,0.1), ylim=c(-
0.1,0.1))
```

3.4. Mantel tests

The *Mantel* and *partial Mantel tests* (Mantel, 1967) were performed using functions of the *vegan* (Legendre & Legendre, 1998; Oksanen, et al., 2015) package.

The *Mantel* and *partial Mantel tests* (Mantel, 1967) were performed in all the possible combinations between the genetic, geographic and linguistic data, taking in consideration the analysis were performed between datasets containing the same populations, that is, only within datasets designed for the populations studied either for the slow- or fast-evolving markers.

Prior to the tests computation, adjustments of the matrices were performed has needed, like ordering of row and columns by a defined order, to ensure the right performance of the tests.

To compute the tests the functions `mantel` and `mantel.partial` for the *Mantel* and *partial Mantel tests* (Mantel, 1967), respectively, were used as follows:

```
mantel(xdist, ydist, method="pearson", permutation=10000)

mantel.partial(xdist, ydist, zdist, method="pearson",
permutation=10000)
```

Where:

- `xdist`, `ydist` and `zdist` are the matrices upon which the analysis will be executed;

- `method` is the method selected to performed the computation;
- `permutation` is the number of permutations chosen to be executed before the result is drawn

To note that the in the vegan package the permutations are only made in one of the matrices (Oksanen, et al., 2015), and that the number of maximum unique permutations allowed for each data was calculated using the `numPerms` function on the *vegan* (Oksanen, et al., 2015) package wielding values of 3.7199E^{41} and 4.0329E^{26} for the fast- and slow-evolving data, respectively. Due to the high computational power needed for this high number, a number of 10000 permutations was chosen after repeated tests to confirm the stability of the results.

The significance level chosen *a priori* for the analysis of the statistical significance of the correlations is $\alpha = 0.05$.

3.5. Scatterplots

With the aim to have a graphical perspective of the correlation between the genomic, geographic and linguistic variables, using the different matrices, scatterplots of these variables against each other were drawn.

The construction of these representations was performed by using two matrices at a time using their distance values as X and Y coordinates, thus computing the distance values distribution of each matrix against the other. These plots will allow to visualize how the distance distribution of a given variable behaves in relation to the distance distribution of another variable, e.g., if the population pairs of a given linguistic family has a wider variability for the Y-chr dataset in comparisons to the mtDNA set.

To facilitate the identification of the data being represented in each scatterplot they will be identified by the data used to compute it, e.g., Ling. vs mtDNA HVR 1 will be used to identify the scatterplot where the linguistic and mtDNA HVR 1 data were used.

These graphical representations were create making use of the `plot` *function* from the *base* package of *R* (R Core Team, 2015) as follows:

```
plot(xdit, ydist, main="X dist vs Y dist", xlab="Xdist",
ylab="Ydist")
```

Where:

- *xdist*, *ydist* are the distance matrices to be used;
- *main* is the argument for the title of the plot;
- *xlab* and *ylab* is the argument to assign a label name to the x and y coordinates;

Results and Analysis

1. Dissimilarity analysis

From the analysis of the variability within each dataset (as described in Methodology: 3.1 Diversity analysis) matrices containing the genomic, geographic and linguistic distances between each population were obtained. Hereafter all the subsequent analysis will be performed using these matrices, this is thus to enable us to correlate the different information with each other. Bellow it follows example matrices for each of the datasets and in the Annex: 2 Dissimilarity matrices can be found the complete matrices.

1.1. Genomic

1.1.1. Fast-evolving markers

Below are present sub-set tables of the dissimilarity matrices obtained by the diversity analysis of the genomic data for the fast-evolving data, mtDNA HVR 1 and Y-chr STRs (Tables 5 and 6, respectively).

Table 5 - Exemplary sub-set of mtDNA HVR 1 dissimilarity matrix

	Ar	**Can**	**cB**	**Fin**	**Hi**	**Ptg**	**Tur**
Ar	0	NA	NA	NA	NA	NA	NA
Can	0.09206	0	NA	NA	NA	NA	NA
cB	0.0214	0.13147	0	NA	NA	NA	NA
Fin	0.01915	0.07794	0.02066	0	NA	NA	NA
Hi	0.06679	0.04419	0.11411	0.06735	0	NA	NA
Ptg	0.00697	0.09018	0.01345	0.00873	0.07294	0	NA
Tur	0.01374	0.04319	0.04027	0.01504	0.0463	0.01482	0

Table 6 - Exemplary sub-set of Y-chr STRs dissimilarity matrix

	Ar	**Can**	**cB**	**Fin**	**Hi**	**Ptg**	**Tur**
Ar	0	NA	NA	NA	NA	NA	NA
Can	0.24028	0	NA	NA	NA	NA	NA
cB	0.26931	0.24323	0	NA	NA	NA	NA
Fin	0.23867	0.24036	0.28434	0	NA	NA	NA
Hi	0.11849	0.29304	0.3095	0.25553	0	NA	NA
Ptg	0.13244	0.16368	0.05097	0.18725	0.20246	0	NA
Tur	0.03186	0.23649	0.29946	0.21943	0.08659	0.12194	0

1.1.2. Slow-evolving markers

Below are present sub-set tables of the dissimilarity matrices obtained by the diversity analysis of the genomic data for the slow-evolving data, mtDNA SNPs and Y-chr SNPs (Tables 7 and 8, respectively).

Table 7 - Exemplary sub-set of mtDNA SNPs dissimilarity matrix

	Ar	Cal	cB	Fin	Grk	Ptg	Tur
Ar	0	NA	NA	NA	NA	NA	NA
Cal	0.00116	0	NA	NA	NA	NA	NA
cB	0.04286	0.0265	0	NA	NA	NA	NA
Fin	0.0174	0.00927	0.02244	0	NA	NA	NA
Grk	0.02225	0.01468	0.02141	0.01231	0	NA	NA
Ptg	0.00576	0.00488	0.03656	0.02517	0.02303	0	NA
Tur	0.03083	0.04364	0.10175	0.04103	0.04886	0.05292	0

Table 8 - Exemplary sub-set Y-chr SNPs dissimilarity matrix

	Ar	Cal	cB	Fin	Grk	Ptg	Tur
Ar	0	NA	NA	NA	NA	NA	NA
Cal	0.06397	0	NA	NA	NA	NA	NA
cB	0.40104	0.35303	0	NA	NA	NA	NA
Fin	0.2991	0.29839	0.52198	0	NA	NA	NA
Grk	0.07114	0.00708	0.38588	0.26941	0	NA	NA
Ptg	0.23513	0.09597	0.08299	0.36476	0.12606	0	NA
Tur	0.0264	0.01037	0.32019	0.24011	0.0205	0.14528	0

1.2. Geographic

Below are present sub-set tables of the distance matrices obtained by the diversity analysis of the geographic data (Table 9).

Table 9 – Exemplary sub-set of Geographical distance matrix

	Ar	Can	cB	Fin	Hi	Ptg	Tur
Ar	0	NA	NA	NA	NA	NA	NA
Can	6818370.535	0	NA	NA	NA	NA	NA
cB	4940066.23	10325193.03	0	NA	NA	NA	NA
Fin	4289431.146	7846633.825	2652271.67	0	NA	NA	NA
Hi	3054432.415	3767050.755	7129602.172	5227074.383	0	NA	NA
Ptg	5427804.486	11045272.77	726784.0295	3364741.844	7784845.113	0	NA
Tur	2134789.481	7737965.14	2977988.732	2316693.094	4223356.487	3584266.376	0

1.3. Linguistic

Below are present sub-set tables of the dissimilarity matrices obtained by the diversity analysis of the linguistic data (Table 10).

Table 10 - Exemplary sub-set of Linguistic distance matrix

	Ar	Can	cB	Fin	Hi	Ptg	Tur
Ar	0	NA	NA	NA	NA	NA	NA
Can	0.522	0	NA	NA	NA	NA	NA
cB	0.324	0.231	0	NA	NA	NA	NA
Fin	0.306	0.32	0.167	0	NA	NA	NA
Hi	0.278	0.25	0.206	0.211	0	NA	NA
Ptg	0.295	0.36	0.194	0.237	0.158	0	NA
Tur	0.424	0.231	0.207	0.125	0.214	0.324	0

2. Multidimensional Scaling

2.1. Fast-evolving dataset

2.1.1. Genomic

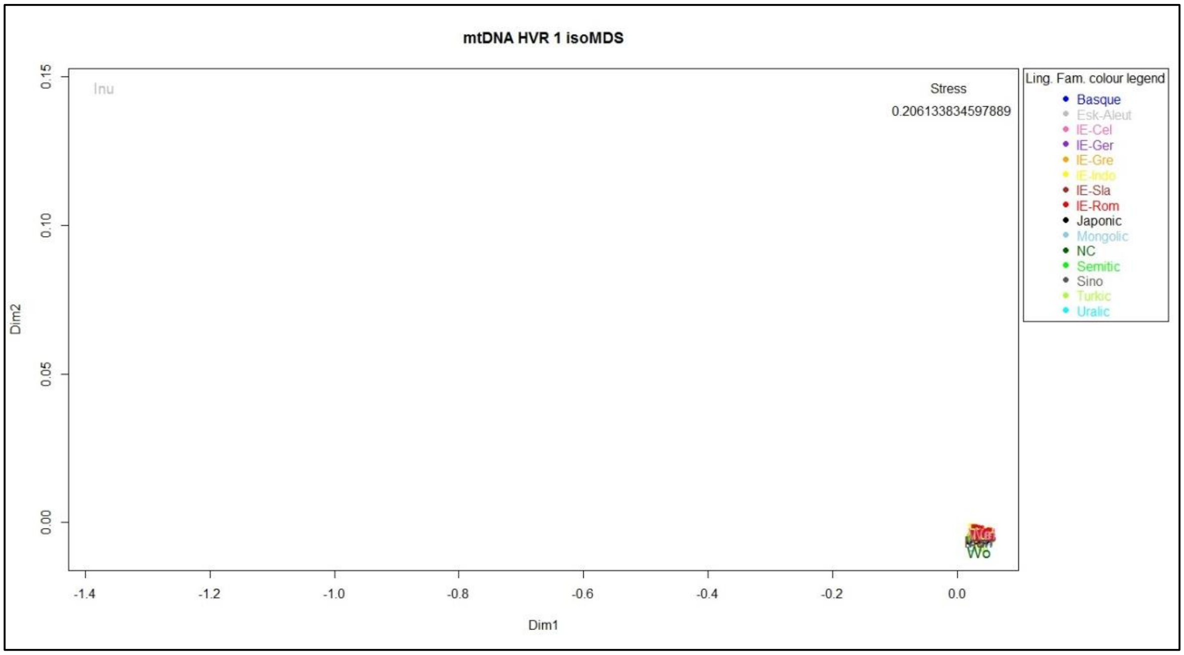

Figure 3 **- Graphical representation of MDS computation for the mtDNA HVR 1 dataset**

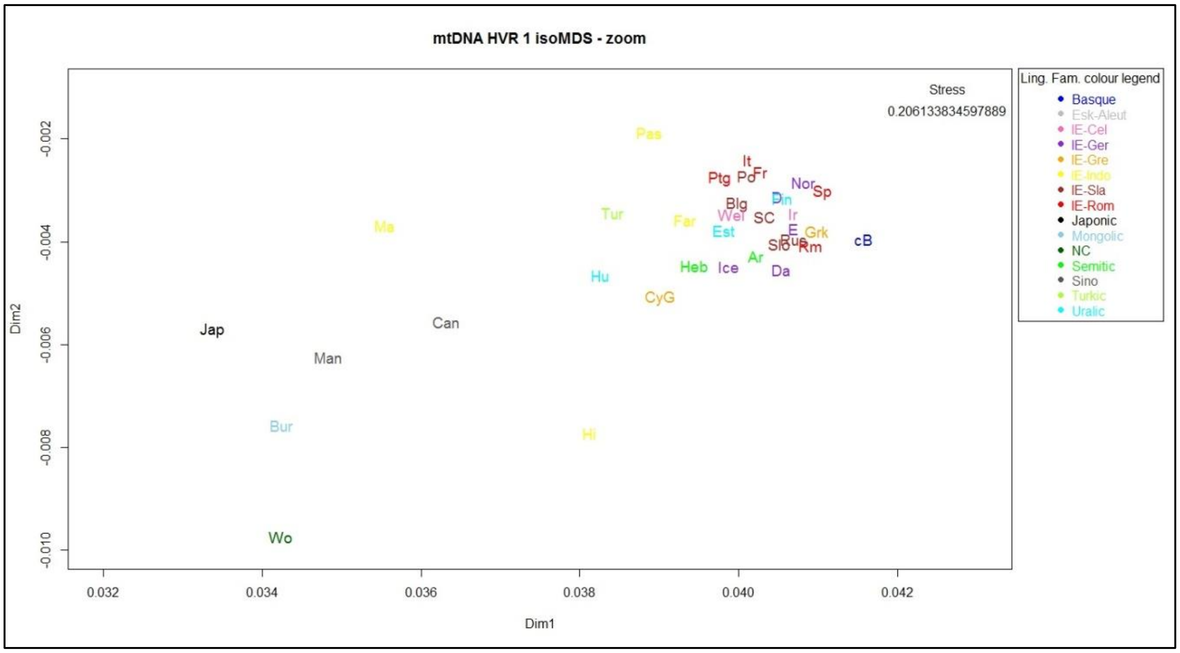

Figure 4 - Graphical representation of zoomed MDS computation for the mtDNA HVR 1 dataset

With an excellent *stress value* of 0.21%, the MDS computation of the mtDNA HVR 1 dataset presents little distortion from the input data. When analysing the graphical representation of the MDS computation for the mtDNA HVR 1 dataset (Figure 3) the Inuit population is clearly demarked from all other populations that form together a tight cluster, leaving the Inuits as outliers. When zooming on the population cluster a different picture appears (Figure 4), were its possible to see, with some exceptions, a clustering to the right of the graphic of more western populations and to the left of eastern populations, with exception for the Farsi and Pashto, intruding in the western cluster, and the Wolof, intruding on the eastern cluster.

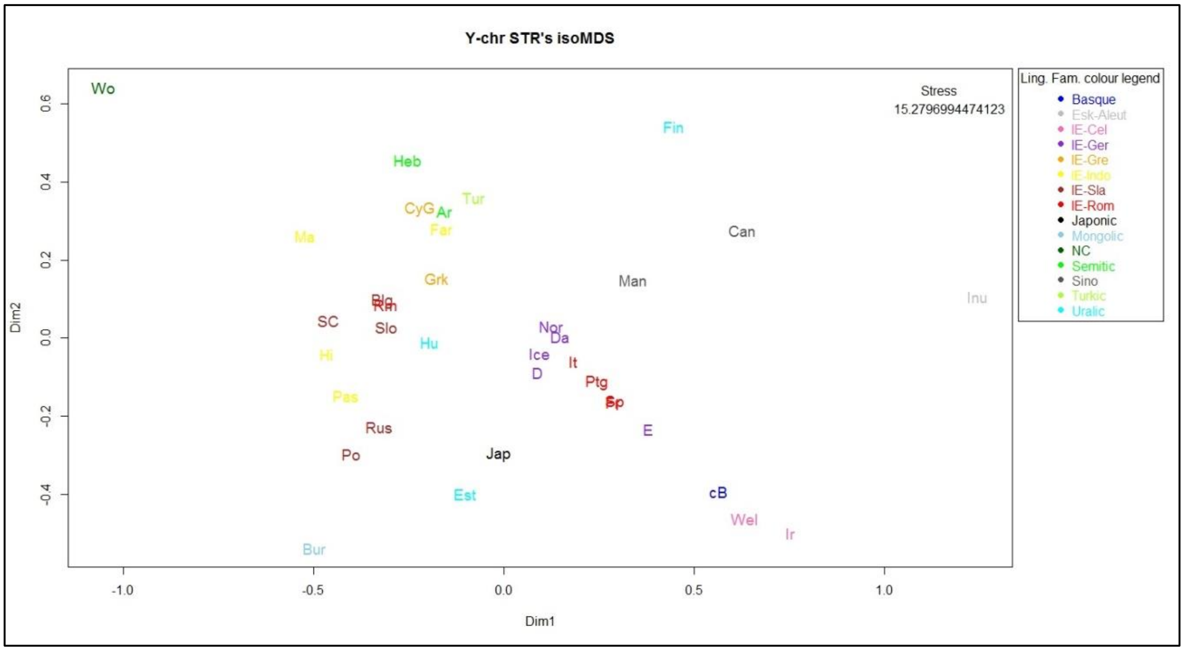

Figure 5 - Graphical representation of MDS computation for the Y-chr STRs dataset

With a *stress value* of 15.28%, the MDS computation for the Y-chr STRs dataset (Figure 5) even though acceptable, presents some degree of distortion in relation to the matrix input data. In a first look Wolof and Inuit appear as outliers and some geographical structure is apparent, even though not very clear. It is possible to see that some populations clustering together with their geographical neighbours rather than with their linguistic neighbours, such as Romanians with Bulgarians instead with the IE-Romance languages, Farsi and Cypriot Greeks with Arabs and Turks instead with IE-Indo-Iranians and IE-Greeks respectively. Japan here is placed surrounded by European populations and the Uralic populations appear dispersed throughout not clustering together or even with their geographical neighbours.

2.1.2. Geographic

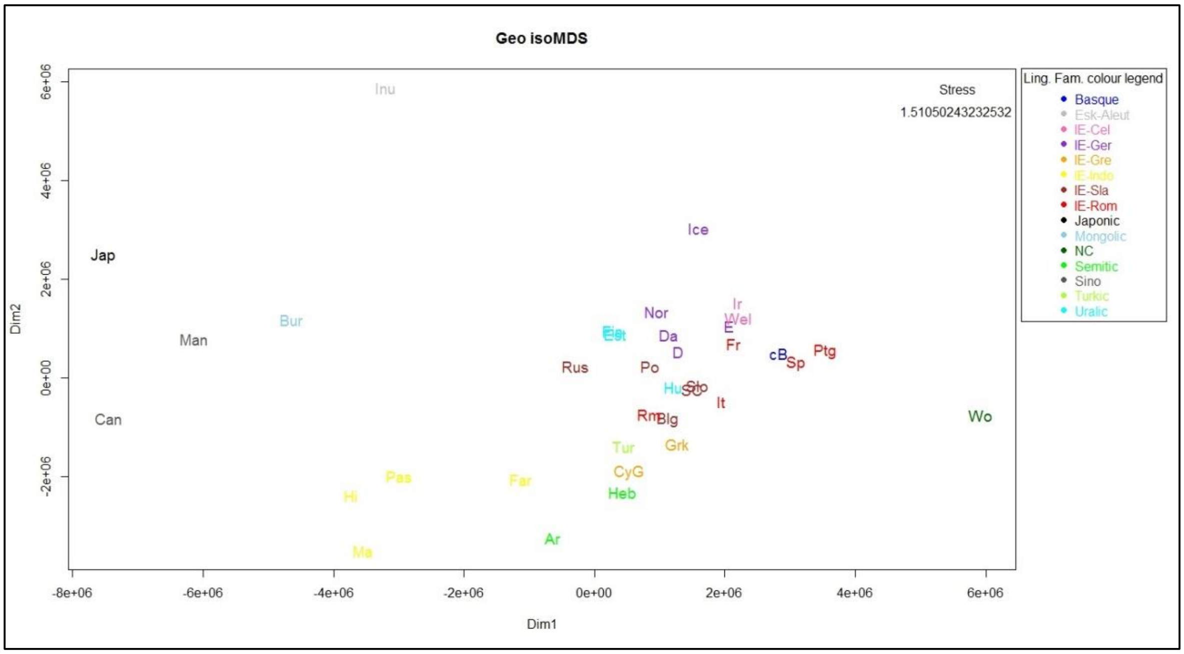

Figure 6 - Graphical representation of MDS computation for the Geographic fast dataset

By looking at the graphical representation of the MDS computation for the geographical data of the fast-evolving dataset (Figure 6) it is possible to see a clear resemblance to the actual geographical representation of each country of the dataset, with a clear east-west and north-south orientation of the samples (here, west-east orientation mirrors that of a real map), with just little dissimilarities visible. This goodness-of-fit can be verified by the low *stress value* given by the computation, 1.51%, which is consider to be an excellent *stress value*, indicative that there is little distortion of the data input to achieve this MDS result.

2.1.3. Linguistic

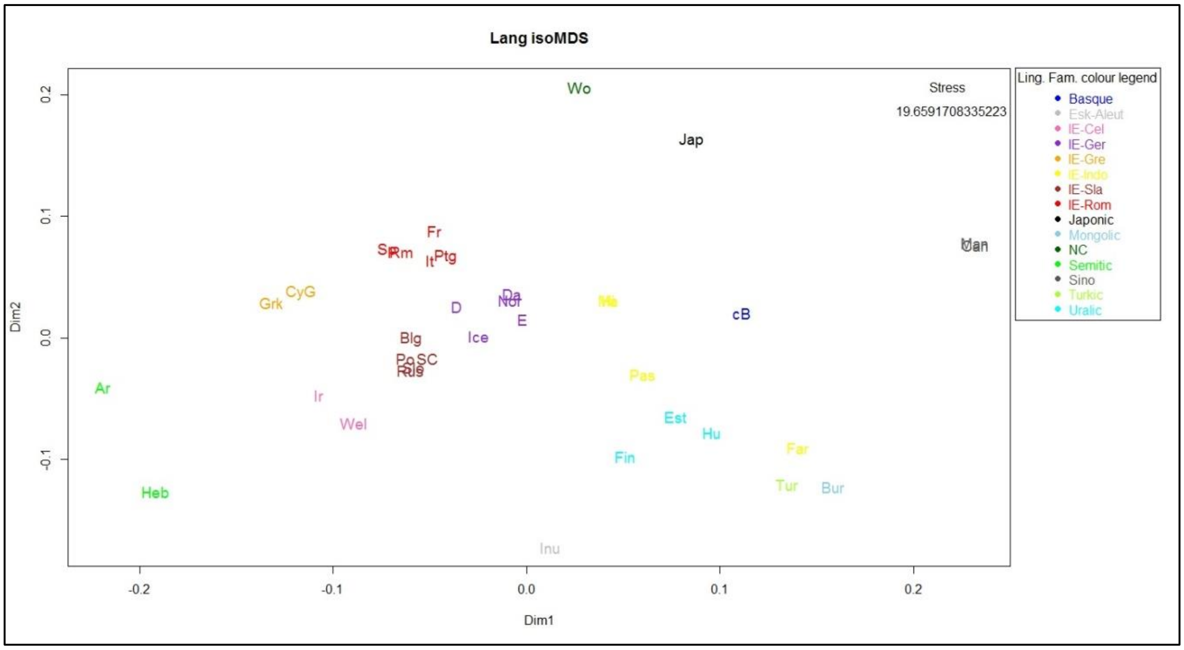

Figure 7 - Graphical representation of MDS computation for the Linguistic fast dataset

By the observation of the graphical representation of the MDS computation of the linguistic data for the fast-evolving dataset (Figure7), it is clear that languages belonging to a same linguistic family, and in the case of the Indo-European family tend to cluster together with some notable exceptions, such has the position of Farsi more close to Turkish and Buriat than the remaining IE-Indo-Iranian languages, having the Uralic linguistic group between the two clusters. It is also notable the isolation of the Central Basque language from its geographical neighbours.

No apparent geographical structuring is visible in the overall structure of the MDS.

A certain caution is need when analysing this graphic, and these “misplacements” of some populations out of what would be expected of them can be interpreted as data distortion since the *stress value* obtained, even though it can be consider fair, is still high: 19.66%.

2.2. Slow-evolving dataset

2.2.1. Genomic

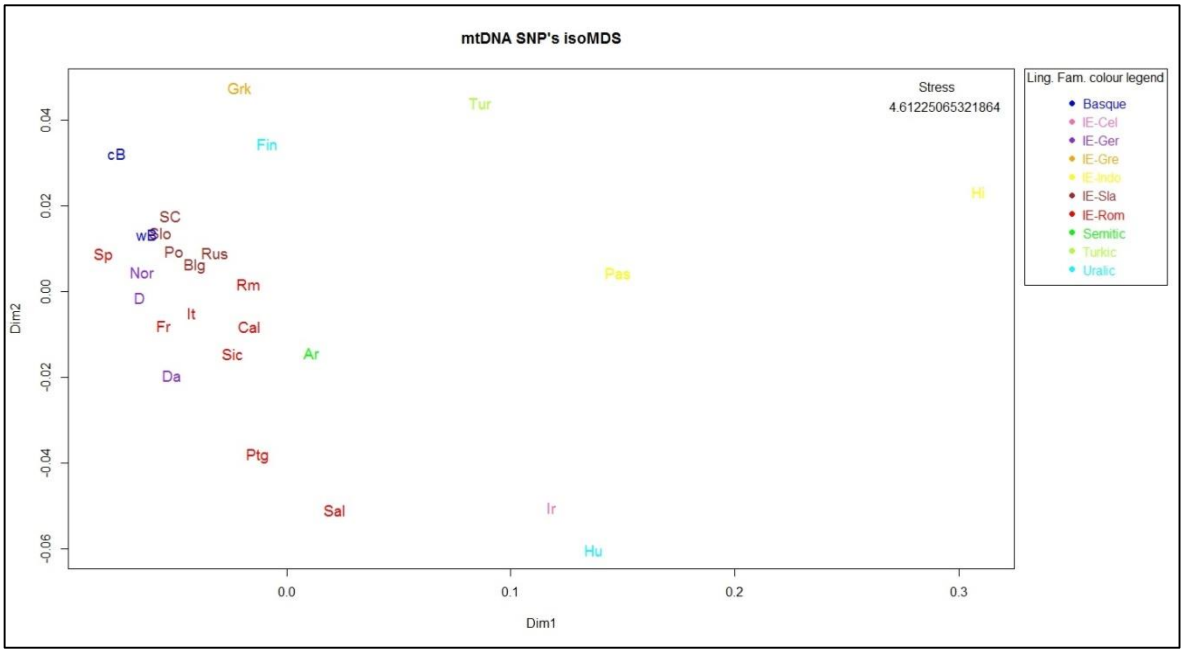

Figure 8 - Graphical representation of MDS computation for the mtDNA SNPs dataset

The good 4.61% *stress value* achieved for the MDS computation on the mitochondrial DNA SNPs dataset indicates that a small distortion of the results in relation to the input data is expected.

In the graphical representation of the MDS computation for the mtDNA SNPs dataset (Figure 8) it is visible a clustering of almost all of the European populations at the left side of the graphic, with the Indo-Iranian, Irish, Hungarian and Turkish populations appearing positioned as outliers. Within the outliers is noteworthy the case of the Hindi population which appears further apart than all the other outliers at the extreme right of the graph. Positioned next to the majority of the European populations appear the Basque, Arabic and Finnish languages. Within the European population languages tend to be positioned close to their linguistic neighbours with some notable exceptions. The IE-Germanic linguistic group appears divided by France, with Danish appear isolated from the rest. In the IE-Romance linguistic group, Spanish appears distant from Portuguese and separated from the main IE-Romance group by the Norwegian and Germanic populations, while Portuguese and Salentino appear as outliers of the IE-Romance cluster. Another interesting positioning is of the Western Basque populations that appear closer to IE-Slavic than to the Central Basque populations.

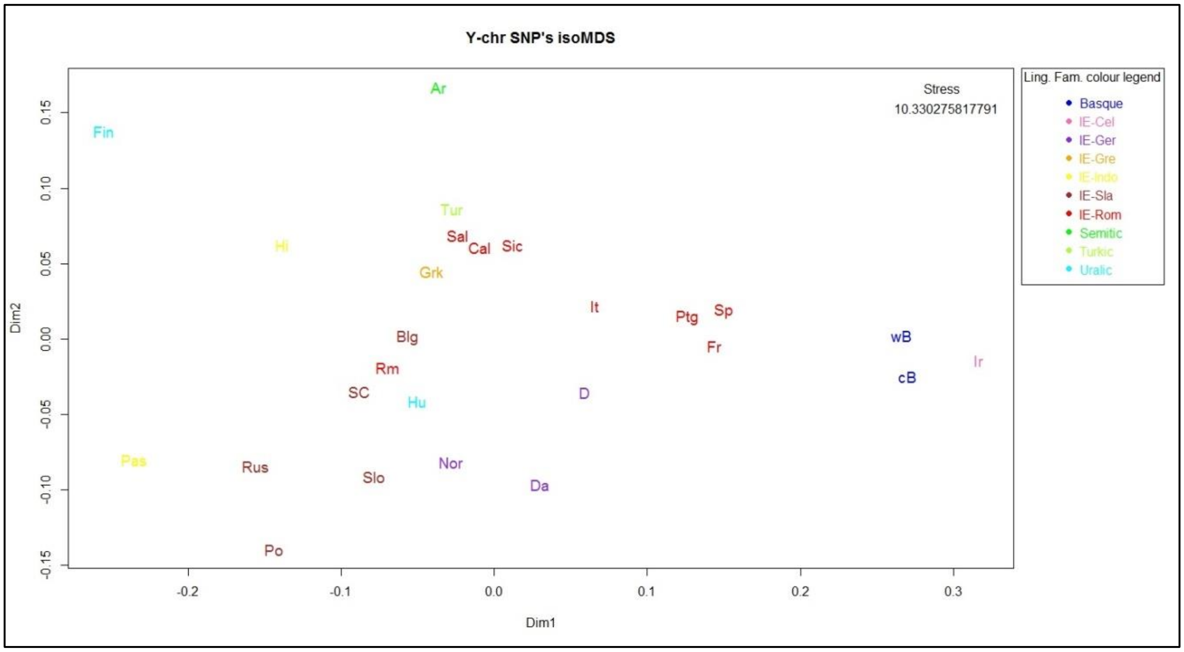

Figure 9 - Graphical representation of the MDS computation for the Y-chr SNPs dataset

The MDS computation for the Y chromosome SNPs dataset presents *stress value* of 10.33%, which indicates that there is a fair of goodness-of-fit between the computational results obtained and the input that, even though some distortion is present.

By analysing the graphical representation of the MDS computation for the Y chromosome SNPs dataset (Figure 9) it is possible to see that up to some extent populations tend to appear close to their geographical and historical neighbours rather than their linguistic. For instance, Salentino, Calabrese and Sicilian populations appear closer to Turkish and Greek population than to Italian population, which is half way between these and a cluster of southern IE-Romance languages. Romanian, the other IE-Romance language of the dataset appears in a cluster formed by its geographical neighbours. The Basque and Irish forming a cluster at the right of the graphic and Finnish alone at the left appear as outliers. With some clear distortions of the geographical reality, populations appear to be present with an east-west (left to right side of the graphic) and north-south (bottom to top side of the graphic) orientation.

2.2.2. Geographic

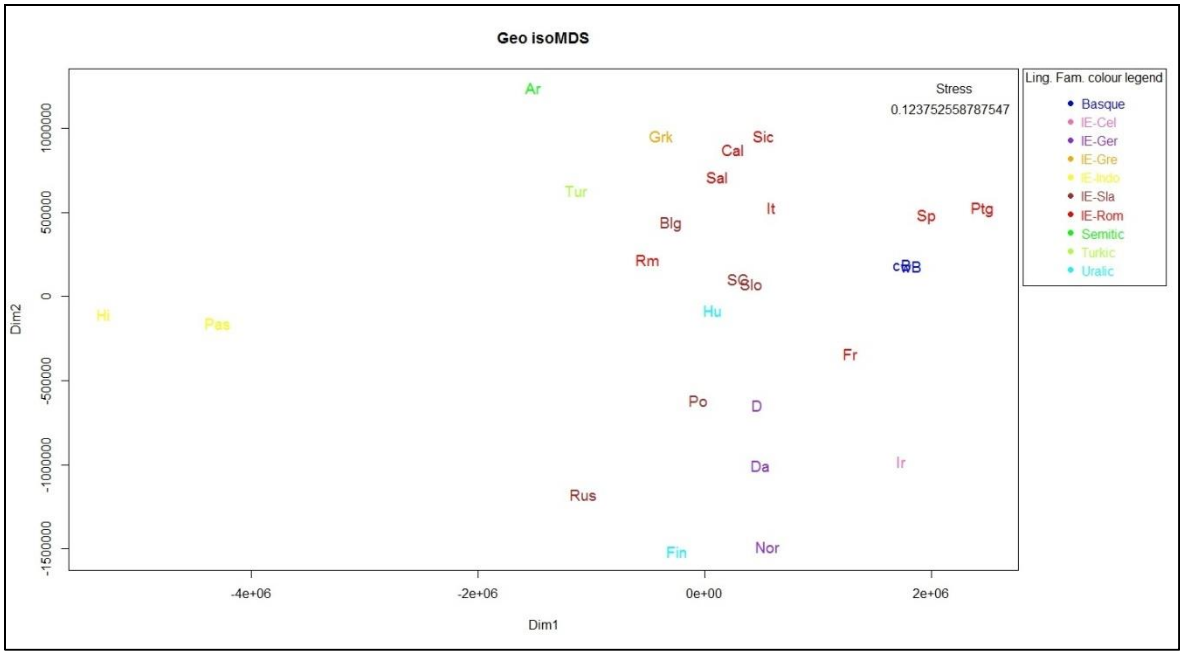

Figure 10 - Graphical representation of the MDS computation for the Geographic slow dataset

With an excellent *stress value* of 0.12%, the MDS computation of the Geographic slow dataset present minimal distortion in relation to the input data.

In the graphical representation of the MDS for the Geographic slow dataset (Figure 10) populations tend lay close to their geographical neighbours rather than their linguistic ones. A clear pattern of geographic orientation is visible with populations being clustered from east to west (left to right side of the graphic) and north to south (bottom to top side of the graphic).

2.2.3. Linguistic

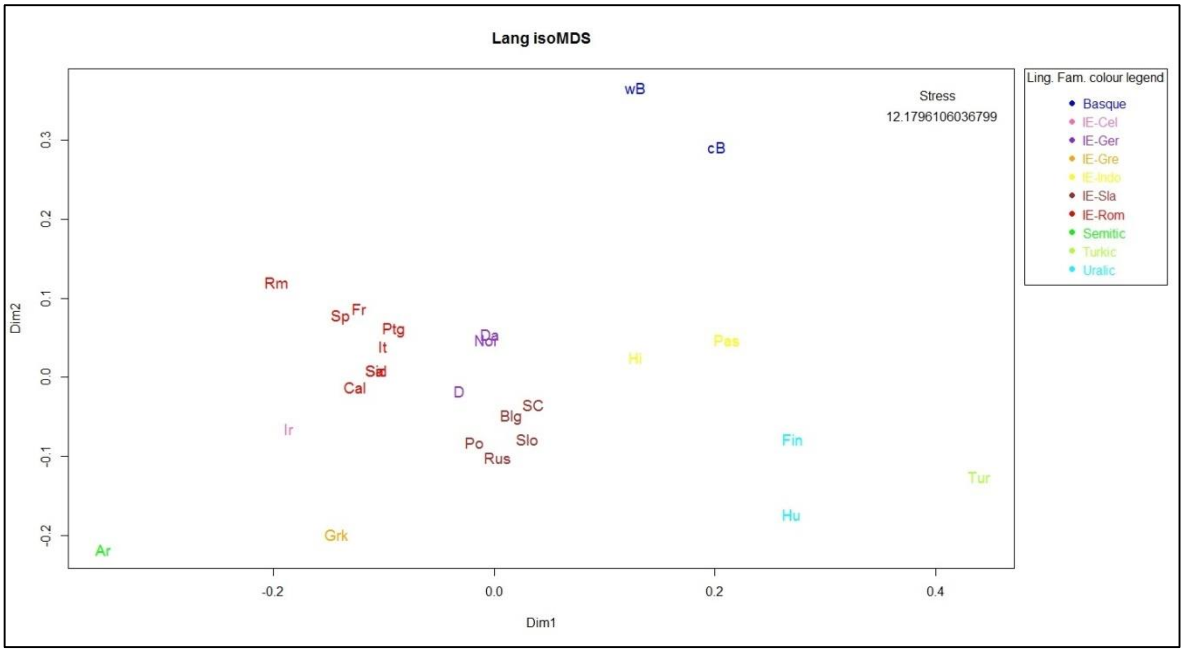

Figure 11 - Graphical representation of the MDS computation for the Linguistic slow dataset

The MDS computation for the Linguistic slow dataset presents a fair *stress value* of 12.18%, which lets us now that some distortion is present in relation to the input data.

On the MDS graphical computation of the Linguistic slow dataset (Figure 11) Turkish, Basque and Arabic languages appear as outliers. The languages cluster together primarily with their linguistic families rather than their geographical neighbours, with the IE-Romance, IE-Celtic, IE-Germanic and IE-Slavic languages clustering fairly close to each other.

3. Mantel and partial Mantel tests

It is important to note beforehand that the statistical significance of the following tests will not be performed using a modified/corrected α value. Corrections are usually made when multiple correlations are performed over the same data, for example, to reduce the risk of Type I error, that is, the rejection of the null hypothesis when it is true. Example of corrections to control for this problematic are the well know classical Bonferroni or the sequential Bonferroni (Rice, 1989). But has pointed by Nakagawa, using correction tests such has Bonferroni leads to an increase of Type II errors, that is, acceptance of the null hypothesis when it is false (Nakagawa, 2004). The problematic of incurring in Type I error here is avoided by the presence of a variant of the permutation test (see Materials: 2.8 Permutation test) in the function used to calculate our Mantel and partial Mantel tests.

3.1. Fast-evolving dataset

Table 11 -Mantel tests for the fast-evolving dataset. Green underline is used to indicate correlations that are significant at α = 0.05.

Test	Mantel r	*p-value*
mtDNA HVR 1 - Lang	0.3483	0.00024
mtDNA HVR 1 - Geo	0.5772	0.00001
Y-chr STRs - Lang	0.3489	0.00024
Y-chr STRs - Geo	0.4269	0.00013
Lang - Geo	0.5192	0.00001
mtDNA HVR 1 - Y-chr STRs	0.5227	0.00004

Table 12 - Partial Mantel tests for the fast-evolving dataset. Green underline is used to indicate correlations that are significant at α = 0.05.

Test	Mantel r	*p-value*
mtDNA HVR 1 - Lang / Geo	0.06971	0.26354
mtDNA HVR 1 - Geo / Lang	0.4948	0.00003
Y-chr STRs - Lang / Geo	0.1647	0.04874
Y-chr STRs - Geo / Lang	0.3068	0.00352
Lang - Geo / mtDNA HVR 1	0.4156	0.00004
Lang - Geo / Y-chr STRs	0.4369	0.00001
mtDNA HVR 1 - Y-chr STRs / Lang	0.4567	0.00040
mtDNA HVR 1 - Y-chr STRs / Geo	0.3742	0.00426

All the correlations for our Mantel test for the fast-evolving dataset (Table 11) are significant at our α = 0.05, while in the partial Mantel tests (Table 12) all except one (mtDNA HVR1 – Lang / Geo) are significant.

By analysing both Mantel tests it is possible to see that for this dataset, both mtDNA and Y-chr variation present a higher correlation with geographical distance than with linguistic distance. The results indicate that 12.13% and 33.32% of the genetic diversity observed for mtDNA can be explained by linguistic and geographical distance, respectively, with the percentage explained by geographic distance dropping slowly, to 24.48%, when controlling for the linguistic distance. For the Y-chr diversity, linguistic and geographic distances explain 12.17% and 18.22% of the diversity observed, respectively, with values dropping for 2.71% and 9.41% when controlling their effects for geographic and linguistic distance, respectively. The results also indicate a significant high correlation between both the female and male lineage (Mantel *r* = 0.5227), even when controlling for the geographic and linguistic distance, were the higher correlation occurs when controlling for the last (Mantel *r* = 0.3742 and *r* = 0.4567 and, respectively).

The results also show a high correlation between linguistic and geographic distance (Mantel *r* = 0.5192, *r* = 0.4156 and *r* = 0.4369 for Mantel test and partial Mantel tests when controlling for mtDNA and Y-chr, respectively), with geographic distance explaining 26.96% of the linguistic diversity, value that remains fairly high even when controlling for the mtDNA and Y-chr variability, 17.27% and 19.09%, respectively.

3.2. Slow-evolving dataset

Table 13 - Mantel tests for the slow-evolving dataset. Green underline is used to indicate correlations that are significant at α = 0.05.

Test	**Mantel r**	***p-value***
mtDNA SNPs - Lang	0.1553	0.16321
mtDNA SNPs - Geo	0.6719	0.00082
Y-chr SNPs - Lang	0.3483	0.0032
Y-chr SNPs - Geo	0.4366	0.00153

Test	**Mantel r**	***p-value***
Lang - Geo	0.1545	0.14811
mtDNA SNPs - Y-chr SNPs	0.1255	0.18129

Table 14 - Partial Mantel tests for the slow-evolving dataset. Green underline is used to indicate correlations that are significant at α = 0.05.

Test	**Mantel r**	**p-value**
mtDNA SNPs - Lang / Geo	0.07031	0.27374
mtDNA SNPs - Geo / Lang	0.6639	0.00122
Y-chr SNPs - Lang / Geo	0.3159	0.00954
Y-chr SNPs - Geo / Lang	0.4133	0.00141
Lang - Geo / mtDNA SNPs	0.06859	0.27409
Lang - Geo / Y-chr SNPs	0.00292	0.43327
mtDNA SNPs - Y-chr SNPs / Lang	0.07716	0.24561
mtDNA SNPs - Y-chr SNPs / Geo	-0.2519	0.98952

For the Mantel and partial Mantel tests on the slow-evolving dataset (Tables 13 and 14, respectively) not all correlations were statistically significant at our α = 0.05. For the Mantel test only for the correlation between mtDNA diversity with geographic distance and the correlations between the Y-chr diversity and both linguistic and geographic distance are significant. For the partial Mantel tests, only the correlations between the maternal lineage and geographical distance controlling for linguistic distance, and the correlations between the paternal lineages and the linguistic and geographic distances when controlling for the other non-genetic distance were statistically significant.

When analysing these results is clear that both genomic markers are highly correlated with the geographical distances, more so in the case of the maternal lineage (Mantel r = 0.6719 and r = 0.4366, for maternal and paternal lineages respectively), even when controlling for linguistic distance (Mantel r = 0.6639 and r = 0.4133), which indicates that 45.14% and 19.06% of the diversity present for maternal and paternal lineages, respectively, can be explained by geographic distance, value that keeps almost at the same level when controlling for linguistic distance (44.08% and 17.08%, respectively).

From these correlations can also be seen a statistically significant level of correlation between Y-chr diversity and linguistic diversity even when controlling for geographic distance (Mantel *r* = 0.3489 and *r* = 0.3159) which results in 12.13% and 9.98% of the Y-chr diversity being able to be explained by linguistic distance, even when controlling for geographic distance.

4. Scatterplots

4.1. Fast-evolving dataset

4.1.1. mtDNA HVR 1 vs Ling

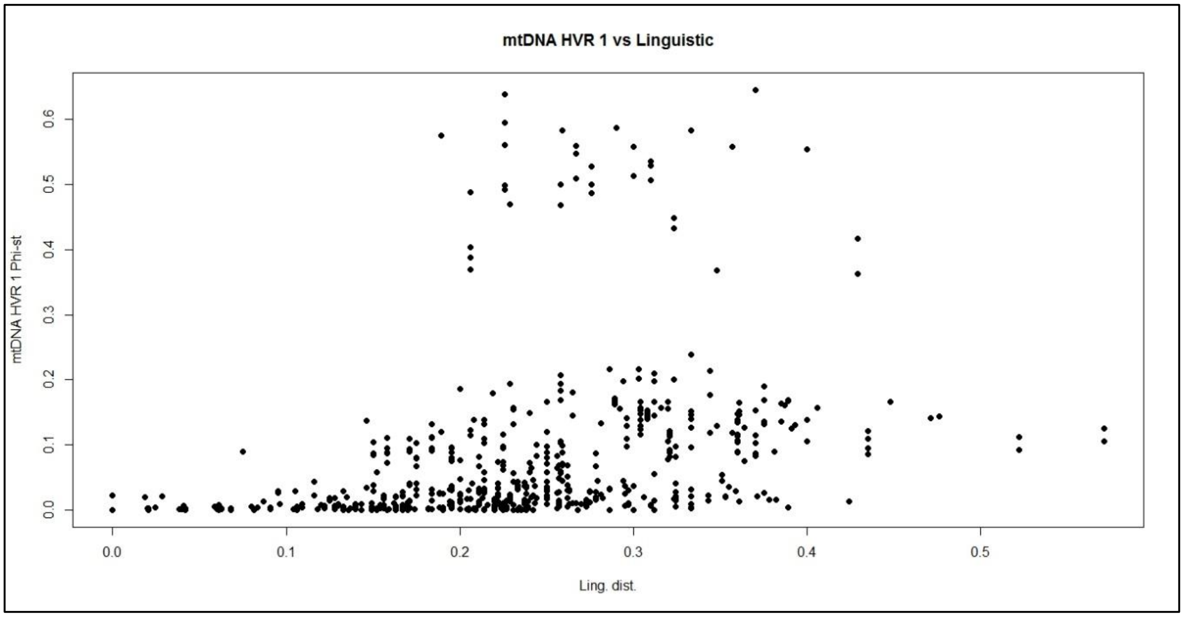

Figure 12 - Scatterplot of mtDNA HVR 1 distances versus Linguistic distances

Analysing Figure 12 scatterplot of mtDNA HVR 1 distances versus Linguistic distances a correlation between both traits is visible, with the degree of differentiation between populations' increasing slowly with the increase of linguistic distance. The strength of this correlation is not constant though, and this is seen in the fact that the level of variability in the genetic distance is ever higher with the increase of geographical distance. On this scatterplot it is also visible the existence of a low level of genomic differentiation between populations, even when looking for populations with higher linguistic distance, which indicates a high degree of homogeneity in the maternal lineage across vast linguistic distances. It is notable that the higher degree of genetic differentiation between pairs of populations occurs at mid-range linguistic distance.

4.1.2. mtDNA HVR 1 vs Geo

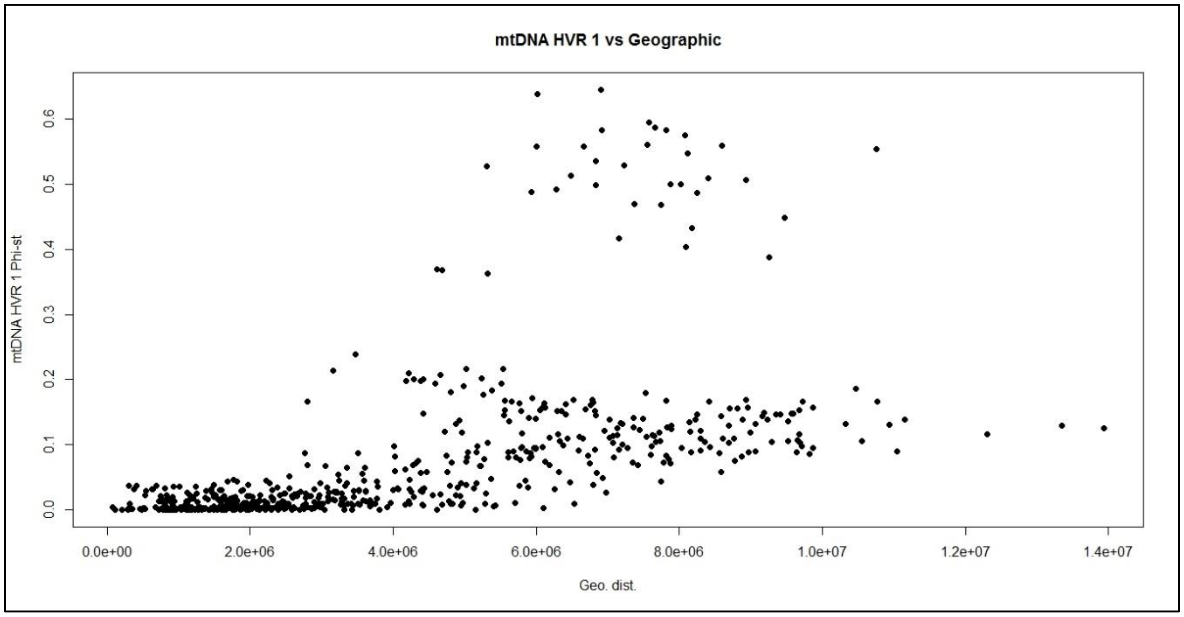

Figure 13 - Scatterplot of mtDNA HVR 1 distances versus Geographic distances

In Figure 13 scatterplot of the mtDNA HVR 1 distances versus the Geographic distances it is visible that a correlation between both traits exists with the genomic distance between populations increasing as the distance between each population increases, with this correlation being stronger for geographically proximate populations than for increasingly distant ones. It is also possible to verify that a low level of genetic differentiation exists along the geographic trait, with the level of genomic homogeneity being higher for geographical closer populations, and the higher level of genomic diversity being higher for geographical distant populations.

4.1.3. Y-chr STRs vs Ling

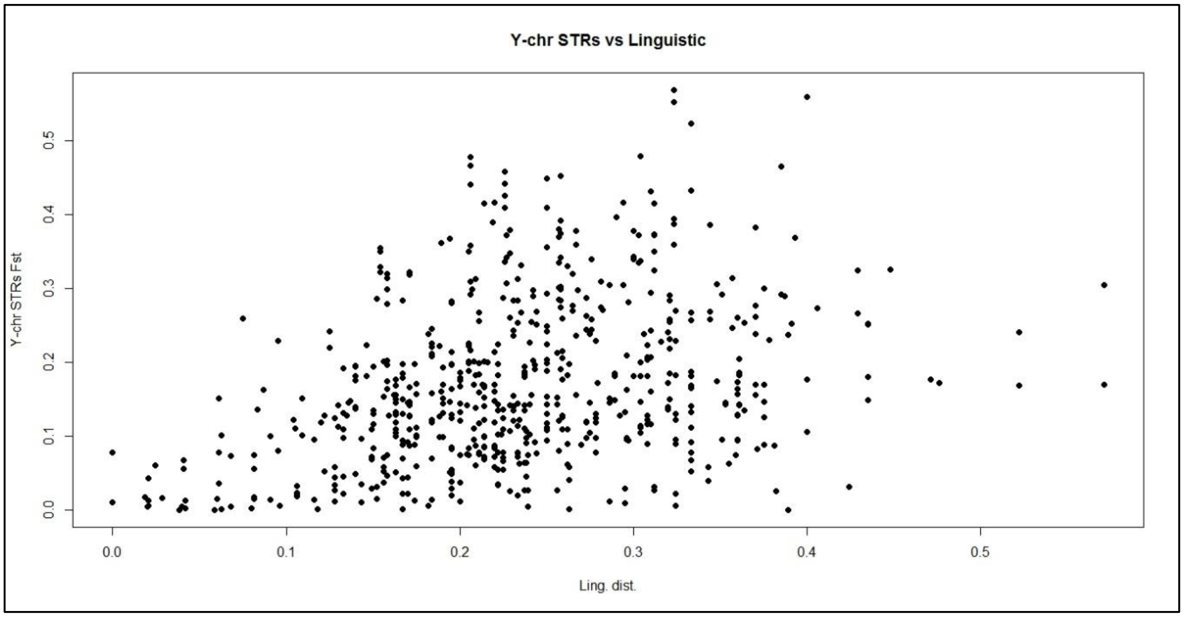

Figure 14 - Scatterplot of Y-chr STRs distances versus Linguistic distances

By analysing the scatterplot of the Y-chr STRs distances versus the Linguistic counterpart (Figure 14) it is possible to see that a correlation between the two variables exists but that this is very small, with the degree of variability between genetic differentiation of populations varying widely has more linguistically distant the populations are from each other, which indicates a decrease of correlation between the two variables with the increase of linguistic distance.

4.1.4. Y-chr STRs vs Geo

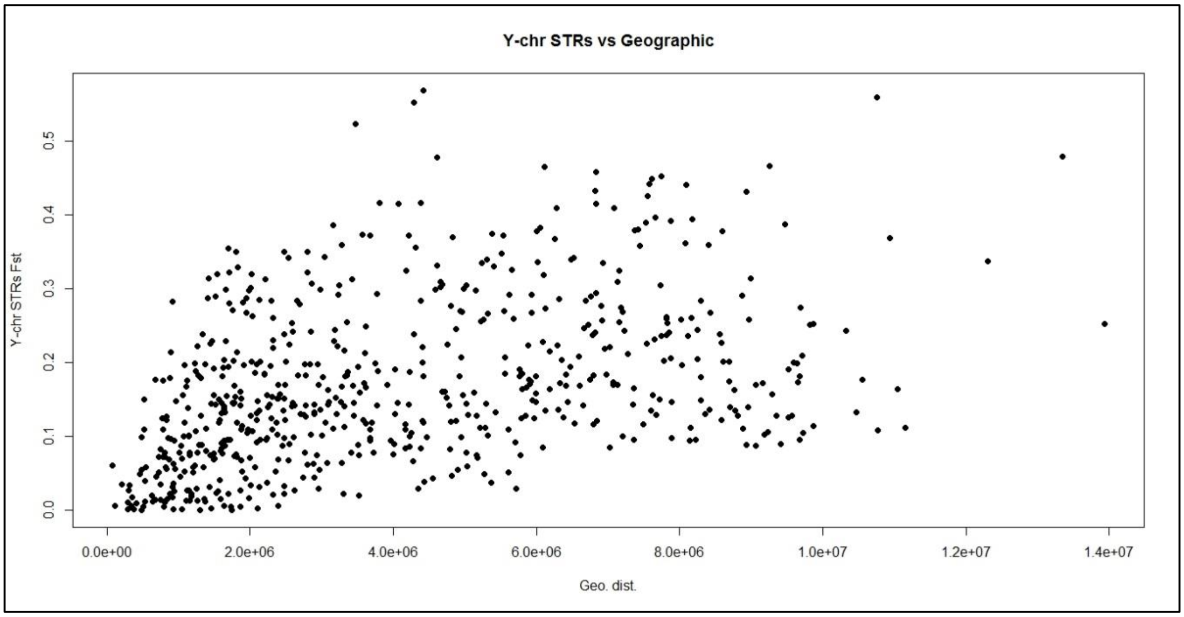

Figure 15 - Scatterplot of Y-chr STRs distances versus Geographic distances

By analysing the scatterplot of the Y-chr STRs distances versus the Geographic counterpart it is visible that a small correlation exists between the genetic and geographical traits. This can be seen by the fact that even when there is an increase in genomic differentiation along of the geographical axis, this same increase is accompanied by an increase of the variability in the genomic differentiation, and that this progressive increase in genomic differentiation between populations along the geographical axis increases ever less.

4.1.5. Ling vs Geo

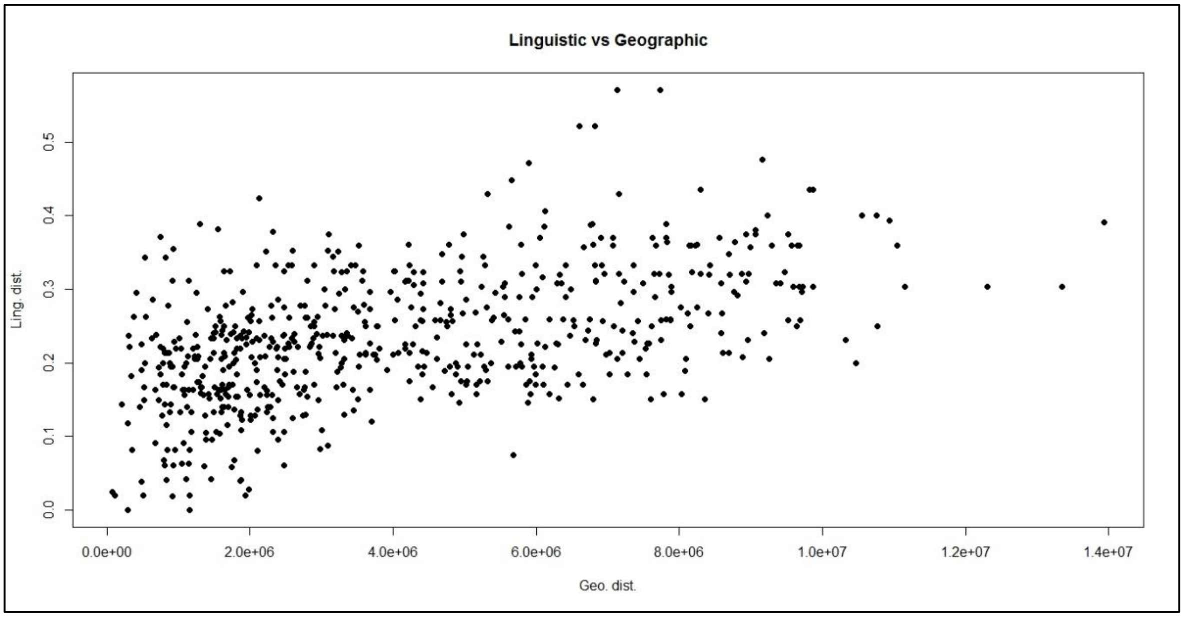

Figure 16 - Scatterplot of Linguistic distances versus Geographic distances

In the scatterplot of Linguistic distances versus Geographic distances (Figure 16) is possible to observe that a correlation between both variables exists and is kept almost constant, that is, the linguistic distance between population pairs increases as the geographical distance also increases, but the distance variability for the linguistic trait for any given geographical distance is kept almost constant as the geographical distance between populations increases, which is indicative of a constant level of correlation between both variables.

4.1.6. mtDNA HVR 1 vs Y-chr STRs

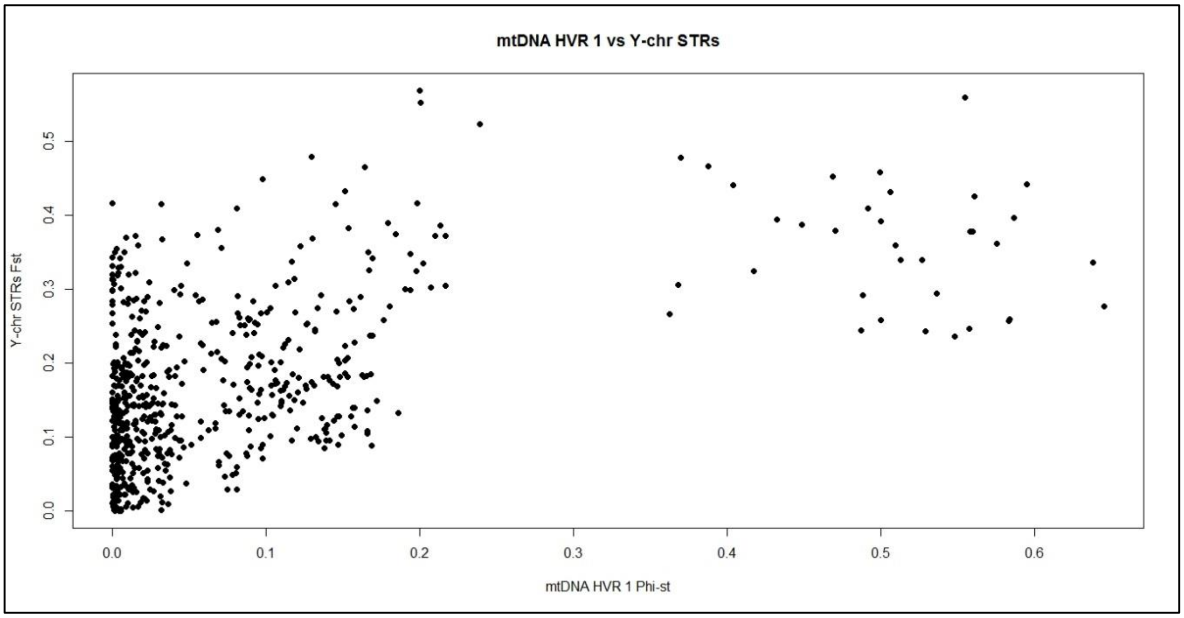

Figure 17 - Scatterplot of mtDNA HVR 1 distances versus Y-chr STRs distances

When looking for the mtDNA HVR 1 distances versus Y-chr STRs distances scatterplot (Figure 17) is clearly visible that a contrast exists between the maternal and paternal lineages, with the maternal lineages presenting low levels of genetic differentiation between populations, with a great number of population pairs presenting near genetic homogeneity between them, while for the paternal lineage the levels of genetic differentiation between populations tends to vary greatly for any given maternal distance. Analysing the scatterplot is also visible that some correlation between them exists, with both lineages genomic differentiation level increasing as the same occurs in the other lineage. It is noteworthy that for a group of population pairs with higher genomic differentiation on the paternal lineage, an also high degree of genomic differentiation exists on the maternal lineage, in contrast with the more general low level of differentiation patent in this lineage.

4.2. Slow-evolving dataset

4.2.1. mtDNA SNPs vs Ling

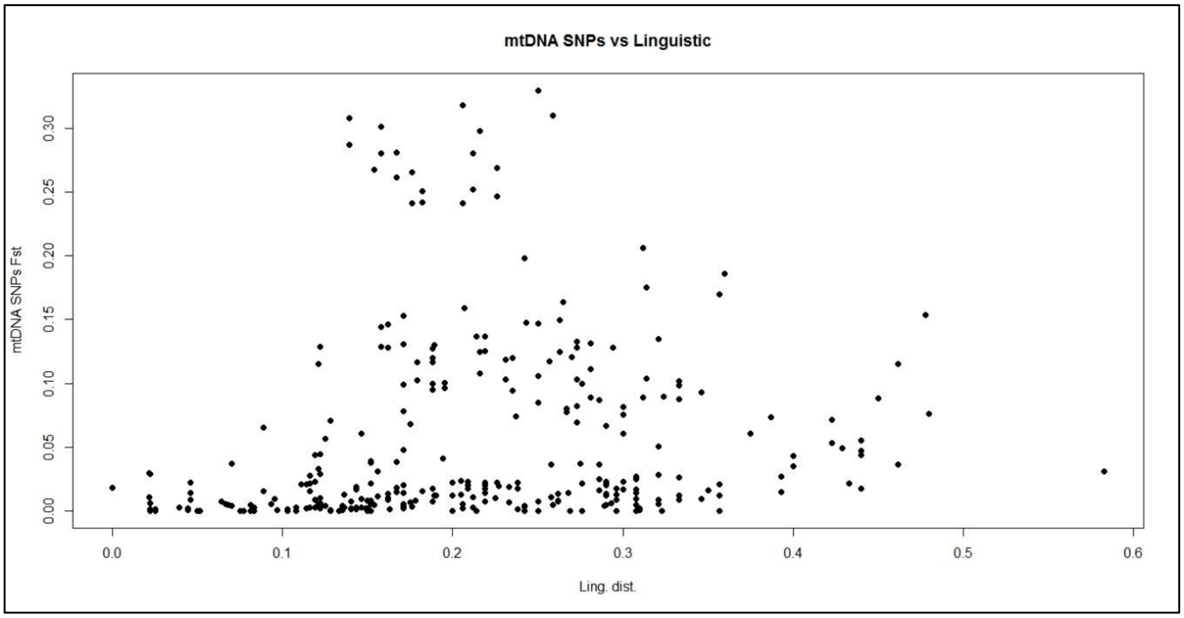

Figure 18 – Scatterplot mtDNA SNPs distances versus Linguistic distances

In Figure 18 scatterplot of the mtDNA SNPs and Linguistic distances it is possible to see a very low, to almost null, correlation between both variables exists, this can be seen by the fact that the increase in the linguistic distance between populations, a reciprocal tendency not being seen in the genomic variable, that is, even though there are population pairs presenting higher genomic differentiation for higher linguistic distances, this is not observed as a general pattern, with the media of the population pairs keep presenting low levels of genomic differentiation. In the scatterplot is also evident a low level of genomic diversity for the maternal lineage across the linguistic variable, that is, even though the variability of genomic differentiation increases, its media tends to remain mainly low even for ever linguistically distance population pairs.

4.2.2. mtDNA SNPs vs Geo

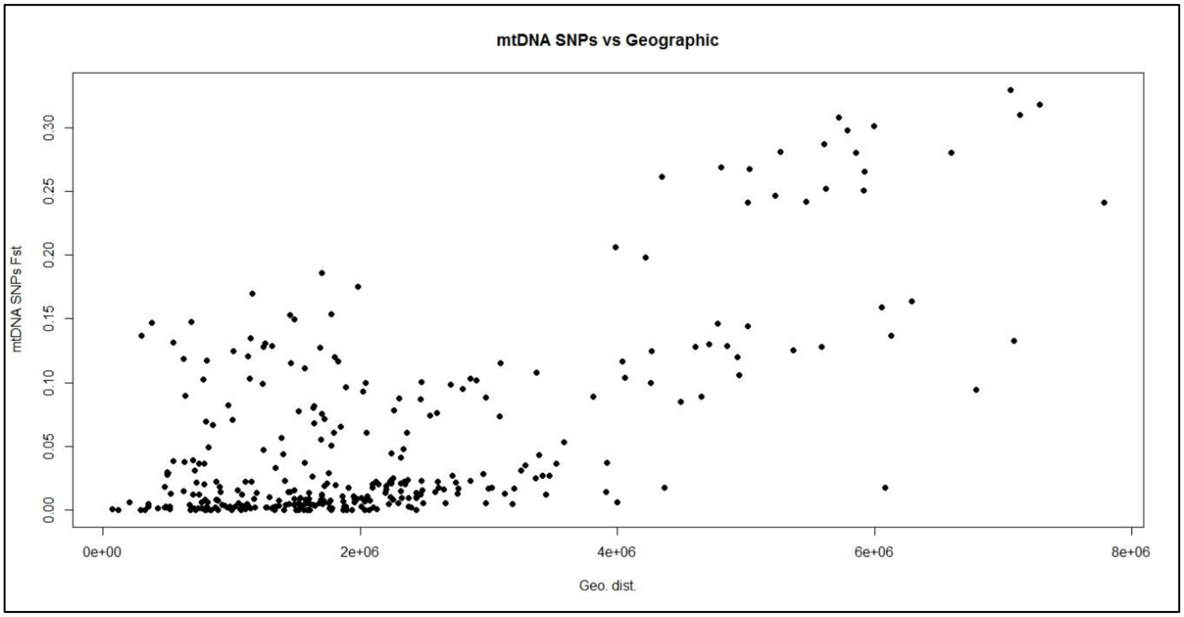

Figure 19 – Scatterplot of mtDNA SNPs distances versus Geographic distances

By plotting the mtDNA SNPs distance versus the Geographic distance in a scatterplot (Figure 19) it is possible to see a low level of genetic differentiation for geographically close population pairs, even though some degree of variability in the genomic differentiation exists. With the increase of the geographical distance between pairs a higher genomic differentiation is also visible, more for further geographically distant population pairs, and very small for close ones, which is indicative that some degree of correlation exists between both variables.

4.2.3. Y-chr SNPs vs Ling

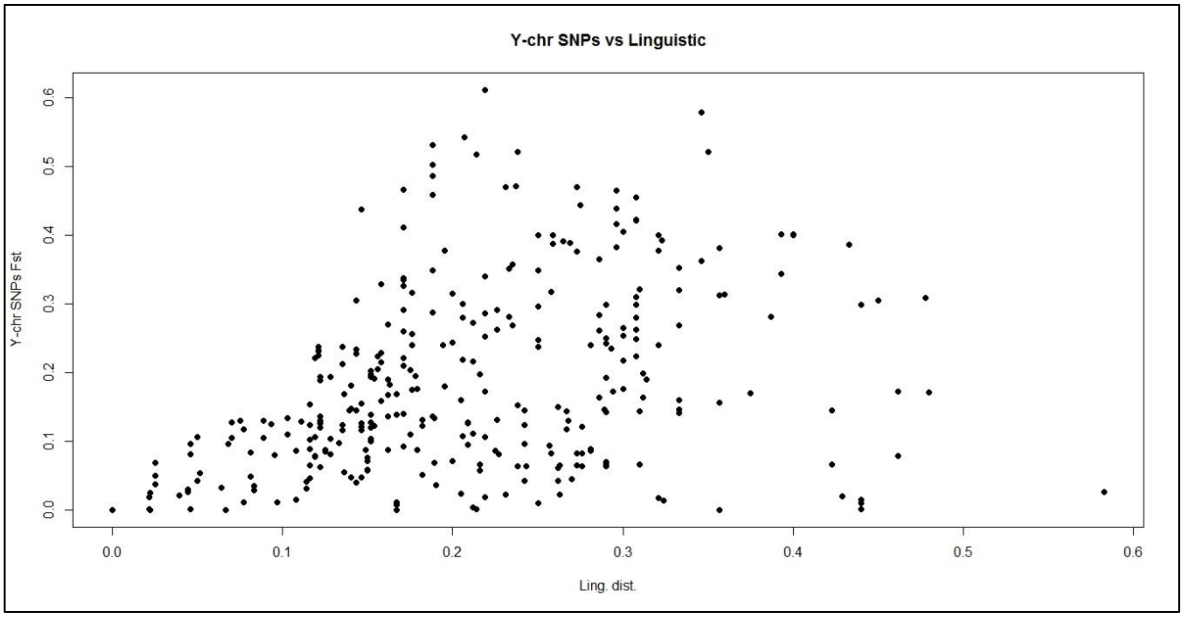

Figure 20 – Scatterplot of Y-chr SNPs distances versus Linguistic distances

By analysing the scatterplot of the Y-chr SNPs distance vs Linguistic distance (Figure 20) it is possible to see that some correlation exists between the two variables but that it is not strong and loses power as more linguistically distance the population pairs are, this can be seen on the increasing in the variability of the genomic differentiation levels along the linguistic axis.

4.2.4. Y-chr SNPs vs Geo

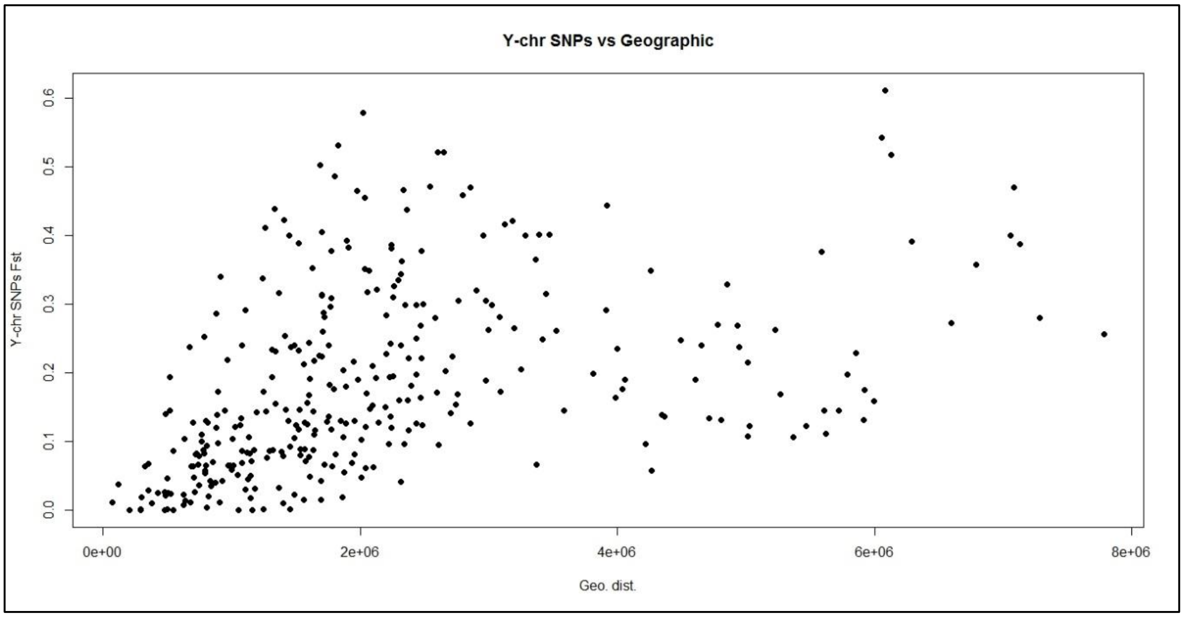

Figure 21 – Scatterplot of Y-chr SNPs distances versus Geographic distances

When comparing the Y-chr SNPs distances with the Geographic distances (Figure 21) it is possible to see that a correlation between both variables exists, but decreases in power for ever further apart populations, until there is not a visible correlation between both variables (from mid-range distance onwards), which can be seen with the increasing in variability in the genomic differentiation levels up to mid-range of the geographical axis, point which onwards the genomic differentiation levels between populations tends to vary within the same *Fst* values.

4.2.5. Ling vs Geo

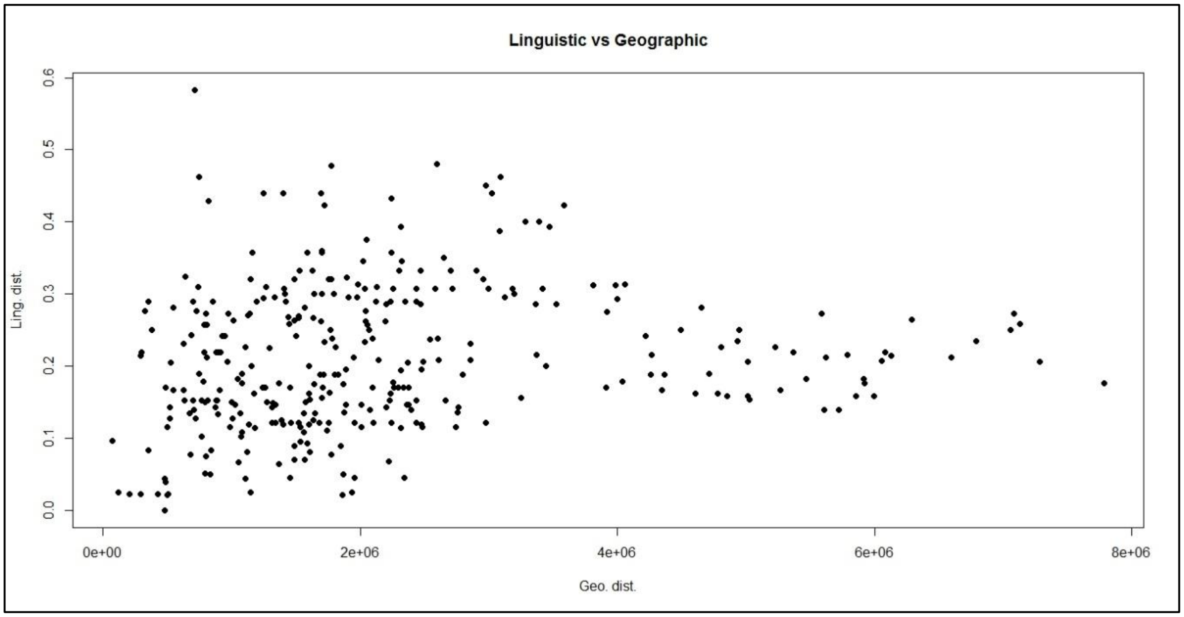

Figure 22 – Scatterplot of Linguistic distances versus Geographic distances

In the scatterplot of Figure 22 it is possible to observe that some correlation exists between both variables in study, but this is not very powerful and is only kept until mid-range geographical distances, and being null onward, with the linguistic distances between pairs keeping at the same range while the geographical distance increases.

4.2.6. mtDNA SNPs vs Y-chr SNPs

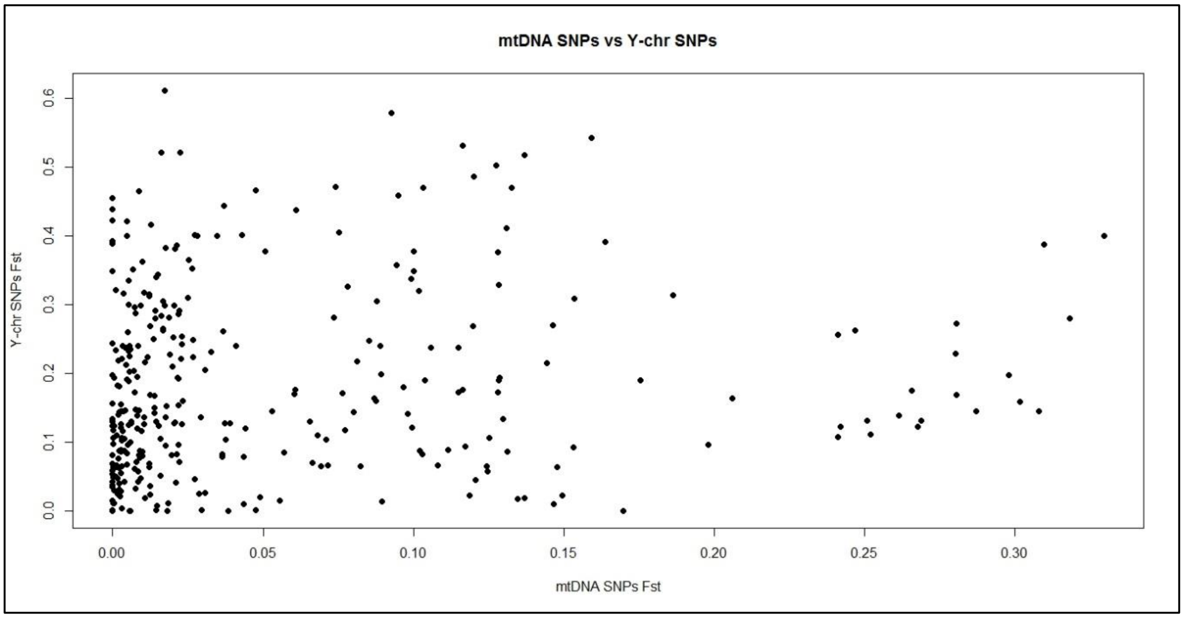

Figure 23 – Scatterplot of mtDNA SNPs distances versus Y-chr SNPs distances

In Figure 23 scatterplot between the mtDNA SNPs distances and Y-chr SNPs distances it is possible to observe that overall the maternal lineage presents a smaller range in genetic differentiation between populations when compared with the paternal lineage, seen that the maximum of genomic differentiation in the maternal lineage is given by *Fst* values lower than 0.4, while for the paternal lineage some population pairs present *Fst* values around 0.6. In this scatterplot is also possible to see that almost no correlation exists between both variables, with the levels of genomic differentiation between population pairs varies greatly between both lineages for any selected distance on the counterpart.

Discussion and conclusion

When working with genetic markers to infer demographic patterns some aspects must be taken in account, such as which kind of markers to use, if autosomal or uniparental markers. Due to recombination, the demographic history on autosomal markers doesn't allow to detect differences in the demographic history of both sexes over time, which in contrast can be detected using uniparental markers. But these are not 100% bullet proof either, with results being potentially skewed due to differences in effective population size in each sex or different mutation rates on the chosen markers. Different effective population sizes can hinder the results, for in a small population size a demographic event, such as population decrease, can lead to a higher mark in the demographic history than the same event in a population with higher effective size. The same goes for different mutation rates, where using markers with a high-mutation rate for one parent and a slow-mutation rate for the other may skew the results, for the observed demographic patterns may be marks of different demographic events that occurred at different periods in time.

Taking this in account and looking at our genomic databases, even though, for example, we possess information on either parental line on the fast-evolving marker, the comparison between each uniparental marker data has its limitations, such as the comparison of the divergence time between populations, due to our markers for each sex (even though being both considered fast-evolving markers) presenting different mutation rates. This can be specially seen within the Y-chr STRs database), this due to the different authors using different STR markers within their own work and when compared with other authors for that dataset, and as shown in Gusmão et al. (2005), different STR markers will present different mutation rates between each other. In the end, comparison can be made between the patterns observed in both parental lines within each rate-mutation dataset, but always taking in consideration that the events may not have occurred at the exact same time in each parental line.

Another consideration to make on our dataset regards the mtDNA HVR 1 dataset. Even though a large dataset can be collected for this particular marker, as stated in the Methodology, some problems arise when using this marker due to not just different methods used for sequencing it, but also the length of the sequence used by each author, which varied widely in some cases. And even if we tried to correct some sequencing errors and tried to keep the maximum length of this genomic region without hindering the results, that fact is that the use of such a small portion of the mitochondrial genome may not reveal all the history of that parental lineage.

When comparing the structure, demographic history and thus, the correlation between different variables, such as genomic and linguistic, assumptions must be made carefully, for even when a correlation may exist between two variables that does not imply a direct cause. And even when it is identified the existence of a possible cause between two variables, caution must be taken for multiple variables may be causing influence that can be masked, for example, a group of populations diverging genetically separated by mountains from each other and each possessing divergent languages, on this case we may obtain a correlation between the genomic and the linguistic data, but assuming that the linguistic component is the main cause or even the sole cause of the genomic differentiation would be an error, and so the geographical property must also be taking in consideration. In this work, as in many others, there is a limit to the numbers of variables that can be studied, and so approaches to overcome this must be taken when analysing the data, for instance, the use of partial Mantel tests, so when analysing the correlation between two given variables we can take in account the effect of a third variable. In this study, for example, when analysing the correlation between the genomic and linguistic variables we also analysed this correlation taking in account the effects that may be due to the geographic component. And even then, caution must be taken for other variables not taken in account, such has other cultural traits diverging between populations may be playing a role too.

1. Correlation between linguistic and geographic data

In our Multidimensional Scaling plots a clear clustering of each language with their close linguistic relatives was obtained agreeing with what is expected for, showing that languages from a same linguistic family tend to keep closer to each other than to their geographical neighbours, even when they belong to a different linguistic family or sub-families. Even more, this shows the power of the syntax methodology used in this project to classify languages, by showing evidence of a resistance of syntax to borrowing from neighbour languages.

Even though no correlation with geographic distances can be observed in the MDS computations, this can be observed in the results obtained with the Mantel tests on the fast-evolving dataset. In this analysis a significant correlation was obtained between the linguistic and geographical variables, even when controlling for the effects that may be due to the genomic variable, with the geographical variable being able to explain up to 26.92% of the linguistic diversity. The percentage of linguistic variability, that can be explained by the geographic distance, lessens when controlling for the genomic variables, to values of 17.27% and 19.09%, for mtDNA HVR 1 and Y-chr STRs respectively, which indicates already that some part of the linguistic diversity may be due in part to the genomic diversity.

This correlation between the geographical and linguistic variables can be further noticed in the scatterplots between both data where a small correlation is visible between them, more so in the fast-evolving data where the correlation strength is kept almost constant along the distance, but it decreases in the slow-evolving dataset after mid-range geographical distance between populations.

The differences observed between the correlation of geographic and linguistic data for both datasets may be due to the use of a different set of populations, that is, has indicated in the methodology for the genomic data-mining (Methodology: 1.3 – Genetic data-mining) and shown in Table 4, both the fast- and slow-evolving datasets do not present the complete list of populations used in this study, rather, each possess a subset of that total, and so, possible discrepancies on the history of each language may have led to this differences in our results.

2. Correlation between genomic ad geographic data

When observing the results of the MDS computations for both markers in either the paternal and maternal datasets a clear underlying geographical structuring of the populations is visible, with populations tending to be more close to their geographical neighbours than to their linguistic ones. To note, that this geographical structuring observed is different in the different datasets. For example in the mtDNA HVR 1 dataset a clear east-west orientation is visible with all populations making a tight clusters apart from the Inuits, the clear outlier in this dataset, and even within this cluster this orientation is visible in part, where only a separation between Far East and Middle East and European populations is made, with the populations of this last two groups forming a big cluster with no clear geographical orientation. When looking for the mtDNA SNPs dataset there's a west-east orientation of the populations with the majority of the European populations clustering together apart from the populations of the Middle East, with the notable exception of the Arabic populations that cluster together with the European cluster, and the Icelandic and Hungarian populations that are positioned outside of it. The Inuits divergence from all the other populations may be due to their demographic history and long divergence form the rest of the populations, including the Far East populations in our study.

On the Y-chr STRs dataset, no particular geographical orientation is visible, even though a geographic structuring underlying the populations is visible by the clustering of populations with their geographical neighbours rather than their linguistic neighbours, with some notable exceptions, such has the position of the Japanese (in the middle of European populations), and the position of the Central Basque (closer to the Celtic languages rather than the Spanish and French populations). The Celtic populations also show a demarcation from their geographical neighbour, the English populations. In contrast with the pattern observed for the Y-chr STRs dataset, the SNPs dataset of the paternal lineage show clear west-east orientation of the populations with only some populations being left out of this orientation, such has the case of the Finnish that appear here as outliers, the close proximity of the Indo-Iranian populations to the European ones, the demarcation of the Basque languages from their geographical neighbours and closeness to the Irish populations, which may be caused by clear cut in mating between the Basques and the French and Spanish, or be also indicative of a demographic event between the Irish and Basque leading to their closeness. In this analysis is also visible that the geographical discrimination in both parental markers is made differently, on the maternal lineage the analysis manages to resolve a differentiation between populations belonging to geographically distant population, clustering them in west-east clusters, and it fails to discriminate any geographical structure within these geographical close clusters. In the other side, the MDS analysis for the paternal lineage were able to obtain a clear structuring for close proximity populations in both markers and give a west-east geographical orientation for the SNPs dataset. It is important though to note that the MDS results on the paternal lineage, even if presenting good *stress* values, these are very high when compared to those obtained for the maternal lineage, which is indicative that some of the underlying geographical structure occurring in the paternal lineage may not be possible to observe here due to distortion of the real data.

Looking to the Mantel tests for both genomic datasets we again obtain results indicating that a correlation between the genomic and geographic variables, which can be seen by the significant correlations obtained by both parental data in both markers even when controlling by the effects of linguistic distance. And once again we can see a higher correlation between the maternal lineage and geographic distance than the paternal lineage, with the geographic variable explaining 33.32% and 45.14% of the mtDNA HVR 1 and mtDNA SNPs, values that keep significant and close to this when controlling for the linguistic effect (24.48% and 44.08%, respectively). This correlation is noticeably low for the paternal lineage, where only 18.22% and 19.06% of its diversity for the STRs and SNPs markers can be explained by the geographical variable, values that even though significant drop when controlling the linguistic effect, more so for the STRs marker, with it being just explained by 9.41% (for contrast the value drops little, to 17.08%, for the SNPs marker when controlling for linguistic effect). These Mantel results show that the maternal lineage presents a higher correlation with the geographic distance in comparison with the paternal lineage, with the results being even clearer when taking in account the effects of linguistic distance which reduce even more the correlation between the paternal data.

These previous results are also visible and supported by the Scatterplot results, with a visibly higher correlation existing between the maternal lineages and geographical distance, especially on the HVR1 dataset. In this analysis a low values of genomic differentiation between populations along the geographical axis indicates a low level of genomic diversity between populations even at great geographical distances, values, that when compared to the paternal results, where an increasingly higher genomic differentiation values were obtained along the geographical axis is indicative of a higher mobility of females across populations.

By looking at both markers and comparing each results for their different markers, some divergences are observed. These can be caused either by the population's samples for each marker presenting different demographic histories or that at different periods of the human history the geographical barrier had different levels of importance in either parental marker. If we ignore the effect that may be caused by the different subset of populations used for each time-related marker, a clear difference is visible for both time scales with both parental lineages, being less affected by the geographical barrier in more recent periods than it was further in the past, more so in the maternal lineage were a lower genomic diversity is observed for the slow-evolving dataset, which may have been caused by a higher mobility of the females further in the past.

3. Correlation between genomic and linguistic data

When looking to the MDS results of the genomic datasets the only computation where a small underlying linguistic structure is perceived is in the mtDNA SNPs dataset were the Indo-Europeans, with some distortion, appear to cluster almost all with their linguistic sub-families.

When looking instead to the Mantel correlations, mainly for the fast-evolving dataset where both parental markers present a significant correlation (whereas in the slow-evolving dataset only the paternal data shows a significant correlation), it is possible to see that a correlation between the linguistic and genomic data exists for the populations in study, even when controlling for the effects of geographic distance with both parental markers scoring an almost equal correlation value, indicating that approximately 12% of their variability can be explained by the linguistic variable, which would point for both parental lineages being affected by the linguistic variable in the same way, thing which can't be confirmed for only the correlation with the paternal marker when controlling for the geographical distance is significant, with the result being also at the 2.71% of genomic variability being able to be explained by the linguistic barrier, indicating that at least for the paternal lineage the effect observed in the correlation between the genomic and linguistic data is affected by the geographic variable.

For slow-evolving markers the only parental lineage with statistically significant correlations was the paternal lineage, even when controlling for the geographical distance. The results point for 12.13% of the variability observed for the paternal SNPs dataset being explained by linguistic diversity.

When looking for the correlation between the paternal markers and the linguistic diversity of populations and controlling for the effect of the geographic barrier there is a higher drop in the percentage that the linguistic diversity can explain in the STRs dataset which isn't so accentuated in the SNPs data, with the values dropping to 2.71% and 9.98%. These results are indicative the effect of the linguistic variable in the paternal demographic history was more important in distant periods of human history.

This higher correlation between the paternal and linguistic distances when comparing to the maternal lineage can be further seen in the Scatterplot analysis with a higher correlation of the linguistic data with the paternal lineage, even though not a strong one. In this analysis, and supporting the closer resemblance between the patterns of population differentiation, a resemblance between the scatterplots of the Y-chr and linguistic distances with the geographical distances can be observed. This same resemblance does not occur for the maternal markers.

Overall, the analysis above show that some degree of correlation between both genomic data and the linguistic data exists, but this is perceptually higher for the paternal markers whose geographical genomic distribution resembles the geographical distribution of the linguistic distances.

4. Correlation between parental markers

Taking in account the results explained above, some degree of the difference between both parental datasets is expected. This can be perfectly seen in the different MDS results obtained for both datasets on both markers, more so for the fast-evolving data, even when taking in account the different levels of distortion presented by the different parental lineages.

When looking for the Mantel test only for the fast-evolving dataset a statistically significant correlation between both parental markers was obtained with each diversities explaining up to 27.32% of each other's diversity, value that drops to 20.86% and 14 when controlling for the linguistic and geographic diversity.

This small correlation between both parental datasets can be further observed on their scatterplots with both data showing a low correlation in the fast-evolving dataset and a barely null one on the slow-evolving dataset. On both datasets is also visible a clear higher overall diversity on the paternal lineage, with this being clearer on the slow-evolving dataset.

Overall these results indicate a small correlation between both parental lineages and an overall higher diversity on the paternal side, which indicates a higher dispersal rate of the females comparing to the males.

When taking in account the totality of the results it is possible to see that both language and geography played a role in human demographic history working as barriers, and that both acted differently for both parental lineages and while both the parental lineages were more affected by the geographic constraint, the maternal lineage was the one that presented a higher correlation with this variable; while the paternal lineage presented the higher correlation with the linguistic trait at both time scales studied. And even if we care to take a leap and disregard the possibility of an effect on the results of the different time-related datasets due to differences in the populations used for each, we can see that different demographic patterns occurred along the human history, with both geographic and linguistic variables playing rules with different importance, with their importance being higher in the distant past than in more recent times.

With these results we are able to answer the questions proposed for this project:

- There is a correlation between both the genomic (both in the maternal and paternal lineages) and linguistic diversity with the geographic distance. That is, the genomic distance increases with the increase of geographical distance, a same effect visible for languages that get increasingly distant from each other as they get geographically distant. And this effect can be observed on the genomic data in different time periods of human history.

- There is a correlation between the genomic and linguistic diversity for both parental markers, with linguistically closer populations being genomically closer. This evidence supports the hypothesis that language also works as a barrier on the demographic history of humanity. An effect that is more visible in the paternal lineage at both time scales.

- There is a difference in the demographic history of both parental lineages with the female lineage showing a lower genomic variability across the linguistic and geographic variables, indicating a higher rate of mobility for females than for males. And has for the other question, here also this is evident at different time scales.

A key point of the results obtained in this work is that they go in agreement with previous results obtained by others.

Since the Sokal et al. (1988) and Cavalli-Sforza et al. (1988) papers, further studies revealed that some level of correlation between the linguistic differences of populations and their genomic differentiation exists. And these results are not just confined to the European populations but show the same effects of linguistic diversity on genomic diversity across populations in different points of the globe, remarking that this is not a local effect that just happens in Europe, has is seen in this work, were the effects of language has a barrier is spread throughout the Eurasian linguistic diversity.

For instance, when looking to the recent work of Creanza et al. (2015) both results show a clear correlation between the genomic and linguistic data with the geographic data, indicating that geographic distance played a role in both the genomic and linguistic structuring of populations. Our results also go in agreement with the results obtained by Creanza et al. (2015) for the correlation between genomic and linguistic data, with linguistically closer population presenting a smaller degree of differentiation between them. The differences seen between both studies on the degree of correlation observed can be explained by the differences on the exact data used by each, for instance, whereas our linguistic data is based on syntactic characteristics of languages the Creanza et al. (2015) linguistic data is based on phonemes, that is, a lexical characteristic. The differences seen between both works may also be caused by the different genomic data used. And if doubts may arise on the reason for these differences, when comparing our results to those obtained by Colonna et al. (2010), were both linguistic data is based in syntactic characteristics of languages (and both works using the PCM methodology) and based on STR and SNPs dataset, it is clearly possible to see that our results keep being in accordance with previous works.

Furthermore, the inference made from our results that there is a higher correlation between the genomic structure of the female lineage and the geographical diversity of human populations is not a result unique of our study, already other studies using uniparental markers to study differences in human demographic history that might be observed between populations came to the same conclusion. For instance, both the works of Seielstad et al. (1998) and Pérez-Lezaum et al. (1999), when comparing the genomic diversity on STRs with that on the mtDNA found that the paternal lineage had a higher diversity across populations, with both works pointing to the patrilocal characteristics of our species, were the bride tends to move to the groom's residence.

5. Future

As obvious, a survey of these datasets using the complete kit of phylogenetic methods was not performed, due to either time constrains or even constrains caused by the datasets themselves as pointed above in this discussion. But still, some more analyses can be performed for these data which may help enhance the resolution of this results and help a better understanding of the dynamics in play in both genomic and linguistic evolution of human populations.

For instances, the differential colouring of the Scatterplot analysis by linguistic populations and its analysis may result in a better understanding in the dynamics that might be playing within different linguistic families and how different linguistic families behave in relation to either their genomic, geographical or linguistic closer populations.

A study on both the genomic and linguistic boundaries might also help to verify if they overlap and thus confirm the results already obtained by Barbujani and Sokal (1990). More importantly, with the increase of studies available using the genomic markers chosen for this study, our datasets can be enhance in more information, and make possible to have the same populations for both fast- and slow-evolving mutations, thus allowing us to compare the results obtained at different time scales without the constrain of these being caused by the use of different population sets.

And as the number of languages sampled for their syntactic characteristics using the PCM method increases, also a higher number of populations can be added to the study, ultimately covering the complete set of populations of the Eurasian linguistic diversity in order to study if the results here obtained of a correlation between both linguistic and genomic data with each other and the geographic data, and of a higher maternal geographical structure can be observed for the complete set of Eurasian populations.

Bibliography

Achilli, A., Olivieri, A., Pala, M., Metspalu, E., Fornarino, S., Battaglia, V., . . . Torroni, A. (2007). Mitochondrial DNA variation of modern Tuscans supportsthe Near Eastern origin of Etruscans. *American Journal of Human Genetics*, 759-768.

Allentoft, M., Sikora, M., Sjögren, K., Rasmussen, S., Rasmussen, M., Stenderup, J., . . . Willerslev, E. (2015). Population genomics of Bronze Age Eurasia. *Nature*, 167-172.

Altschul, S., Gish, W., Miller, W., Myers, E., & D.J., L. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 403-410.

Al-Zahery, N., Pala, M., Battaglia, V., Grugni, V., Hamod, M., Kashani, B., . . . Semino, O. (2011). 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*.

Anderson, S., Bankier, A., Barrell, B., Bruijn, M., Coulson, A., Drouin, J., . . . Young, I. (1981). Sequence and organization of the human mitochondrial genome. *Nature*, 457-465.

Andrews, R., Kubacka, I., Chinnery, P., Lightowlers, R., Turnbull, D., & Howell, N. (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature*, 147.

Armitage, S., Jasim, S., Marks, A., Parker, A., Usik, V., & Uerpmann, H. (2011). The southern route "Out of Africa": evidence for an early expansion of the modern humans into Arabia. *Science*, 453-456.

Atkinson, Q. (2013). The descent of words. *Proceedings of the National Academy of Sciences of the U.S.A.*, 4159-4160.

Atkinson, Q., & Gray, R. (2005). Curious parallels and curious connections - Phylogenetic thinking in biology and historical linguistics. *Systematic Biology*, 513-526.

Babalini, C., Martínez-Labarga, C., Tolk, H.-V., Kivisild, T., Giampaolo, R., Tarsi, T., . . . Rickards, O. (2005). The population history of the Croatian linguistic minority of Molise (shouthern Italy): a maternal view. *European Journal of Human Genetics*, 902-912.

Badro, D., Douaihy, B., Haber, M., Youhanna, S., Salloum, A., Ghassibe-Sabbagh, M., . . . Consortium, T. G. (2013). Y-chromosome and mtDNA genetics reveal significant contrasts in affinities of modern Middle Eastern populations with European and African populations. *PLoS One*, e54616.

Balter, M., & Gibbons, A. (2015). Indo-European languages tied to herders. *Science*, 814-815.

Bandelt, H., & Kivisild, T. (2006). Quality assessment of DNA sequence data: autopsy of a mis-sequenced mtDNA population sample. *Annals of Human Genetics*, 314-326.

Bandelt, H., Lahermo, P., Richards, M., & Macaulay, V. (2001). Detecting errors in mtDNA data by phylogenetic analysis. *International Journal of Legal Medicine*, 64-69.

Bandelt, H., Quintana-Murci, L., Salas, A., & Macaulay, V. (2002). The fingerprint of phantom mutations in mitochondrial DNA data. *American Journal of Human Genetics*, 1150-1160.

Barbujani, G. (1997). DNA variation and language affinities. *American Journal of Human Genetics*, 1011-1014.

Barbujani, G. (2013). Genetic evidence for prehistoric demographic changes in Europe. *Human Heredity*, 133-141.

Barbujani, G., & Bertorelle, G. (2001). Genetics and the population history of Europe. *Proceedings of the National Academy of Sciences of the U.S.A.*, 22-25.

Barbujani, G., & Colonna, V. (2010). Human genome diversity: frequently asked questions. *Trends in Genetics*, 285-295.

Barbujani, G., & Pilastro, A. (1993). Genetic evidence on origin and dispersal of human populations speaking languages of the Nostratic macrofamily. *Proceedings of the National Academy of the U.S.A.*, 4670-4673.

Barbujani, G., & Sokal, R. (1990). Zones of sharp genetic change in Europe are also linguistic boundaries. *Proceedings of the National Academy of Sciences of the U.S.A.*, 1816-1819.

Barbujani, G., & Sokal, R. (1990). Zones of sharp genetic change in Europe are also linguistic boundaries. *Proceedings of the National Academy of Sciences of the U.S.A.*, 1816-1819.

Barnabas, S., Shouche, Y., & Suresh, C. (2006). High-resolution mtDNA studies of the Indian population: implications for the palaeolithic settlement of the Indian subcontinent. *Annals of Human Genetics*, 42-58.

Barreiro, L., Laval, G., Quach, H., Patin, E., & Quintana-Murci, L. (2008). Natural selection has driven population differentiation in modern humans. *Nature Genetics*, 340-345.

Behar, D., Metspalu, E., Kivisild, T., Rosset, S., Shay, T., Hadid, Y., . . . Skorecki, K. (2008). Counting the founders: the matrilineal genetic ancestry of the Jewish diaspora. *PLos ONE*, e2062.

Benazzi, S., Slon, V., Talamo, S., NEgrino, F., Peresani, M., Bailey, S., . . . Hublin, J. (2015). Archaeology. The makers of the Protoaurignacian and implications for Neandertal extinction. *Science*, 793-796.

Benson, D., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D., Ostell, J., & Sayers, E. (2013). GenBank. *Nucleic Acids Research*, D36-D42.

Berger, L., Hawks, J., Ruiter, D., Churchill, S., Schmid, P., Delezene, L., . . . Zipfel, B. (2015). Homo naledi, a new species of the genus Homo from the Dinaledi Chamber, South Africa. *eLife*, e09560.

Bertranpetit, J., Sala, J., Calafell, F., Underhill, P., Moral, P., & Comas, D. (1995). Human mitochondrial DNA variation and the origin of Basques. *Annals of Human Genetics*, 63-81.

Blast. (2015). Obtido de http://blast.ncbi.nlm.nih.gov/Blast.cgi

Boattini, A., Martinez-Cruz, B., Sarno, S., Harmant, C., Useli, A., Sanz, P., . . . Consortium, t. G. (2013). Uniparental markers in Italy reveal a sex-biased genetic structure and different historical strata. *PLoS ONE*, e65441.

Bosch, E., Calafell, F., González-Neira, A., Flaiz, C., Mateu, E., Scheil, H., . . . Comas, D. (2006). Paternal and maternal lineages in the Balkans show a homogeneous landscape over linguistic barriers, except for the isolated Aromuns. *Annals of Human Genetics*, 459-487.

Bouchard-Côté, A., Hall, D., Griffiths, T., & Klein, D. (2013). Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences of the U.S.A.*, 4224-4229.

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S., Alekseyenko, A., Drummond, A., . . . Atkinson, Q. (2012). Mapping the Origins and Expansion of the Indo-European language family. *Science*, 957-960.

Brandstätter, A., Sänger, T., Lutz-Bonengel, S., Parson, W., Béraud-Colomb, E., Wen, B., . . . Bandelt, H. (2005). Phantom mutation hotspots in human mitochondrial DNA. *Electrophoresis*, 3414-3429.

Brandt, G., Haak, W., Adler, C., Roth, C., Szécsényi-Nagy, A., Karimnia, S., . . . The Genographic Consortium. (2013). Ancient DNA reveals key stages in the formation of central european mitochondrial genetic diversity. *Science*, 257-261.

Brisighelli, F., Álvarez-Iglesias, V., Fondevila, M., Blanco-Verea, A., Carracedo, Á., Pascali, V., . . . Salas, A. (2012). Uniparental markers of contemporary Italian population reveals details on its pre-Roman heritage. *PLoS ONE*, e50794.

Burckhardt, F., von Haeseler, A., & Meyer, S. (1999). HvrBase: compilation of mtDNA control region sequences from primates. *Nucleic Acids Research*, 138-142.

Calafell, F., Underhill, P., Tolun, A., Angelicheva, D., & Kalaydjieva, L. (1996). From Asia to Europe: mitochondrial DNA sequence variability in Bulgarians and Turks. *Annals of Human Genetics*, 35-49.

Cann, R. (2001). Genetic clues to dispersal in human populations: retracing the past from the present. *Science*, 1742-1748.

Caramelli, D., Lalueza-Fox, C., Vernesi, C., Lari, M., Casoli, A., Mallegni, F., . . . Bertorelle, G. (2003). Evidence for a genetic discontinuity between neandertals and 24,000-year-old anatomically modern europeans. *Proceedings of the National Academy of Sciences of the U.S.A.*, 6593-6597.

Cardoso, S., Valverde, L., Alfonso-Sánchez, M., Palencia-Madrid, L., Elcoroaristizabal, X., Algorta, J., . . . Pancorbo, M. (2013). The expanded mtDNA phylogeny of the Franco-Cantabrian region upholds the pre-neolithic genetic substrate of Basques. *PLoSONE*, e67835.

Cardoso, S., Villanueva-Millán, M., Valverde, L., Odriozola, A., Aznar, J., Piñeiro-Herminda, S., & Pancorbo, M. (2012). Mitochondrial DNA control region variation in an autochthonous

Basque population sample from the Basque Country. *Forensic Population Genetics*, e106-e108.

CARTA. (2015). Obtido de http://carta.anthropogeny.org/

Cavalli-Sforza, L., Piazza, A., Menozzi, P., & Mountain, J. (1988). Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proceedings of the National Academy of Sciences of the U.S.A.*, 6002-6006.

Chang, W., Cathcart, C., Hall, D., & Garret, A. (2015). Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language*, 194-244.

Chen, F., Wang, S.-Y., Zhang, R.-Z., Hu, Y.-H., Gao, G.-F., Liu, Y.-H., & Kong, Q.-P. (2008). Analysis of mitochondrial DNA polymorphisms in Guangdong Han Chinese. *Forensic Science International Genetics*, 150-153.

Chikhi, L., Destro-Bisol, G., Bertorelle, G., Pascali, V., & Barbujani, G. (1998). Clines of nuclear DNA markers suggest a largely Neolithic ancestry of the European gene pool. *Proceedings of the National Academy of Sciences of the U.S.A.*, 9053-9058.

Colonna, V., Boattini, A., Guardiano, C., Dall'Ara, I., Pettener, D., Longobardi, G., & Barbujani, G. (2010). Long-range comparison between genes and languages based on syntactic distances. *Human Hereditary*, 245-254.

CRAN. (2015). Obtido de http://cran.r-project.org/

Creanza, N., Ruhlen, M., Pemberton, T., Rosenberg, N., Feldman, M., & Ramachadran, S. (2015). A comparison of worldwide phonemic and genetic variation in human populations. *Proceedings of the National Academy of Sciences of the U.S.A.*, 1265-1272.

Cunningham, F., Amode, M., Barrell, D., Beal, K., Billis, K., Brent, S., . . . Flicek, P. (2015). Ensembl 2015. *Nucleic Acids Research*, D662-D669.

Darwin, C. (1859). *The origin of species by means of natural selection.* Oxford, U.K.: Oxford University Press.

Darwin, C. (1871). *The descent of man.* London, U.K.: Murray.

Derenko, M., Grzybowski, T., Malyarchuk, B., Dambueva, I., Denisova, G., Czarny, J., . . . Zakharov, I. (2003). Diversity of mitochondrial DNA Lineages in South Siberia. *Annals of Human Genetics*, 391-411.

Derenko, M., Malyarchuk, B., Grzybowski, T., Denisova, G., Dambueva, I., Perkova, M., . . . Zakharov, I. (2007). Phylogeographic Analysis of mitochondrial DNA in Northern Asian populations. *The American Journal of Human Genetics*, 1025-1041.

Di Benedetto, G., Ergüven, A., Stenico, M., Castrì, L., Bertorelle, G., Togan, I., & Barbujani, G. (2001). DNA diversity and population admixture in Anatolia. *American Journal of Physical Anthropology*, 144-156.

Dirks, P., Berger, L., Roberts, E., Kramers, J., Hawks, J., Randolph-Quinney, P., . . . Tucker, S. (2015). Geological and taphonomic context for the new hominin species Homo naledi from the Dinaledi Chamber, South Africa. *eLife*, e09561.

Dunn, M., Terrill, A., Reesink, G., Foley, R., & Levinson, S. (2005). Structural phylogenetics and the reconstruction of ancient language history. *Science*, 2072-2075.

Edgar, R. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 1792-1797.

Ennafaa, H., Cabrera, V., Abu-Amero, K., González, A., Amor, M., Bouhaha, R., . . . Larruga, J. (2009). Mitochondrial DNA haplogroup H structure in North Africa. *BMC Genetics*.

Ensembl. (2015). Obtido de http://www.ensembl.org/index.html

Ethnologue. (2015). Obtido de http://www.ethnologue.com/

Excoffier, L., & Lisher, H. (2010). Arlequin suite ver 3.5: A new series of programs population genetics analysis under Linux and Windows. *Molecular Ecology Resources*, 564-567.

Excoffier, L., Smouse, P., & Quattro, J. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics*, 479-491.

Falchi, A., Giovannoni, L., Calo, C., Piras, I., Moral, P., Paoli, G., . . . Varesi, L. (2006). Genetic history of some western Mediterranean human isolates through mtDNA HVR1 polymorphisms. *Journal of Human Genetics*, 9-14.

Finnilä, S., Lehtonen, M., & Majamaa, K. (2001). Phylogenetic Network for European mtDNA. *American Journal of Human Genetics*, 1475-1484.

Fisher, R. (1925). *Statistical methods for the research workers.* Edinburgh, UK: Oliver and Boyd.

Fregel, R., & Delgado, S. (2011). HaploSearch: A tool for haplotype-sequence two-way transformation. *Mitochondrion*, 366-367.

Fu, Q., Rudan, P., Pääbo, S., & Krause, J. (2012). Complete mitochondrial genomes reveal neolithic expansion into Europe. *PLoS ONE*, e32743.

García, O., Fregel, R., Larruga, J., Álvarez, V., Yurrebaso, I., Cabrera, V., & González, A. (2011). Using mitochondrial DNA to test the hypothesis of a European post-glacial human recolonization from the Franco-Cantabrian refuge. *Heredity*, 37-45.

GenBank. (2015). Obtido de http://www.ncbi.nlm.nih.gov/genbank/

Ghirotto, S., Penso-Dolfin, L., & Barbujani, G. (2011). Genomic evidence for an African expansion of anatomically modern humans by a southern route. *Human Biology*, 477-489.

Gibbons, A. (2015). Deep roots for the genus Homo. *Science*, 1056-1057.

Gibert, M., Theves, C., Ricaut, F., Dambueva, I., Bazarov, B., Moral, P., . . . Sevin, A. (2010). mtDNA variation in the Buryat population of the Barguzin Valley: new insights into the micro-evolutionary history of the Baikal area. *Annals of Human Biology*, 501-523.

González, A., Brehm, A., Pérez, J., Maca-Meyer, N., Flores, C., & Cabrera, V. (2003). Mitochondrial DNA affinities at the Atlantic fringe of Europe. *American Journal of Physical Anthropology*, 391-404.

Gravel, S. (2012). Population genetics models of local ancestry. *Genetics*, 607-619.

Gray, R., & Atkinson, Q. (2003). Language tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 435-439.

Gray, R., Atkinson, Q., & Greenhill, S. (2011). Language evolution and human history: what a difference a data makes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 1090-1100.

Gray, R., Drummond, A., & Greenhill, S. (2009). Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science*, 479-483.

Grzybowski, T., Malyarchuk, B., Derenko, M., Perkova, M., Bednarek, J., & Woźniak, M. (2007). Complex interactions of the Eastern and Western Slavic populations with other European groups as revealed by mitochondrial DNA analysis. *Forensic Science International: Genetics*, 141-147.

Guardiano, C., & Longobardi, G. (2005). Parametric Comparison and Language Taxonomy. Em M. Batllori, M.-L. Hernanz, C. Picallo, & F. (. Roca, *Grammaticalization and Parametric Variation* (pp. 149-174). Oxford, UK: Oxford University Press.

Gusmão, L., Sánchez-Diz, P., Calafell, F., Martín, P., Alonso, C., Álvarez-Fernández, F., . . . Amorim, A. (2005). Mutation rates at Y chromosome specific microsatellites. *Human mutation*, 520-528.

Haak, W., Balanovsky, O., Sanchez, J., Koshel, S., Zaporozhchenko, V., Adler, C., . . . the Genographic Consortium. (2010). Ancient DNA from European early neolithic farmers reveals their Near Eastern affinities. *PLoS Biology*, e1000536.

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., . . . Reich, D. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*, 207-211.

Haber, M., Youhanna, S., Balanovsky, O., Saade, S., Martínez-Cruz, B., Ghassibe-Sabbagh, M., . . . Zalloua, P. (2012). mtDNA lineages reveal coronary artery disease-associated structures in the Lebanese population. *Annals of Human Genetics*, 1-8.

Hall, T. (1999). BioEdit: a user-friendly biological sequence alignment and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, 95-98.

Haplosite. (2015). Obtido de http://www.haplosite.com/

Harpending, H., Batzer, M., Gurven, M., Jorde, L., Rogers, A., & Sherry, S. (1998). Genetic traces of ancient demography. *Proceedings of the National Academy of Sciences of the U.S.A.*, 1961-1967.

Hedman, M., Brandstätter, A., Pimenoff, V., Sistonen, P., Palo, J., Parson, W., & Sajantila, A. (2007). Finnish mitochondrial DNA HVS-I and HVS-II population data. *Forensic Science International*, 171-178.

Helgason, A., Hickey, E., Goodacre, S., Bosnes, V., Stefánson, K., Ward, R., & Sykes, B. (2001). mtDNA and the islands of the North Atlantic: estimating the proportions of Norse and Gaelic ancestry. *American Journal of Human Genetics*, 723-737.

Helgason, A., Pálsson, G., Pedersen, H., Angulalik, E., Gunnarsdóttir, E., Ygvadóttir, B., & Stefánsson, K. (2006). mtDNA variation in Inuit populations of Greenland and Canada: migration history and population structure. *American Journal of Physical Anthropology* , 123-134.

Helgason, A., Sigurðardóttir, S., Gulcher, J., Ward, R., & Stefánsson, K. (2000). mtDNA and the origin of the Icelanders: deciphering signals of recent population history. *American Journal of Human Genetics*, 999-1016.

Hoffecker, J. (2009). The spread of modern humans in Europe. *Proceedings of the National Acadmy of Sciences of the U.S.A.*, 16040-16045.

Hofmann, S., Jaksch, M., Bezold, R., Mertens, S., Aholt, S., Paprotta, A., & Gerbitz, K.-D. (1997). Population genetics and disease susceptibility: characterization of central European haplogroups by mtDNA gene mutations, correlation with D loop variants and association with disease. *Human Molecular Genetics*, 1835-1846.

Horai, S., Murayama, K., Hayasaka, K., Matsubayashi, S., Hattori, Y., Fucharoen, G., . . . Pan, I.-H. (1996). mtDNA polymorpism in East Asian populations, with special reference to the peopling of Japan. *American Journal of Human Genetics*, 579-590.

Hublin, J. (2014). Paleoanthropology: Homo erectus and the limits of paleontological species. *Current Biology*, R82-R84.

Hunley, K. (2015). Reassessment of global gene-language coevolution. *Proceedings of the National Academy of Sciences of the U.S.A.*, 1919-1920.

HvrBase++. (2015). Obtido de http://hvrbase.cibiv.univie.ac.at/index.html

Ingmann, M., Kaessmann, H., Pääbo, S., & Gyllensten, U. (2000). Mitochondrial genome variation and the origin of modern humans. *Nature*, 708-712.

Irwin, J., Egyed, B., Saunier, J., Szamosi, G., O'Callaghan, J., Padar, Z., & Parsons, T. (2007). Hungarian mtDNA population databses from Budapest and the Baranya county Roma. *International Journal of Legal Medicine*, 377-383.

Irwin, J., Saunier, J., Strouss, K., Paintner, C., Diegoli, T., Sturk, K., . . . Parsons, T. (2008). Mitochondrial control region sequences from northern Greece and Greek Cypriots. *International Journal of Legal Medicine*, 87-89.

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 223-270.

Jaworska, N., & Chupetlovska-Anastosova, A. (2009). A review of Multidimensional Scaling (MDS) and its utility in various psychological domains. *Tutorials in Quantitative Methods for Psychology*, 1-10.

Karachanak, S., Carossa, V., Nesheva, D., Olivieri, A., Pala, M., Hooshiar, B., . . . Torroni, A. (2012). Bulgarians vs the other European populations: a mitochondrial DNA perspective. *International Journal of Legal Medicine*, 497-503.

Kivisild, T., Bamshad, M., Kaldma, K., Metspalu, M., Metspalu, E., Reidla, M., . . . Villems, R. (1999). Deep common ancestry of Indian and Western-Eurasian mitochondrial DNA lineages. *Current Biology*, 1331-1334.

Kivisild, T., Tolk, H.-V., Parik, J., Wang, Y., Papiha, S., Bandelt, H.-J., & Villems, R. (2002). The emerging limbs and twigs of the East Asian mtDNA tree. *Molecular Biology and Evolution*, 1737-1751.

Klein, R. (2000). Archeology and evolution of human behavior. *Evolutionary Anthropology*, 17-36.

Klein, R. (2008). Out of Africa and the evolution of Human Behavior. *Evolutionary Anthropology*, 267-281.

Kohl, J., Paulsen, I., Laubach, T., Radtke, A., & von Haeseler, A. (2005). HvrBase++: a phylogenetic database for primate species. *Nucleic Acids Research*, D700-D704.

Kouvatsi, A., Karaiskou, N., Apostolidis, A., & Kirmizidis, G. (2001). Mitochondrial DNA sequence variation in Greeks. *Human Biology*, 855-869.

Kruskal, J. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 1-27.

Kruskal, J. (1964b). Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 115-129.

Laland, K., Odling-Smee, J., & Myles, S. (2010). How culture shaped the human genome: bringing genetics and the human sciences together. *Nature Reviews Genetics*, 137-148.

LanGeLin. (2015). Obtido de https://www.york.ac.uk/language/research/projects/langelin/

Lansing, J., Cox, M., Downey, S., Gabler, B., Hallmark, B., Karafet, T., . . . Hammer, M. (2007). Coevolution of languages and genes on the islanf of Sumba, eastern Indonesia. *Proceedings of the National Academy of Sciences of the U.S.A.*, 16022-16026.

Lao, O., Lu, T., Nothnagel, M., Junge, O., Freitag-Wolf, S., Caliebe, A., . . . Kayser, M. (2008). Correlation between genetic and geographic structure in Europe. *Current Biology*, 1241-1248.

Larruga, J., Díez, F., Pinto, F., Flores, C., & González, A. (2001). Mitochondrial DNA characterisation of European isolates: the Maragatos from Spain. *European Journal of Human Genetics*, 708-716.

Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., . . . Krause, J. (2014). Ancient human genomes suggest three ancestral populations for present-day europeans. *Nature*, 409-413.

Legendre, P., & Legendre, L. (1998). *Numerical Ecology. 2nd English Edition.* Elsevier.

Lewis, M., Simons, G., & Fennig, C. (. (2015). *Ethnologue: Languages of the World, Eighteenth edition.* Dallas, USA: SIL International.

Longobardi, G. (2003). Methods in parametric linguistics and cognitive history. *Linguistic Variation Yearbook*, 101-138.

Longobardi, G., & Guardiano, C. (2009). Evidence for syntac as a signal of historical relatedness. *Lingua*, 1679-1706.

Lutz, S., Weisser, H.-J., Heizmann, J., & Pollak, S. (1998). Location and frequency of polymorphic positions in the mtDNA control region of individuals from Germany. *International Journal of Legal Medicine*, 67-77.

M., M., Sørensen, E., Rasmussen, E., & Morling, N. (2010). Mitochondrial DNA HV1 and HV2 variation in Danes. *Forensic Science International: Genetics*, e87-e88.

Mabuchi, T., Susukida, R., Kido, A., & Oya, M. (2007). Typing the 1.1kb control region of Human mitochondrial DNA in Japanese individuals. *Journal of Forensic Sciences*, 355-363.

Malyarchuk, B., Grzybowski, T., Derenko, M., Czarny, J., Drobnič, K., & Miścicka-Śliwka, D. (2003). Mitochondrial DNA variability in Bosnians Slovenians. *Annals of Human Genetics*, 412-425.

Mantel, N. (1967). The detection of disease clustering and generalized regression approach. *Cancer Research*, 209-220.

Marques-Bonet, T., Kidd, J., Ventura, M., Graves, T., Cheng, Z., Hillier, L., . . . Eichler, E. (2009). A burst of segmental duplications in the genome of the african great ape ancestor. *Nature*, 877-881.

Masel, J. (2011). Genetic drift. *Current Biology*, R837-R838.

McEvoy, B., Richards, M., Forster, P., & Bradley, D. (2004). The Longue Durée of genetic ancestry: multiple genetic markers systms and Celtic origins on the Atlantic facade of Europe. *American Journal of Human Genetics*, 693-702.

Metspalu, M., Kivisild, T., Metspalu, E., Parik, J., Hudjashov, G., Kaldma, K., . . . Villems, R. (2004). Most of the extant mtDNA boundaries in South and Southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. *BMC Genetics*.

Meyer, M., Arsuaga, J.-L., Nagel, S., Martínez, I., Gracia, A., Castro, J., . . . Pääbo, S. (2015). Nuclear DNA sequences from the hominin remains of Sima de los Huesos, Atapuerca, Spain. *Proceedings of the European Society for the study of Human Evolution 4*, (p. 161). London, U.K.

Mielnik-Sikorska, M., Daca, P., Malyarchuk, B., Derenko, M., Skonieczna, K., Perkova, M., . . . Grzybowski, T. (2013). The history of Slavs inferred from complete mitochondrial genome sequences. *PLoS ONE*, e54360.

Miller, R., & Kwok, P. (2001). The birth and death of human single-nucleotide polymorphisms: new experimental evidence and implications for human history and medicine. *Human Molecular Genetics*, 2195-2198.

Mogentale-Profizi, N., Chollet, L., Stévanovitch, A., Dubut, V., Poggi, C., Pradié, M., . . . Béraud-Colomb, E. (2001). Mitochondrial DNA sequences diversity in two groups of Italian Veneto speakers from Veneto. *Annals of Human Genetics*, 153-166.

Morozova, I., Evsyukov, A., Kon'kov, A., Grosheva, A., Zhukova, O., & Rychkov, S. (2012). Russian ethnic history inferred from mitochondrial DNA diversity. *American Journal of Physical Anthropology*, 341-351.

Myres, N., Rootsi, S., Lin, A., Järve, M., King, R., Kutuev, I., . . . Underhill, P. (2011). A major Y-chromossome haplogroup R1b Holocene era founder effect in central and western Europe. *European Journal of Human Genetics*, 95-101.

Nakagawa, S. (2004). A farewell to Bonferroni: the problems of low statistical power and publication bias. *Behavioral Ecology*, 1044-1045.

NASA. (2015). *Visible Earth*. Obtido de http://visibleearth.nasa.gov/

NCBI. (2015). Obtido de http://www.ncbi.nlm.nih.gov/

Norton, H., Kittles, R., PArra, E., McKeigue, P., Mao, X., Cheng, K., . . . Shriver, M. (2007). Genetic evidence for the convergent evolution of light skin in Europeans and East Asians. *Molecular Biology and Evolution*, 710-722.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A., Auton, A., . . . Bustamante, C. (2008). Genes mirror geography within Europe. *Nature*, 98-101.

Oksanen, J., Blanchet, F., Kindt, R., Legendre, P., Minchin, P., O'Hara, R., . . . Stevens, M. W. (2015). vegan: Comunity Ecology Package. R package version 2.3-0. http://CRAN.R-project.org/package=vegan.

Oota, H., Kitano, T., Jin, F., Yuasa, I., Wang, L., Ueda, S., . . . Stoneking, M. (2002). Extreme mtDNA homogeneity in Continental Asian populations. *American Journal of Physical Anthropology*, 146-153.

Ottoni, C., Martinez-Labarga, C., Vitelli, L., Scano, G., Fabrini, E., Contini, I., . . . Rickards, O. (2009). Human mitochondrial DNA variation in Southern Italy. *Annals of Human Biology*, 785-811.

Pajnič, I., Balažic, J., & Komel, R. (2004). Sequence polymorphism of the mitochondrial DNA control region in the Slovenian population. *International Journal of Legal Medicine*, 1-4.

Pakendorf, B. (2014). Coevolution of languages and genes. *Current Opinion in Genetics & Development*, 39-44.

Pakendorf, B., & Stoneking, M. (2005). Mitochondrial DNA and human evolution. *Annual Review of Genomics and Human Genetics*, 165-183.

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., . . . Reich, D. (2012). Ancient admixture in human history. *Genetics*, 1065-1093.

Pereira, L., Cunha, C., & Amorim, A. (2004). Predicting sampling saturation of mtDNA haplotypes: an application to an enlarged Portuguese database. *International Journal of Legal Medicine*, 132-136.

Pereira, L., Dupanloup, I., Rosser, Z., Jobling, M., & Barbujani, G. (2001). Y-chromosome mismatch distributions in Europe. *Molecular Biology and Evolution*, 1259-1271.

Pérez-Lezaun, A., Calafell, F., Comas, D., Mateu, E., Bosch, E., Martínez-Arias, R., . . . Bertranpetit, J. (1999). Sex-specific migration patterns in Central Asian populations, revealed by Analysis of Y-chromosome Short Tandem Repeats and mtDNA. *American Journal of Human Genetics*, 208-219.

Prieto, L., Zimmermann, B., Goios, A., Rodriguez-Monge, A., Paneto, G., Alves, C., . . . Parson, W. (2011). The GHEP-EMPOP collaboration on mtDNA population data - A new resource for forensic casework. *Foresnci Science International: Genetics*, 146-151.

Prugnolle, F., & Meeus, T. (2002). Inferring sex-biased dispersal from population genetic tools: a review. *HEredity*, 161-165.

R Core Team. (2015). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna, Austria: http://www.R-project.org/.

Rakha, A., Shin, K.-J., Yoon, J., Kim, N., Siddique, M., Yang, I., . . . Lee, H. (2011). Forensic and genetic characterization of mtDNA from Pathans of Pakistan. *International Journal of Legal Medicine*, 841-848.

Ramachandran, S., Deshpande, O., Roseman, C., Rosenberg, N., Feldman, M., & Cavalli-Sforza, L. (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences of the U.S.A.*, 15942-15947.

Rando, J., Pinto, F., González, A., Hernández, M., Larruga, J., Cabrera, V., & Bandelt, H.-J. (1998). Mitochondrial DNA analysis of Northwest African populations reveals genetic exchange with European, Near-Eastern, and Sub-Saharan populations. *Annals of Human Genetics*, 531-550.

Raule, N., Sevini, F., Li, S., Barbieri, A., Tallaro, F., Lomartire, L., . . . Franceschi, C. (2014). The co-occurence of mtDNA mutations on different oxidative phosphorylation subunits, not detected by haplogroup analysis, affects human longevity and is population specific. *Aging Cell*, 401-407.

Reich, D., Thangaraj, K., Patterson, N., Price, A., & Singh, L. (2009). Reconstructing indian population history. *Nature*, 489-494.

Relethford, J. (2001). Ancient DNA and the origin of modern humans. *Proceedings og the National Academy of Sciences of the U.S.A.*, 390-391.

Relethford, J. (2008). Genetic evidence and the modern human origins debate. *Heredity*, 555-563.

Reyes-Centeno, H., Ghirotto, S., Détroit, F., Grimaud-Hervé, D., Barbujani, G., & HArvati, K. (2014). Genomic and cranial phenotype data support multiple modern human dispersals from Africa and a southern route into Asia. *Proceedings of the National Academy of Sciences of the U.S.A.*, 7248-7253.

Rice, R. (1989). Analyzing tables of statistical tests. *Evolution*, 223-225.

Richards, M., Côrte-Real, H., Forster, P., Macaulay, V., Wilkinson-Herbots, H., Demaine, A., . . . Sykes, B. (1996). Paleolithic and neolithic lineages in the Euopean mitochondrial gene pool. *American Journal of Human Genetics*, 185-203.

Sajantila, A., Lahermo, P., Anttinen, T., Lukka, M., Sistonen, P., Savontaus, M.-L., . . . Pääbo, S. (1995). Genes and Languages in Europe: an analysis of mitochondrial lineages. *Genome Research*, 42-52.

Sajantila, A., Salem, A.-H., Savolainen, P., Bauer, K., Gierig, C., & Pääbo, S. (1996). Paternal and maternal DNA lineages reveal a bottleneck in the founding of the Finnish population. *Proceedings of the National Academy of Sciences of the U.S.A.*, 12035-12039.

Schönberg, A., Theunert, C., Li, M., Stoneking, M., & Nasidze, I. (2011). High-throughput sequencing of complete human mtDNA genomes from the Caucasus and West Asia: high diversity and demographic inferences. *European Journal of Human Genetics*, 988-994.

Schwartz, J., & Tattersall, I. (2015). Defining the genus Homo. *Science*, 931-932.

Seguin-Orlando, A., Korneliussen, T., Sikora, M., Malaspinas, A., Manica, A., Moltke, I., . . . Willerslev, E. (2014). Genomic structure in europeans dating back at least 36,200 years. *Science*, 1113-1118.

Seielstad, M., Minch, E., & Cavalli-Sforza, L. (1998). Genetic evidence for a higher female migration rate in humans. *Nature Genetics*, 278-280.

Serre, D., Langaney, A., Chech, M., Techler-Nicola, M., Paunovic, M., Mennecier, P., . . . Pääbo, S. (2004). No evidence of Neandertal mtDNA contribution to early modern Humans. *PLoS Biology*, 0313-0317.

Sharma, G., Tamang, R., Chaudhary, R., Singh, V., Shah, A., Anugula, S., . . . Thangaraj, K. (2012). Genetic affinities of the Central Indian tribal populations. *PLoS ONE* , e32546.

Shepard, R. (1962). The analysis of proximities: multidimensional scaling with an unknown distance function. *Psychometrika*, 219-246.

Skoglund, P., Malmström, H., Raghavan, M., Storå, J., Hall, P., Willerslev, E., . . . Jakobsson, M. (2012). Origins and genetic legacy of neolithic farmers and hunter-gatherers in Europe. *Science*, 466-469.

Sokal, R. (1988). Genetic, geographic, and linguistic distances in Europe. *Proceedings of the National Academy of Sciences of the U.S.A.*, 1722-1726.

Sokal, R., Oden, N., Legendre, P., Fortin, M., Kim, J., Thomson, B., . . . Barbujani, G. (1990). Genetics and language in european populations. *American Naturalist*, 157-175.

Starikovskaya, E., Sukernik, R., Derbeneva, O., Volodko, N., Ruiz-Pesini, E., Torroni, A., . . . Wallace, D. (2005). Mitochondrial DNA diversity in indigenous populations of the Southern Extent of Siberia, and the origins of Native American haplogroups. *Annals of Human Genetics*, 67-89.

Stumpf, M., & Goldstein, D. (2001). Genealogical and evolutionary inference with the human Y chromosome. *Science*, 1738-1742.

Templeton, A. (2002). Out of Africa again and again. *Nature*, 45-51.

Tetzlaff, S., Brandstätter, A., Wegener, R., Parson, W., & Weirich, V. (2007). Mitochondrial DNA population data of HVS-I and HVS-II sequences from a northeast German sample. *Forensic Science International* , 218-224.

Thalmann, O., Fischer, A., Lankester, F., Pääbo, S., & Vigilant, L. (2007). The complex evolutionary history of gorillas: insights from genomic data. *Molecular Biology and Evolution*, 16-158.

Thangaraj, K., Naidu, B., Crivellaro, F., Tamang, R., Upadhyay, S., Sharma, V., . . . Singh, L. (2010). The influence of natural barriers, in shaping the genetic structure of Maharashtra populations. *PLoS ONE*, e15283.

The Chimpanzee Sequencing and Analysis Consortium. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 69-8.

Tishkoff, S., Reed, F., Friedlaender, F., Ehret, C., Ranciaro, A., Froment, A., . . . Williams, S. (2009). The genetic structure and history of Africans and African Americans. *Science*, 1035-1044.

Varki, A., Geschwind, D., & Eichler, E. (2008). Explaining the human uniqueness: genome interactions with environment, behaviour and culture. *Nature Reviews Genetics*, 749-763.

Venables, W., & Ripley, B. (2002). *Modern applied Statistics with S. Forth Edition.* New York, USA: Springer. Obtido de http://CRAN.R-project.org/package=MASS

Vernesi, C., Di Benedetto, G., Caramelli, D., Secchieri, E., Simoni, L., Katti, E., . . . Barbujani, G. (2001). Genetic characterization of the body attributed to the evangelist Luke. *Proceedings of the National Academy of Sciences of the U.S.A.*, 13460-13463.

Wang, C., Farina, S., & Li, H. (2013). Neanderthal DNA and modern human origins. *Quaternary International*, 126-129.

Warnes, G., Bolker, B., Gorjanc, G., Grothendieck, G., Korosec, A., Lumley, T., . . . et al. (2015). gdata: Various R Programming Tools for Data Manipulation. R package version 2.17.0. http://CRAN.R-project.org/package=gdata.

Weisstein, E. (2015). *Great Circle*. Obtido de Math World - A Wolfram Web Resource: http://mathworld.wolfram.com/GreatCircle.html

Welch, W. (1990). Construction of Permutation Tests. *Journal of the American Statistical Association*, 693-698.

Wen, B., Xie, X., Gao, S., Li, H., Shi, H., Song, X., . . . Jin, L. (2004). Analyses of genetic structure of Tibeto-Burman populations reveals sex-biased admixture in southern Tibeto-Burmans. *American Journal of Human Genetics*, 856-865.

Wickelmaier, F. (2003). *An introduction to MDS.* Aalborg University: Department of Acoustics, Institute of Electronic Systems.

Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics*, 323-354.

Yao, Y., Bravi, C., & Bandelt, H. (2004). A call for mtDNA data quality control in forensic science. *Forensic Science International* , 1-6.

Yao, Y., Salas, A., Logan, I., & Bandelt, H. (2009). mtDNA data mining in GenBank needs surveying. *American Journal of Human Genetics*, 929-933.

Yao, Y.-G., Kong, Q.-P., Bandelt, H.-J., Kisild, T., & Zhang, Y.-P. (2002). Phylogeographic differentiation of mitochondrial DNA in Han Chinese. *American Journal of Human Genetics*, 635-651.

ANNEX

1. Data-mining

Table 15 - mtDNA HVR 1 references per population and author

Language	References	Nº of Samples
Arabic	Al Balwi (unpublished)	1
	Al-Zahery, et al., 2011	315
	Badro, et al., 2013	1213
	Haber, et al., 2012	363
Bulgarian	Calafell, Underhill, Tolun, Angelicheva, & Kalaydjieva, 1996	30
	Karachanak, et al., 2012	853
Buriat	Derenko, et al., 2003	91
	Derenko, et al., 2007	295
	Gibert, et al., 2010	61
	Ingmann, Kaessmann, Pääbo, & Gyllensten, 2000	1
	Starikovskaya, et al., 2005	25
Cantonese	Chen, et al., 2008	106
	Kivisild, et al., 2002	69
	Yao, Kong, Bandelt, Kisild, & Zhang, 2002	30
central Basque	Bertranpetit, et al., 1995	45
	Cardoso, et al., 2012	34
	Cardoso, et al., 2013	210
	García, et al., 2011	115
	Prieto, et al., 2011	3
Cypriot Greek	Irwin, et al., 2008	91
Danish	Leicester (personal communication)	20
	M., Sørensen, Rasmussen, & Morling, 2010	201
	Raule, et al., 2014	429
	Richards, et al., 1996	33
English	García, et al., 2011	9
	Helgason, et al., 2001	142
	Ingmann, Kaessmann, Pääbo, & Gyllensten, 2000	1
	Leicester (personal communication)	20
Estonian	Sajantila, et al., 1995	28
	Sajantila, et al., 1996	20
Farsi	Metspalu, et al., 2004	435
	Schönberg, Theunert, Li, Stoneking, & Nasidze, 2011	30

Language	References	Nº of Samples
Finish	Finnilä, Lehtonen, & Majamaa, 2001	192
	Hedman, et al., 2007	200
	Raule, et al., 2014	146
	Sajantila, et al., 1995	49
French	Badro, et al., 2013	790
	García, et al., 2011	33
	Ingmann, Kaessmann, Pääbo, & Gyllensten, 2000	1
German	García, et al., 2011	11
	Hofmann, et al., 1997	67
	Leicester (personal communication)	20
	Lutz, Weisser, Heizmann, & Pollak, 1998	200
	Richards, et al., 1996	156
	Tetzlaff, Brandstätter, Wegener, Parson, & Weirich, 2007	213
Greek	Irwin, et al., 2008	317
	Kouvatsi, Karaiskou, Apostolidis, & Kirmizidis, 2001	54
	Leicester (personal communication)	20
	Raule, et al., 2014	14
	Vernesi, et al., 2001	48
Hebrew	Behar, et al., 2008	233
Hindi	Barnabas, Shouche, & Suresh, 2005	9
	Kivisild, et al., 1999	68
	Sharma, et al., 2012	143
Hungarian	Irwin, et al., 2007	415
	Leicester (personal communication)	20
Icelandic	Helgason, Sigurðardóttir, Gulcher, Ward, & Stefánsson, 2000	394
	Sajantila, et al., 1995	39
Inuit	Helgason, et al., 2006	96
	Simonson (unpublished)	46
Irish	Leicester (personal communication)	20
	McEvoy, Richards, Forster, & Bradley, 2004	299
Italian	Achilli, et al., 2007	321
	Boattini, et al., 2013	600
	Brisighelli, et al., 2012	352
	Falchi, et al., 2006	61
	Ingmann, Kaessmann, Pääbo, & Gyllensten, 2000	1
	Leicester (personal communication)	20
	Mogentale-Profizi, et al., 2001	68

Language	References	Nº of Samples
Italian	Ottoni, et al., 2009	92
Japanese	Horai, et al., 1996	62
	Ingmann, Kaessmann, Pääbo, & Gyllensten, 2000	2
	Mabuchi, Susukida, Kido, & Oya, 2007	124
	Oota, et al., 2002	89
Mandarin	Oota, et al., 2002	85
	Yao, Kong, Bandelt, Kisild, & Zhang, 2002	146
Marathi	Barnabas, Shouche, & Suresh, 2006	30
	Thangaraj, et al., 2010	185
Norwegian	Helgason, et al., 2001	323
	Leicester (personal communication)	20
Pashto	Rakha, et al., 2011	230
Polish	Grzybowski, et al., 2007	413
	Mielnik-Sikorska, et al., 2013	404
Portuguese	González, et al., 2003	299
	Pereira, Cunha, & Amorim, 2004	549
	Prieto, et al., 2011	240
Romanian	Bosch, et al., 2006	105
Russian	Grzybowski, et al., 2007	157
	Morozova, et al., 2012	365
Serbo-Croat	Babalini, et al., 2005	96
	Fu, Rudan, Pääbo, & Krause, 2012	38
	Leicester (personal communication)	20
Slovenian	Malyarchuk, et al., 2003	104
	Pajnič, Balažic, & Komel, 2004	129
Spanish	Falchi, et al., 2006	66
	García, et al., 2011	5
	Larruga, Díez, Pinto, Flores, & González, 2001	196
	Leicester (personal communication)	20
	Prieto, et al., 2011	316
Turkish	Di Benedetto, et al., 2001	17
	Schönberg, Theunert, Li, Stoneking, & Nasidze, 2011	29
Welsh	Richards, et al., 1996	92
Wolof	Ennafaa, et al., 2009	11
	Rando, et al., 1998	48

2. Dissimilarity matrices

2.1. Genomic matrices

Table 16 - mtDNA HVR 1 dissimilarity matrix

	Ar	Blg	Bur	Can	cB	CyG	D	Da	E	Est	Far	Fin	Fr	Grk	Heb	Hi	Hu	Ice
Ar	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Blg	0.00576	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Bur	0.14141	0.15737	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Can	0.09206	0.09652	0.06437	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
cB	0.0214	0.01388	0.17906	0.13147	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CyG	0.01292	0.01731	0.14028	0.08262	0.05392	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
D	0.0077	-0.00059	0.17182	0.10926	0.01244	0.02529	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Da	0.00667	0.00283	0.16347	0.10321	0.01374	0.0246	0.00083	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
E	0.00676	0.00071	0.15421	0.089	0.01305	0.02633	-0.00281	0.00056	0	NA	NA	NA	NA	NA	NA	NA	NA	NA
Est	0.00914	-0.0017	0.13701	0.07068	0.01783	0.01906	-0.00083	0.0046	0.00192	0	NA	NA	NA	NA	NA	NA	NA	NA
Far	0.00401	0.00893	0.12064	0.068	0.02384	0.01648	0.00924	0.00798	0.00831	0.00819	0	NA	NA	NA	NA	NA	NA	NA
Fin	0.01915	0.01024	0.13161	0.07794	0.02066	0.0211	0.01261	0.01613	0.01316	0.00349	0.01711	0	NA	NA	NA	NA	NA	NA
Fr	0.00771	0.00065	0.16899	0.1072	0.01288	0.01804	0.00031	0.00334	0.00091	0.00078	0.0126	0.01074	0	NA	NA	NA	NA	NA
Grk	0.00321	0.00078	0.14622	0.0868	0.01592	0.01968	0.00134	0.00293	0.00085	5.00E-05	0.00493	0.01211	0.0036	0	NA	NA	NA	NA
Heb	0.0184	0.03055	0.15669	0.10574	0.05504	0.03635	0.0318	0.02637	0.02945	0.02724	0.01524	0.04459	0.03665	0.02442	0	NA	NA	NA
Hi	0.06679	0.08054	0.06404	0.04419	0.11411	0.06241	0.09465	0.08925	0.08289	0.0668	0.05134	0.06735	0.09053	0.07417	0.08241	0	NA	NA
Hu	0.02943	0.03112	0.09551	0.05745	0.04662	0.04325	0.03741	0.0355	0.03225	0.0284	0.02293	0.02688	0.03589	0.02662	0.04445	0.04762	0	NA
Ice	0.00919	0.00856	0.16245	0.09719	0.02379	0.02295	0.00867	0.00558	0.00654	0.01267	0.01074	0.02229	0.00783	0.00786	0.02468	0.08495	0.03367	0
Inu	0.44841	0.49969	0.36948	0.417	0.58347	0.55964	0.53605	0.5127	0.58294	0.63767	0.43218	0.48811	0.52886	0.50925	0.50612	0.40395	0.4701	0.52693
Ir	0.00563	0.00162	0.15128	0.09434	0.015	0.01497	0.00117	0.00147	0.00166	0.00543	0.00863	0.01045	-0.00012	0.00112	0.03175	0.08087	0.03024	0.00591
It	0.00583	0.00063	0.16528	0.10397	0.01168	0.02067	0.00017	0.00266	-6.00E-05	0.00028	0.00986	0.01259	6.00E-04	0.00161	0.02991	0.08751	0.0324	0.00762
Jap	0.12951	0.14902	0.02615	0.03134	0.18558	0.1313	0.16848	0.15599	0.14724	0.124	0.10412	0.12685	0.16561	0.13637	0.14349	0.04497	0.0898	0.15528
Ma	0.08716	0.10232	0.08266	0.07099	0.12235	0.06922	0.11561	0.10981	0.10022	0.07862	0.06921	0.08137	0.11084	0.0973	0.09748	0.02288	0.07599	0.10364
Man	0.11217	0.12613	0.01926	0.02092	0.15725	0.10327	0.14132	0.13234	0.1197	0.09875	0.08851	0.10586	0.1387	0.11469	0.12528	0.03541	0.07246	0.12868
Nor	0.00909	0.00316	0.16771	0.10933	0.01175	0.02746	0.00109	0.00135	5.00E-04	0.00458	0.01089	0.01517	0.00252	0.0045	0.02828	0.09398	0.03831	0.00549
Pas	0.02155	0.03125	0.08764	0.05625	0.05844	0.02959	0.04031	0.03845	0.03477	0.02846	0.01489	0.03159	0.03679	0.0277	0.04054	0.02209	0.02255	0.03833
Po	0.00976	0.00132	0.15288	0.09136	0.01249	0.02221	0.0016	0.00416	0.0019	-0.00259	0.00965	0.00713	0.00248	0.00342	0.03296	0.07807	0.02926	0.0096
Ptg	0.00697	0.00182	0.14596	0.09018	0.01345	0.01725	0.003	0.00544	0.00282	-0.00069	0.00941	0.00873	0.00224	0.00296	0.03273	0.07294	0.0252	0.01044
Rm	0.00301	0.00177	0.15143	0.08824	0.01751	0.01975	0.00084	0.00084	0.00186	0.00774	0.00412	0.01878	0.00441	0.00177	0.01878	0.07283	0.0321	0.00627
Rus	0.00749	0.00269	0.14765	0.09271	0.01403	0.01735	0.00289	0.00353	0.0033	0.00045	0.0087	0.01156	0.00345	0.00433	0.02536	0.07467	0.03015	0.0066
SC	0.0135	0.00387	0.15115	0.08153	0.02256	0.02845	0.00305	0.0088	0.00217	0.00113	0.01523	0.01469	0.00362	0.00667	0.04069	0.08002	0.03757	0.01354
Slo	0.00978	0.00062	0.15138	0.08873	0.01534	0.02051	0.00018	0.00452	1.00E-04	-0.00178	0.01238	0.00966	-0.00017	0.00398	0.03377	0.08104	0.03175	0.00921
Sp	0.01036	0.0022	0.16747	0.10589	0.00894	0.02472	0.0018	0.00475	0.00175	0.00222	0.01367	0.00922	0.00099	0.00438	0.04499	0.09518	0.03764	0.01147
Tur	0.01374	0.01625	0.08929	0.04319	0.04027	0.02254	0.01947	0.01929	0.02049	0.01112	0.00812	0.01504	0.02173	0.0144	0.03537	0.0463	0.02968	0.02767
Wel	-0.02896	-0.03863	0.14495	0.08569	-0.02317	0.00034	-0.02503	-0.03986	-0.05685	-0.01041	-0.0372	-0.00855	-0.02854	-0.04206	-0.01933	0.04825	0.01591	-0.02442
Wo	0.16146	0.19006	0.13004	0.12961	0.23887	0.14535	0.21674	0.20177	0.19793	0.1534	0.13335	0.16387	0.20999	0.18024	0.16666	0.1028	0.11859	0.19377

Table 16 - mtDNA HVR 1 dissimilarity matrix - continuation

	Inu	Ir	It	Jap	Ma	Man	Nor	Pas	Po	Ptg	Rm	Rus	SC	Slo	Sp	Tur	Wel	Wo
Ar	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Blg	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Bur	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Can	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
cB	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CyG	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
D	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Da	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
E	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Est	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Far	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Fin	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Fr	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Grk	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Heb	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Hi	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Hu	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Ice	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Inu	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Ir	0.55753	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
It	0.49993	0.00105	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Jap	0.3679	0.14747	0.1575	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Ma	0.38728	0.09778	0.1108	0.07186	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Man	0.36229	0.12122	0.13404	0.00313	0.05849	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Nor	0.55793	0.00215	0.00267	0.16572	0.10905	0.13787	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Pas	0.46852	0.03192	0.03346	0.07369	0.04329	0.0592	0.0415	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Po	0.49908	0.00389	0.00246	0.1437	0.09668	0.12168	0.00326	0.03294	0	NA	NA	NA	NA	NA	NA	NA	NA	NA
Ptg	0.4869	0.00236	0.00181	0.13831	0.09455	0.11647	0.00663	0.02667	0.00244	0	NA	NA	NA	NA	NA	NA	NA	NA
Rm	0.58644	0.00489	0.00214	0.13789	0.08775	0.11312	0.00328	0.03073	0.00397	0.004	0	NA	NA	NA	NA	NA	NA	NA
Rus	0.49171	0.00355	0.00382	0.13947	0.08883	0.11708	0.00184	0.03091	0.00157	0.00381	0.00251	0	NA	NA	NA	NA	NA	NA
SC	0.59504	0.00742	0.00314	0.14594	0.09463	0.11406	0.0072	0.0373	0.00418	0.0059	0.00962	0.00665	0	NA	NA	NA	NA	NA
Slo	0.56094	0.00238	0.00062	0.14681	0.0965	0.1183	0.00227	0.03469	0.00078	0.00247	0.00483	0.00234	0.00018	0	NA	NA	NA	NA
Sp	0.54781	0.00195	0.00177	0.16552	0.11223	0.13881	0.00488	0.04159	0.00275	0.00266	0.00828	0.0061	0.00593	0.00193	0	NA	NA	NA
Tur	0.57546	0.01664	0.01718	0.07558	0.05752	0.05717	0.02708	0.02044	0.01666	0.01482	0.02092	0.01968	0.02215	0.02112	0.0197	0	NA	NA
Wel	0.64513	-0.04132	-0.03075	0.1528	0.06847	0.11008	-0.03443	0.00302	-0.0258	-0.023	-0.06219	-0.03369	-0.02013	-0.02357	-0.03037	-0.04753	0	NA
Wo	0.55436	0.19988	0.19783	0.12591	0.10549	0.11671	0.21677	0.11807	0.18393	0.16649	0.17651	0.16913	0.20732	0.19343	0.21335	0.13583	0.2005	0

Table 17 - Y-chr STRs dissimilarity matrix

	Ar	Blg	Bur	Can	cB	CyG	D	Da	E	Est	Far	Fin	Fr	Grk	Heb	Hi	Hu	Ice
Ar	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Blg	0.07432	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Bur	0.1772	0.2279	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Can	0.24028	0.26733	0.25462	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
cB	0.26931	0.28483	0.38915	0.24323	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CyG	0.00488	0.05781	0.1819	0.26117	0.29186	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
D	0.11633	0.08895	0.14916	0.16304	0.08049	0.12257	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Da	0.11917	0.13755	0.18169	0.20164	0.14459	0.12717	0.01757	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
E	0.20694	0.20174	0.28371	0.19923	0.02581	0.22264	0.03566	0.05636	0	NA	NA	NA	NA	NA	NA	NA	NA	NA
Est	0.18715	0.16085	0.18094	0.20641	0.24176	0.18224	0.08761	0.12633	0.16814	0	NA	NA	NA	NA	NA	NA	NA	NA
Far	-0.00189	0.06694	0.16056	0.21568	0.30897	0.00588	0.08862	0.09454	0.19988	0.14658	0	NA	NA	NA	NA	NA	NA	NA
Fin	0.23867	0.28689	0.24577	0.24036	0.28434	0.24203	0.16698	0.17925	0.21436	0.05993	0.21678	0	NA	NA	NA	NA	NA	NA
Fr	0.16015	0.14509	0.23741	0.17373	0.03473	0.16959	0.02258	0.04527	0.00534	0.14191	0.14288	0.1967	0	NA	NA	NA	NA	NA
Grk	0.02681	0.01216	0.16885	0.23799	0.2297	0.01776	0.07563	0.09585	0.1701	1.51E-01	0.02194	0.23852	0.12217	0	NA	NA	NA	NA
Heb	0.01154	0.0836	0.27313	0.30443	0.37293	0.00892	0.12981	0.15272	0.24913	0.2295	0.02506	0.30413	0.1815	0.02882	0	NA	NA	NA
Hi	0.11849	0.05943	0.21312	0.29304	0.3095	0.10967	0.12452	0.16668	0.25169	0.11151	0.08989	0.25553	0.20842	0.07843	0.12969	0	NA	NA
Hu	0.07437	0.01947	0.15787	0.22649	0.20198	0.07217	0.04533	0.0784	0.14378	0.1102	0.05514	0.22929	0.10725	0.02623	0.09519	0.03744	0	NA
Ice	0.11762	0.11885	0.18391	0.20942	0.14247	0.12138	0.00548	0.00287	0.05273	0.08727	0.0918	0.1574	0.0375	0.08382	0.14469	0.13496	0.06348	0
Inu	0.3877	0.39204	0.47846	0.3245	0.25977	0.37833	0.29427	0.34005	0.25719	0.33626	0.39384	0.29219	0.24361	0.35944	0.43188	0.44073	0.37968	0.33979
Ir	0.33087	0.34957	0.43205	0.25248	0.02642	0.3719	0.12793	0.18821	0.04883	0.30129	0.36935	0.3203	0.07202	0.3076	0.41466	0.40912	0.28087	0.1921
It	0.09777	0.09596	0.18255	0.15745	0.07798	0.10498	0.01832	0.02307	3.14E-02	0.13517	0.07797	0.18426	1.27E-02	0.07008	0.10968	0.17388	0.07227	0.02243
Jap	0.17436	0.10228	0.14612	0.13212	0.13307	0.16932	0.08882	0.13988	0.12794	0.14609	0.12964	0.25364	0.10413	0.12536	0.17232	0.12747	0.08702	0.12803
Ma	0.07789	0.10124	0.15192	0.35549	0.35801	0.06641	0.13554	0.14619	0.26908	0.17092	0.06158	0.26793	0.22065	0.0715	0.0901	0.07797	0.07465	0.1307
Man	0.1684	0.16469	0.15575	0.01609	0.13999	0.16993	0.09522	0.09967	0.11172	0.12517	0.10978	0.17256	0.09517	0.15592	0.16974	0.18863	0.14321	0.0981
Nor	0.11969	0.14778	0.18541	0.20109	0.1859	0.13003	0.01469	-7.00E-04	7.44E-02	0.10933	0.0947	0.17575	0.05862	0.09881	0.16686	0.14646	0.07663	-0.00359
Pas	0.14108	0.09984	0.1939	0.28427	0.28593	0.14523	0.09833	0.14384	0.22328	0.08355	0.11905	0.2214	0.19078	0.10461	0.18214	0.01393	0.0425	0.11641
Po	0.19881	0.10059	0.20732	0.2841	0.26231	0.20113	0.10948	0.17702	0.22535	0.12225	0.18056	0.2822	0.1978	0.13114	0.22492	0.04834	0.05809	0.14158
Ptg	0.13244	0.12891	0.20423	0.16368	0.05097	0.14193	0.0209	0.03312	0.01477	0.14085	0.11208	0.18725	0.00214	0.1025	0.14544	0.20246	0.09802	0.02877
Rm	0.06293	0.00082	0.18617	0.26102	0.26059	0.05095	0.08872	0.12239	0.19809	0.15467	0.05188	0.27076	0.1508	0.0139	0.07683	0.04652	0.01192	0.10984
Rus	0.15914	0.0731	0.1814	0.25502	0.24388	0.15408	0.0918	0.15039	0.20156	0.06846	0.13485	0.21395	0.16981	0.09732	0.18291	0.02895	0.04123	0.11317
SC	0.11763	0.0142	0.22274	0.29021	0.28948	0.10902	0.12497	0.17578	0.23847	0.19316	0.11054	0.32167	0.19696	0.05193	0.14324	0.05134	0.02715	0.16305
Slo	0.07805	0.00445	0.20296	0.25796	0.28743	0.07039	0.07247	0.11814	1.99E-01	0.14106	0.06446	0.28044	0.1474	0.02154	0.10274	0.02886	0.00154	0.10073
Sp	0.15621	0.13913	0.23725	0.17664	0.03344	0.16745	0.02732	0.05788	0.0133	0.14271	0.14186	0.19747	0.00106	0.11881	0.17169	0.21202	0.10942	0.04401
Tur	0.03186	0.09814	0.25963	0.23649	0.29946	0.03973	0.10736	0.12412	0.19812	0.15529	0.02198	0.21943	0.14556	0.05835	0.06332	0.08659	0.08008	0.12045
Wel	0.29794	0.31245	0.4153	0.25162	0.03149	0.31305	0.09963	0.13807	0.03508	0.26692	0.33197	0.29793	0.0546	0.25311	0.41675	0.33528	0.22889	0.14566
Wo	0.28936	0.2999	0.36834	0.47868	0.52339	0.26992	0.30484	0.33457	0.41599	0.38296	0.27464	0.46464	0.37247	0.27657	0.32604	0.27416	0.26828	0.3475

Table 17 - Y-chr STRs dissimilarity matrix - continuation

	Inu	Ir	It	Jap	Ma	Man	Nor	Pas	Po	Ptg	Rm	Rus	SC	Slo	Sp	Tur	Wel	Wo
Ar	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Blg	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Bur	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Can	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
cB	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CyG	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
D	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Da	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
E	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Est	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Far	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Fin	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Fr	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Grk	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Heb	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Hi	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Hu	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Ice	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Inu	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Ir	0.24694	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
It	0.25807	0.11345	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Jap	0.30576	0.20045	0.11367	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Ma	0.46685	0.44947	0.16438	0.1764	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Man	0.26622	0.17973	0.09456	0.09236	0.22406	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Nor	0.37849	0.22172	0.03356	0.13656	0.14135	0.08442	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Pas	0.4521	0.368	0.15985	0.13466	0.09502	0.19021	0.1278	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Po	0.45813	0.32963	0.1791	0.1219	0.16426	0.21865	0.15544	0.03811	0	NA	NA	NA	NA	NA	NA	NA	NA	NA
Ptg	0.24443	0.08762	0.00461	0.11148	0.19628	0.09491	0.04406	0.18386	0.19742	0	NA	NA	NA	NA	NA	NA	NA	NA
Rm	0.39622	0.34227	0.10154	0.11036	0.07455	0.17384	0.12745	0.07596	0.09448	0.13621	0	NA	NA	NA	NA	NA	NA	NA
Rus	0.40897	0.32171	0.15099	0.11573	0.12869	0.18529	0.13303	0.01964	0.01311	0.16968	0.06931	0	NA	NA	NA	NA	NA	NA
SC	0.44199	0.3499	0.14981	0.12843	0.12455	0.23092	0.18155	0.08238	0.07803	0.18544	0.01062	0.06722	0	NA	NA	NA	NA	NA
Slo	0.42527	0.35526	0.09884	0.08916	0.08434	0.14965	0.11807	0.05501	0.05553	0.13182	9.00E-04	0.04295	0.00614	0	NA	NA	NA	NA
Sp	0.23631	0.07412	0.01522	0.10821	0.22568	0.10615	0.0684	0.19374	0.19559	0.00444	0.1511	0.1685	0.19465	0.14544	0	NA	NA	NA
Tur	0.36193	0.35901	0.09561	0.13498	0.09878	0.12116	0.12986	0.12818	0.2036	0.12194	0.08273	0.15318	0.14346	0.09371	0.14297	0	NA	NA
Wel	0.27681	0.01009	0.09097	0.18088	0.38064	0.1487	0.18297	0.31899	0.29844	0.06835	0.28385	0.27854	0.31933	0.3138	0.05995	0.34263	0	NA
Wo	0.55883	0.56894	0.32493	0.25248	0.19125	0.33758	0.37224	0.31411	0.37478	0.35031	0.25787	0.34178	0.30254	0.29945	0.38561	0.29152	0.55294	0

Table 18 - mtDNA SNPs dissimilarity matrix

	Ar	Blg	Cal	cB	D	Da	Fin	Fr	Grk	Hi	Hu	Ir	It	Nor	Pas	Po	Ptg	Rm	Rus	Sal	SC	Sic	Slo	Sp	Tur	wB
Ar	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Blg	0.01443	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Cal	0.00116	0.0015	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
cB	0.04286	0.01045	0.0265	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
D	0.02662	0.0021	0.00976	0.00877	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Da	0.01674	0.00376	0.00284	0.01878	0.00267	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Fin	0.0174	0.01071	0.00927	0.02244	0.02208	0.02194	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Fr	0.01686	0.00153	0.00537	0.01217	0.00223	0.00271	0.01764	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Grk	0.02225	0.01249	0.01468	0.02141	0.0195	0.02032	0.01231	0.0178	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Hi	0.206	0.2676	0.25201	0.32955	0.29788	0.28014	0.24668	0.28042	0.24105	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Hu	0.06046	0.11869	0.08242	0.18615	0.14765	0.12444	0.1149	0.12814	0.1206	0.12508	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Ir	0.03712	0.10016	0.06093	0.16973	0.12874	0.09927	0.0927	0.10212	0.10328	0.13269	-0.0314	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
It	0.01376	0.00019	0.00214	0.01373	0.00219	0.0026	0.01606	0.00053	0.01575	0.26554	0.11733	0.0966	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Nor	0.02683	0.00203	0.00957	0.00667	-0.00076	0.00241	0.0202	0.00285	0.01766	3.01E-01	0.14932	0.13091	0.00296	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Pas	0.07364	0.11633	0.08896	0.15906	0.14638	0.12836	0.09998	0.12791	0.10378	0.07093	0.01752	0.01743	0.11987	0.14427	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Po	0.02055	-0.00027	0.00517	0.00875	0.00045	0.0042	0.01442	0.00348	0.01375	0.2806	0.13118	0.11638	0.00102	0.00039	0.12454	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Ptg	0.00576	0.01251	0.00488	0.03656	0.021	0.0154	0.02517	0.01443	0.02303	0.24103	0.08689	0.06798	0.01066	0.02174	0.09442	0.01668	0	NA	NA	NA	NA	NA	NA	NA	NA	NA
Rm	0.00403	-5.00E-05	-0.00593	0.02074	0.01001	0.00776	0.00547	0.00708	0.01243	0.26886	0.08953	0.07399	0.00294	0.00921	0.08909	0.00357	0.00528	0	NA	NA	NA	NA	NA	NA	NA	NA
Rus	0.01356	0.00017	-8.00E-05	0.01265	0.00471	0.00433	0.00728	0.00521	0.01039	0.26144	0.11145	0.09488	0.00302	0.00343	0.10805	0.00107	0.01408	-0.00221	0	NA	NA	NA	NA	NA	NA	NA
Sal	0.00841	0.02736	0.00596	0.07528	0.04362	0.02921	0.02488	0.03736	0.0384	0.2418	0.06914	0.04757	0.02875	0.04419	0.08517	0.03274	0.02196	0.00835	0.02166	0	NA	NA	NA	NA	NA	NA
SC	0.02197	0.00014	0.00599	0.00463	0.00125	0.00483	0.01156	0.00333	1.23E-02	0.28701	0.13699	0.12018	2.67E-03	3.00E-05	0.12832	-0.00036	0.01901	0.00303	0.00028	0.03746	0	NA	NA	NA	NA	NA
Sic	0.00717	0.00294	-0.00022	0.02307	0.00907	0.00617	0.01415	0.00378	0.01841	0.25081	0.10303	0.07822	0.00169	0.01039	0.10572	0.00546	0.00862	0.00175	0.00573	0.01822	0.0082	0	NA	NA	NA	NA
Slo	0.02314	-0.00108	0.00623	-0.00117	-0.00105	0.00285	0.00768	0.00198	0.00902	0.30784	0.14655	0.1274	0.00201	-0.00294	0.12987	-0.00228	0.0199	0.00393	-0.0015	0.03913	-0.00458	0.00814	0	NA	NA	NA
Sp	0.03666	0.00802	0.02068	0.00484	0.00267	0.00745	0.02818	0.0057	0.02341	0.31801	0.17539	0.15324	0.00745	0.00221	0.16366	0.0054	0.02963	0.0226	0.01213	0.06544	0.00505	0.01578	-0.00014	0	NA	NA
Tur	0.03083	0.06646	0.04364	0.10175	0.09946	0.0874	0.04103	7.64E-02	4.89E-02	0.19811	0.0569	0.03476	0.07152	0.09806	0.0308	0.08137	0.05292	0.03647	0.06082	0.04744	0.07722	0.05552	0.07998	0.11497	0	NA
wB	0.0272	0.00093	0.01225	0.00051	-0.00012	0.00736	0.01605	0.00127	0.01511	0.30962	0.15344	0.13467	0.00207	-0.00024	0.13686	-4.00E-04	0.02125	0.00977	0.00471	0.05077	-0.00076	0.00844	-0.00404	0.00012	0.08793	0

Table 19 - Y-chr SNPs dissimilarity matrix

	Ar	Blg	Cal	cB	D	Da	Fin	Fr	Grk	Hi	Hu	Ir	It	Nor	Pas	Po	Ptg	Rm	Rus	Sal	SC	Sic	Slo	Sp	Tur	wB
Ar	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Blg	0.13015	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Cal	0.06397	0.04809	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
cB	0.40104	0.31785	0.35303	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
D	0.22374	0.08785	0.07946	0.14632	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Da	0.26276	0.08758	0.12646	0.28188	0.02818	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Fin	0.2991	0.21635	0.29839	0.52198	0.29097	0.28609	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Fr	0.26494	0.18202	0.12495	0.06365	0.03942	0.12105	0.38256	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Grk	0.07114	0.02327	0.00708	0.38588	0.08094	0.12788	0.26941	0.15307	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Hi	0.16328	0.12197	0.1114	0.40058	0.19807	0.2288	0.26301	0.27264	0.10698	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Hu	0.17033	0.02212	0.06478	0.31411	0.06347	0.06519	0.23777	0.1723	0.04539	0.10671	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Ir	0.44361	0.37722	0.43716	-0.00192	0.19371	0.33787	0.57848	0.08701	0.47043	0.47004	0.39303	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
It	0.14982	0.09755	0.02648	0.14258	0.03086	0.08938	0.28421	0.03056	0.05182	0.17496	0.09332	0.18033	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Nor	0.24915	0.06234	0.11643	0.35193	0.03522	0.02125	0.25272	0.15465	0.09504	1.59E-01	0.02314	0.41107	0.10295	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Pas	0.28126	0.17618	0.24025	0.54228	0.26996	0.32897	0.34839	0.37635	0.19002	0.10379	0.13663	0.61174	0.26856	0.21566	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Po	0.29885	0.13416	0.23214	0.46498	0.19332	0.23765	0.33993	0.31583	0.16718	0.16919	0.08657	0.53106	0.23383	0.12428	0.0578	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Ptg	0.23513	0.16839	0.09597	0.08299	0.04189	0.12364	0.36476	0.00168	0.12606	0.25662	0.16341	0.10981	0.01821	0.15388	0.35746	0.3055	0	NA	NA	NA	NA	NA	NA	NA	NA	NA
Rm	0.14602	1.56E-03	0.05892	0.38115	0.08636	0.07167	0.23956	0.2039	0.0367	0.1318	0.01353	0.47168	0.10644	0.04771	0.1991	0.14534	0.18826	0	NA	NA	NA	NA	NA	NA	NA	NA
Rus	0.24959	0.11716	1.98E-01	0.4164	0.19126	0.21282	0.17284	0.30034	0.13656	0.13821	0.08849	0.45876	0.22189	0.11589	0.06636	0.05021	0.29174	0.12364	0	NA	NA	NA	NA	NA	NA	NA
Sal	0.04285	0.04687	-0.01358	0.40565	0.07881	0.13585	0.31048	0.12735	-0.00482	0.12215	0.06561	0.46642	0.02526	0.12065	0.24794	0.23119	0.09569	0.0578	0.19396	0	NA	NA	NA	NA	NA	NA
SC	0.19252	0.01146	0.1003	0.40044	0.11007	0.08345	0.22426	0.23968	6.82E-02	0.14495	0.01937	0.48619	1.45E-01	4.85E-02	0.19	0.13012	0.22736	0.00347	0.10592	0.10359	0	NA	NA	NA	NA	NA
Sic	0.06186	0.06494	-0.00295	0.25407	0.07788	0.12951	0.27965	0.10507	0.01179	0.13155	0.08227	0.32685	0.02523	0.12594	0.23775	0.22504	0.08099	0.07685	0.20312	-0.01216	0.11986	0	NA	NA	NA	NA
Slo	0.24266	0.05392	0.12734	0.43874	0.08103	0.08629	0.287	0.21873	0.08769	0.14516	0.01027	0.50231	0.1399	0.01537	0.13407	0.04315	0.21056	0.04283	0.06884	0.12788	0.03707	0.13886	0	NA	NA	NA
Sp	0.26197	0.19493	0.1294	0.06702	0.05447	0.14745	0.40043	2.00E-04	0.1605	0.28061	0.18993	0.09305	0.03291	0.18153	0.39104	0.33453	0.00125	0.22192	0.31542	0.13035	0.26018	0.10481	0.24362	0	NA	NA
Tur	0.0264	0.07011	0.01037	0.32019	0.12161	0.16057	0.24011	1.71E-01	2.05E-02	0.09598	0.08455	0.40001	0.06636	0.14153	0.20489	0.21765	0.14528	0.07912	0.17578	0.001	0.11804	0.01494	0.14365	0.17229	0	NA
wB	0.40138	0.32112	0.3126	0.01175	0.15681	0.29623	0.52162	0.06664	0.34415	0.38724	0.30927	0.01724	0.1432	0.34863	0.5182	4.56E-01	0.08211	0.36284	0.42202	0.37773	0.38924	0.24039	0.42221	0.06413	0.30513	0

2.2. Geographic matrices

Table 20 - Geographical distance matrix for the fast-evolving dataset

	Ar	**Blg**	**Bur**	**Can**	**cB**	**CyG**	**D**	**Da**	**E**	**Est**	**Far**	**Fin**	**Fr**	**Grk**	**Heb**	**Hi**	**Hu**	**Ice**
Ar	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Blg	2931534	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Bur	5899291	6084552	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Can	6818371	8430664	3348222	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
cB	4940066	2130431	7525413	10325193	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CyG	1733125	1206293	5993065	7813361	3229848	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
D	4171087	1319867	5942940	8768546	1589927	2492414	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Da	4422037	1638745	5763604	8687916	1770929	2778923	355732.7	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
E	4947923	2016891	6692070	9644110	941204.8	3221757	932799.6	956627.3	0	NA	NA	NA	NA	NA	NA	NA	NA	NA
Est	4226947	1865997	4927975	7872783	2599859	2773768	1042155	837314.1	1785283	0	NA	NA	NA	NA	NA	NA	NA	NA
Far	1301643	2532790	4724740	6189037	4662333	1636050	3510495	3676371	4403799	3271187	0	NA	NA	NA	NA	NA	NA	NA
Fin	4289431	1948194	4879950	7846634	2652272	2849010	1106586	883842.2	1822924	82371.22	3318573	0	NA	NA	NA	NA	NA	NA
Fr	4685761	1759363	6788053	9646913	743824	2953376	878587.5	1027849	343897.3	1860218	4215509	1910628	0	NA	NA	NA	NA	NA
Grk	2626695	527693.3	6413771	8558354	2317028	916806.5	1806051	2140309	2395015	2.39E+06	2472193	2473258	2098789	0	NA	NA	NA	NA
Heb	1374453	1605101	6126409	7736861	3565613	413931.2	2902848	3192255	3612843	3175921	1559179	3249955	3334564	1254187	0	NA	NA	NA
Hi	3054432	5028181	3606933	3767051	7129602	4173486	5787404	5854897	6719240	5212414	2546173	5227074	6594320	5015969	4030368	0	NA	NA
Hu	3515736	628489.8	5994001	8585718	1775131	1813134	691380.6	1016408	1452747	1383190	2968845	1464198	1246826	1124297	2220190	5370606	0	NA
Ice	6527817	3705916	6411903	9707287	2615307	4880166	2390167	2108447	1890624	2452068	5701226	2418662	2234375	4167698	5291858	7601857	3078871	0
Inu	9472267	7877208	4606660	7151133	7811724	8594763	6832708	6482485	6917543	6014205	8176097	5931834	7224654	8398164	8937230	8091229	7375087	5310335
Ir	5405363	2476189	6814269	9870429	1148603	3682065	1318679	1239932	463732	2005161	4825126	2025898	781573.3	2858683	4074980	7086294	1898807	1495132
It	3679426	896507.3	6796675	9296975	1271008	1960301	1184873	1534335	1.44E+06	2128147	3417837	2205016	1.11E+06	1052766	2306668	5924210	809265.9	3304578
Jap	8697398	9184001	3104298	2894923	10466425	9062396	8926071	8699279	9569427	7886201	7668838	7826837	9722954	9517785	9156955	5845371	9057614	8808143
Ma	2767116	5320726	4747403	4307103	7448576	4270738	6299168	6433732	7205463	5905594	2806796	5935284	7022304	5179468	4018312	1161139	5775602	8353089
Man	6602847	7360154	1489924	1984353	8954241	7072728	7364712	7208667	8149258	6371377	5607563	6328392	8225228	7625765	7127518	3784333	7348642	7892152
Nor	4799366	2102056	5558934	8607302	2074767	3204511	843271.3	4.88E+05	1.16E+06	786767.1	3959472	787095.4	1346996	2613350	3618322	5993664	1490102	1746458
Pas	2584236	4227910	3508135	4381981	6312924	3447290	4953911	5021043	5886046	4387468	1812757	4404578	5765211	4251331	3354902	834247.4	4547449	6793648
Po	3753925	1075459	5555825	8291208	2037473	2137500	517254.6	672046.6	1449838	835045.1	3017713	916188	1367884	1601942	2551304	5270188	548186.4	2773929
Ptg	5427804	2757742	8243341	11045273	726784	3768639	2315518	2480986	1587621	3315401	5280950	3364742	1455324	2854487	4065208	7784845	2472810	2952771
Rm	2880463	295671.1	5788906	8162212	2321794	1203516	1295113	1575479	2092778	1672462	2354453	1753350	1871542	747087.8	1616779	4810752	641185.4	3677970
Rus	3534252	1779017	4423995	7158125	3184454	2314364	1610509	1562273	2503397	868658.1	2467569	893695.7	2489051	2235283	2674730	4347930	1572116	3311526
SC	3611320	680348.3	6293907	8871004	1522288	1882719	770625.6	1124818	1339243	1626628	3163770	1705354	1080378	1081882	2274406	5607819	300282.4	3087701
Slo	3723733	794154.6	6357990	8965308	1412375	1993683	723862.4	1079481	1.23E+06	1637275	3279494	1714105	965397.1	1177725	2382801	5719275	380466.8	2999600
Sp	4966769	2253760	7811636	10555286	321473.9	3286518	1868783	2072499	1262668	2893870	4780295	2949221	1051669	2370000	3597026	7279819	1973897	2891909
Tur	2134789	852290.1	5676091	7737965	2977989	534763.7	2036603	2298139	2834461	2242214	1696959	2316693	2598130	820411	934079.8	4223356	1387031	4406506
Wel	5149274	2217751	6835626	9822314	914867.7	3420574	1140342	1137559	211786.4	1952693	4615402	1984993	490966.3	2581890	3808420	6927723	1661723	1780051
Wo	6761457	4980668	10940451	13349537	3469841	5542431	5015287	5235247	4381569	6051466	7175666	6109236	4212527	4799014	5653759	9683846	4961277	5515581

Table 20 - Geographical distances matrix for the fast-evolving dataset - continuation

Inu	Ir	It	Jap	Ma	Man	Nor	Pas	Po	Ptg	Rm	Rus	SC	Slo	Sp	Tur	Wel	Wo
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
6668774	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
8016691	1887774	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
4685719	9595458	9866001	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
9252327	7617702	6186455	6742541	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
5315377	8290802	8134716	2096116	4760554	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
6007966	1268380	2013594	8411409	6656462	7029311	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
7747650	6252049	5124248	6122701	1689808	4026590	5166155	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
6836044	1828648	1318025	8588169	5794325	6948106	1067018	4436662	0	NA	NA	NA	NA	NA	NA	NA	NA	NA
8252133	1643821	1864192	11156129	8032555	9678094	2745301	6978404	2762638	0	NA	NA	NA	NA	NA	NA	NA	NA
7660895	2539985	1139505	8888372	5158445	7067971	2008746	4000278	945494.2	2978334	0	NA	NA	NA	NA	NA	NA	NA
6283143	2797722	2379346	7486826	5041799	5799336	1645736	3520870	1152016	3910891	1500295	0	NA	NA	NA	NA	NA	NA
7573839	1800663	516971.9	9355096	5966585	7648361	1612466	4790679	804640.9	2202478	809477.3	1871202	0	NA	NA	NA	NA	NA
7549622	1692597	490912.9	9410055	6082699	7722523	1566449	4900874	834033	2099063	9.25E+05	1934061	116445.2	0	NA	NA	NA	NA
8119242	1450679	1362459	10771723	7541499	9229263	2391683	6472781	2289777	505666.8	2472670	3441606	1699411	1597741	0	NA	NA	NA
8076459	3285038	1722473	8771973	4470720	6839344	2703410	3444181	1639787	3584266	747303.1	1793353	1519242	1635325	3084526	0	NA	NA
6899978	294409	1597336	9679349	7417243	8300377	1280125	6094253	1657592	1495299	2300841	2694919	1538204	1426954	1230681	3040793	0	NA
10755551	4414988	4175164	13934525	9520932	12308795	5536579	8988407	5384637	2796620	5265864	6521501	4661238	4589828	3162844	5618546	4290464	0

Table 21 - Geographical distance matrix for the slow-evolving dataset

	Ar	Blg	Cal	cB	D	Da	Fin	Fr	Grk	Hi	Hu	Ir	It	Nor	Pas	Po	Ptg	Rm	Rus	Sal	SC	Sic	Slo	Sp	Tur	wB
Ar	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Blg	1444081	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Cal	1780656	707444.6	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
cB	3393803	2054205	1629280	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
D	2720480	1320930	1534675	1531276	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Da	2998191	1638816	1890105	1724249	355653.3	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Fin	3025567	1947907	2437913	2603379	1107743	883899.4	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Fr	3194837	1759965	1585968	700971.3	876996.5	1027840	1910678	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Grk	1154256	527880	630233.7	2241384	1806731	2140282	2473240	2098864	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Hi	3985365	5027298	5622965	7053772	5788583	5854286	5226214	6593982	5016030	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Hu	2046538	630565.4	975994.1	1702446	690364.7	1014282	1461928	1246412	1126433	5370290	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Ir	3920719	2476639	2366174	1159500	1317072	1239919	2025929	781527	2858703	7085720	1897722	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
It	2199885	895748.8	480376.4	1195764	1183740	1533545	2203861	1107045	1051716	5922587	809555.6	1888160	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Nor	3415646	2102268	2376776	2037795	842105	486845.8	791084.7	1343594	2613291	6.00E+06	1487761	1264190	2011284	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Pas	3085829	4037349	4652086	6052493	4783224	4856227	4257116	5586802	4057736	1007228	4364885	6083914	4933054	5016177	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Po	2348129	1075276	1521999	1973723	519521.8	672469.8	915999.3	1368522	1601898	5269102	546133.2	1829209	1317080	1068294	4263884	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Ptg	4001202	2758458	2224239	789452.3	2313957	2480931	3364743	1455271	2854468	7784662	2473346	1643826	1865626	2741439	6790337	2763174	0	NA	NA	NA	NA	NA	NA	NA	NA	NA
Rm	1425096	2.95E+05	1000922	2246297	1296949	1576062	1753675	1872303	746947.3	4809912	642963.6	2540686	1138475	2010058	3812161	945809.4	2978921	0	NA	NA	NA	NA	NA	NA	NA	NA
Rus	2439678	1778224	2.44E+06	3121785	1612088	1562280	893757.5	2489032	2235055	4347107	1570356	2797724	2377692	1649352	3369342	1151369	3910778	1500066	0	NA	NA	NA	NA	NA	NA	NA
Sal	1694647	503381	210088	1699587	1402301	1754692	2255422	1565831	547399.4	5469964	798056.6	2336325	505509.8	2240858	4491278	1340329	2342318	794375	2227497	0	NA	NA	NA	NA	NA	NA
SC	2122439	680609	770485.6	1448559	769339.9	1123196	1703201	1080873	1.08E+06	5606226	299563.5	1800609	5.18E+05	1.61E+06	4604127	802740.7	2204072	809137.7	1868846	633720	0	NA	NA	NA	NA	NA
Sic	2043410	986357.3	294633.4	1415794	1603113	1955668	2587030	1486830	907604.6	5914223	1141824	2265113	427575.3	2436035	4945434	1680148	1958695	1274852	2665776	483443.2	884175.5	0	NA	NA	NA	NA
Slo	2234082	794714.3	813527.5	1337706	723834.4	1079420	1714097	965430.1	1177766	5719035	381432.5	1692573	489764	1565244	4716231	834419	2099011	925767.2	1933926	700820.8	117247.6	888913.7	0	NA	NA	NA
Sp	3524901	2258752	1744298	357245	1873523	2078757	2955516	1.06E+06	2373098	7284082	1979894	1455933	1367885	2394370	6289372	2296258	501009.8	2477919	3447454	1851493	1706187	1491247	1602952	0	NA	NA
Tur	712888.6	853549.9	1401966	2902708	2040440	2300798	2319396	2.60E+06	8.21E+05	4222273	1390636	3287430	1722577	2707390	3250512	1642177	3585661	749095.5	1795527	1249640	1520799	1694954	1637476	3089894	0	NA
wB	3468931	2130521	1702683	76348.7	1588669	1771388	2652750	744340	2316181	7128931	1775204	1149834	1271456	2071882	6127372	2.04E+06	726589.4	2321962	3184494	1774948	1523279	1485828	1411984	326621.3	2979053	0

2.3. Linguistic matrices

Table 22 - Linguistic distance matrix for the fast-evolving dataset

	Ar	Blg	Bur	Can	cB	CyG	D	Da	E	Est	Far	Fin	Fr	Grk	Heb	Hi	Hu	Ice
Ar	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Blg	0.239	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Bur	0.471	0.316	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Can	0.522	0.333	0.241	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
cB	0.324	0.257	0.219	0.231	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CyG	0.239	0.22	0.333	0.37	0.351	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
D	0.31	0.167	0.289	0.296	0.257	0.184	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Da	0.273	0.14	0.289	0.214	0.2	0.188	0.0816	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
E	0.31	0.208	0.231	0.25	0.229	0.188	0.0612	0.0816	0	NA	NA	NA	NA	NA	NA	NA	NA	NA
Est	0.333	0.22	0.146	0.259	0.125	0.262	0.22	0.171	0.214	0	NA	NA	NA	NA	NA	NA	NA	NA
Far	0.389	0.333	0.189	0.259	0.281	0.324	0.27	0.297	0.216	0.194	0	NA	NA	NA	NA	NA	NA	NA
Fin	0.306	0.225	0.184	0.32	0.167	0.25	0.2	0.2	0.22	0.0244	0.206	0	NA	NA	NA	NA	NA	NA
Fr	0.31	0.17	0.389	0.36	0.222	0.213	0.133	0.133	0.182	0.237	0.361	0.243	0	NA	NA	NA	NA	NA
Grk	0.239	0.2	0.333	0.37	0.378	0.0185	0.204	0.208	0.208	2.86E-01	0.324	0.275	0.234	0	NA	NA	NA	NA
Heb	0.128	0.205	0.406	0.571	0.312	0.295	0.25	0.214	0.25	0.324	0.382	0.294	0.3	0.295	0	NA	NA	NA
Hi	0.278	0.175	0.256	0.25	0.206	0.225	0.225	0.171	0.244	0.175	0.262	0.211	0.184	0.225	0.324	0	NA	NA
Hu	0.359	0.233	0.184	0.24	0.242	0.233	0.238	0.233	0.233	0.105	0.222	0.0952	0.238	0.256	0.351	0.2	0	NA
Ice	0.25	0.12	0.289	0.296	0.229	0.2	0.0962	0.08	0.122	0.171	0.243	0.175	0.156	0.22	0.19	0.15	0.238	0
Inu	0.323	0.258	0.206	0.429	0.259	0.267	0.31	0.3	0.333	0.226	0.323	0.206	0.31	0.267	0.31	0.206	0.229	0.276
Ir	0.262	0.205	0.333	0.435	0.312	0.227	0.159	0.205	0.14	0.257	0.257	0.265	0.22	0.227	0.214	0.25	0.297	0.133
It	0.273	0.143	0.361	0.36	0.222	0.163	0.106	0.106	1.52E-01	0.231	0.333	0.237	4.17E-02	0.184	0.262	0.158	0.214	0.17
Jap	0.348	0.24	0.375	0.333	0.2	0.375	0.375	0.32	0.36	0.304	0.36	0.364	0.304	0.375	0.476	0.259	0.381	0.292
Ma	0.278	0.175	0.256	0.25	0.206	0.225	0.225	0.171	0.244	0.175	0.262	0.211	0.184	0.225	0.324	0	0.2	0.15
Man	0.522	0.333	0.241	0.0286	0.231	0.37	0.296	0.214	0.25	0.259	0.259	0.32	0.36	0.37	0.571	0.25	0.24	0.296
Nor	0.273	0.137	0.289	0.214	0.2	0.167	0.0816	3.85E-02	8.16E-02	0.171	0.297	0.2	0.156	0.188	0.214	0.171	0.233	0.0588
Pas	0.333	0.175	0.15	0.258	0.152	0.275	0.175	0.171	0.146	0.15	0.184	0.184	0.243	0.275	0.324	0.116	0.167	0.15
Po	0.211	0.0909	0.308	0.321	0.273	0.156	0.149	0.163	0.205	0.214	0.237	0.195	0.195	0.133	0.222	0.195	0.263	0.13
Ptg	0.295	0.163	0.361	0.36	0.194	0.204	0.106	0.106	0.152	0.231	0.333	0.237	0.0417	0.224	0.286	0.158	0.238	0.149
Rm	0.234	0.118	0.361	0.36	0.229	0.163	0.174	0.104	0.174	0.231	0.333	0.282	0.109	0.184	0.244	0.158	0.286	0.163
Rus	0.211	0.0682	0.308	0.321	0.273	0.156	0.17	0.163	0.205	0.214	0.237	0.195	0.22	0.133	0.222	0.195	0.263	0.13
SC	0.211	0.0909	0.308	0.321	0.242	0.133	0.128	0.14	0.182	0.19	0.237	0.171	0.195	0.156	0.222	0.195	0.237	0.087
Slo	0.211	0.0682	0.308	0.321	0.273	0.156	0.149	0.163	2.05E-01	0.214	0.237	0.195	0.22	0.133	0.222	0.195	0.263	0.109
Sp	0.267	0.14	0.389	0.4	0.222	0.2	0.128	0.128	0.174	0.256	0.361	0.263	0.0625	0.22	0.279	0.184	0.262	0.17
Tur	0.424	0.278	0.075	0.231	0.207	0.343	0.229	0.278	0.222	0.162	0.167	0.125	0.353	0.343	0.355	0.214	0.0952	0.257
Wel	0.268	0.209	0.312	0.435	0.312	0.233	0.163	0.209	0.143	0.235	0.235	0.242	0.225	0.233	0.22	0.257	0.278	0.136
Wo	0.387	0.375	0.393	0.304	0.333	0.265	0.286	0.303	0.294	0.37	0.281	0.385	0.312	0.265	0.448	0.258	0.344	0.229

Table 22 - Linguistic distance matrix for the fast-evolving dataset - continuation

	Inu	Ir	It	Jap	Ma	Man	Nor	Pas	Po	Ptg	Rm	Rus	SC	Slo	Sp	Tur	Wel	Wo
Ar	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Blg	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Bur	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Can	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
cB	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CyG	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
D	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Da	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
E	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Est	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Far	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Fin	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Fr	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Grk	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Heb	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Hi	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Hu	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Ice	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Inu	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Ir	0.357	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
It	0.276	0.233	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Jap	0.348	0.304	0.304	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Ma	0.206	0.25	0.158	0.259	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Man	0.429	0.435	0.36	0.333	0.25	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Nor	0.3	0.205	0.128	0.32	0.171	0.214	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Pas	0.258	0.194	0.211	0.222	0.116	0.258	0.171	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Po	0.226	0.154	0.167	0.308	0.195	0.321	0.163	0.195	0	NA	NA	NA	NA	NA	NA	NA	NA	NA
Ptg	0.276	0.209	0.04	0.304	0.158	0.36	0.128	0.211	0.167	0	NA	NA	NA	NA	NA	NA	NA	NA
Rm	0.29	0.227	0.0625	0.208	0.158	0.36	0.122	0.211	0.167	0.0833	0	NA	NA	NA	NA	NA	NA	NA
Rus	0.226	0.154	0.19	0.308	0.195	0.321	0.163	0.195	0.0204	0.19	0.167	0	NA	NA	NA	NA	NA	NA
SC	0.226	0.154	0.167	0.308	0.195	0.321	0.14	0.195	0.0612	0.167	0.143	0.0408	0	NA	NA	NA	NA	NA
Slo	0.226	0.154	0.19	0.308	0.195	0.321	0.163	0.195	0.0408	0.19	1.67E-01	0.0204	0.0204	0	NA	NA	NA	NA
Sp	0.267	0.205	0.06	0.25	0.184	0.4	0.149	0.237	0.14	0.02	0.0612	0.163	0.14	0.163	0	NA	NA	NA
Tur	0.189	0.323	0.324	0.364	0.214	0.231	0.278	0.135	0.25	0.324	0.371	0.25	0.25	0.25	0.353	0	NA	NA
Wel	0.37	0	0.238	0.304	0.257	0.435	0.209	0.171	0.158	0.214	0.233	0.158	0.158	0.158	0.209	0.3	0	NA
Wo	0.4	0.323	0.312	0.391	0.258	0.304	0.303	0.357	0.258	0.312	0.344	0.258	0.258	0.258	0.344	0.385	0.323	0

Table 23 - Linguistic distance matrix for the slow-evolving dataset

	Ar	Blg	Cal	cB	D	Da	Fin	Fr	Grk	Hi	Hu	Ir	It	Nor	Pas	Po	Ptg	Rm	Rus	Sal	SC	Sic	Slo	Sp	Tur	wB
Ar	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Blg	0.268	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Cal	0.238	0.14	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
cB	0.4	0.258	0.333	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
D	0.308	0.149	0.0952	0.333	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Da	0.308	0.125	0.146	0.233	0.0833	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Fin	0.44	0.212	0.308	0.238	0.226	0.219	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Fr	0.3	0.163	0.093	0.29	0.143	0.146	0.296	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Grk	0.2	0.205	0.167	0.433	0.227	0.209	0.333	0.238	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Hi	0.312	0.154	0.212	0.25	0.216	0.158	0.226	0.212	0.206	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Hu	0.375	0.231	0.273	0.36	0.243	0.263	0.121	0.294	0.27	0.219	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Ir	0.275	0.195	0.146	0.357	0.122	0.171	0.346	0.179	0.231	0.273	0.323	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
It	0.262	0.133	0.0444	0.29	0.114	0.116	0.286	0.0444	0.182	0.176	0.257	0.195	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Nor	0.308	0.122	0.146	0.233	0.0833	0.0392	0.219	0.146	0.209	1.58E-01	0.263	0.171	0.116	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Pas	0.387	0.179	0.281	0.207	0.162	0.158	0.188	0.273	0.314	0.128	0.188	0.219	0.235	0.158	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Po	0.29	0.103	0.121	0.296	0.128	0.135	0.219	0.176	0.162	0.167	0.281	0.188	0.143	0.135	0.216	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Ptg	0.293	0.136	0.0682	0.258	0.114	0.116	0.286	0.0455	0.209	0.176	0.286	0.175	0.0217	0.116	0.235	0.143	0	NA	NA	NA	NA	NA	NA	NA	NA	NA
Rm	0.289	2.14E-01	0.15	0.357	0.225	0.15	0.321	0.175	0.19	0.226	0.324	0.237	0.119	0.146	0.312	0.242	0.122	0	NA	NA	NA	NA	NA	NA	NA	NA
Rus	0.29	0.0769	1.52E-01	0.296	0.154	0.135	0.219	0.206	0.162	0.167	0.281	0.188	0.171	0.135	0.216	0.025	0.171	0.242	0	NA	NA	NA	NA	NA	NA	NA
Sal	0.262	0.116	0.0222	0.3	0.119	0.122	0.308	0.0698	0.167	0.182	0.273	0.171	0.0222	0.122	0.25	0.121	0.0455	0.15	0.152	0	NA	NA	NA	NA	NA	NA
SC	0.29	0.0769	0.152	0.259	0.103	0.0811	0.156	0.176	1.89E-01	0.139	0.219	0.188	1.43E-01	8.11E-02	0.162	0.075	0.143	0.212	0.05	0.152	0	NA	NA	NA	NA	NA
Sic	0.262	0.116	0.0222	0.3	0.119	0.122	0.308	0.0698	0.167	0.182	0.273	0.171	0.0222	0.122	0.25	0.121	0.0455	0.15	0.152	0	0.152	0	NA	NA	NA	NA
Slo	0.29	0.0513	0.152	0.296	0.128	0.108	0.188	0.206	0.162	0.139	0.25	0.188	0.171	0.108	0.189	0.05	0.171	0.242	0.025	0.152	0.025	0.152	0	NA	NA	NA
Sp	0.286	0.178	0.111	0.29	0.136	0.14	0.321	6.67E-02	0.205	0.206	0.314	0.171	0.0638	0.14	0.265	0.171	0.0217	0.119	0.2	0.0889	0.171	0.0889	0.2	0	NA	NA
Tur	0.583	0.29	0.44	0.333	0.276	0.333	0.194	4.80E-01	4.29E-01	0.242	0.125	0.4	0.423	0.333	0.156	0.3	0.423	0.462	0.3	0.44	0.267	0.44	0.267	0.462	0	NA
wB	0.393	0.31	0.357	0.0968	0.357	0.25	0.35	0.31	0.393	0.259	0.478	0.321	0.31	0.25	0.214	3.08E-01	0.276	0.346	0.308	0.321	0.269	0.321	0.308	0.276	0.45	0