UNIVERSITA' DEGLI STUDI DI PARMA

DOTTORATO DI RICERCA IN
"BIOLOGIA EVOLUZIONISTICA ED ECOLOGIA"

CICLO XXX

Inferring past human migrations with genes, languages and bacteria

Coordinatore:
Chiar.mo Prof. Pierluigi Viaroli

Tutore:
Chiar.mo Prof. Guido Barbujani
Dott.ssa Silvia Ghirotto

Dottorando: Andrea Brunelli

Anni 2014/2017

Home is behind, the world ahead
And there are many paths to tread

- J.R.R Tolkien, The Fellowship of the Ring

ABSTRACT - ENGLISH

In this PhD thesis I outline the work that I did over three years, which has led to the publication of three papers in peer-reviewed journals. All of these studies focus on inferring patterns of human migrations from genetic data, but each one differ regarding to the statistical methodologies employed and to the specific case study investigated.

From the first exit out of Africa, the history of our species has always been characterized by large movements of populations that, over thousands of years, led to the colonization of the entire Planet. Even if these migrations have been the object of extensive anthropological and archaeological studies, population genetics has allowed for a considerable advancement of knowledge in this area. Nowadays we can jointly analyse information coming from different sources, in order to identify relationships between ancient and living populations in a multidisciplinary framework. In this light, the study of biological proxies, meaning species that have coevolved with humans, can help to disentangle the effect of a migration in situations where a direct analysis of human DNA returns conflicting results. Chapter 2 provides a comprehensive study of the genetic variation of the gastric bacteria *Helicobacter pylori* in Siberian populations. Employing ABC simulations I determined the origin of the bacterial subpopulations found in this region, subsequently using them to determine patterns of human migration inside Siberia and into the Americas.

In addition to the genetic analysis of non-human organisms, the study of cultural variability in different populations can provide information on their past. Studies employing both cultural and biological markers have highlighted how the variation of these two features can often be shaped by similar processes, such as isolation or contacts between different populations. In chapter 3 I report two case studies focused on gene-language coevolution. The first study employs a new classification method to infer relationships between several languages in Eurasia, which allows to obtain linguistic distances comparable with genetic ones and useful to determine the effect of large scale migrations on biological and cultural diversity. The second case study focuses on the origin of several

populations in northern Thailand, collectively called Khon Mueang, which show linguistic ties with the Tai-Kadai migration in Southeast Asia.

Inferring migrations using only contemporary data, both biological and cultural, can lead to a partial view of the demographic history of the studied populations. However, the use of genetic material retrieved from fossil remains allows to directly observe the variability in the individuals that took part in past migrations. In chapter 4 I outline the results of a study on complete mitochondrial genomes from necropolises associated with the Lombard migration in Europe. The genetic diversity of both medieval Lombard and non-Lombard individuals was used to shed light on previously unresolved hypothesis, which were based only on archaeological records.

ABSTRACT - ITALIANO

In questa tesi presento l'attività di ricerca svolta durante i tre anni di dottorato che ha portato alla pubblicazione di tre articoli su riviste peer-reviewed. Tutti i lavori presentati hanno come obiettivo lo studio delle migrazioni umane a partire da informazioni genetiche, ma ognuno differisce riguardo alle metodologie statistiche impiegate e alle domande prese in esame.

La storia di *Homo sapiens* è caratterizzata da grandi spostamenti che, partendo dall'Africa, lo hanno portato in poche migliaia di anni a raggiungere e colonizzare tutti gli angoli del Pianeta. Sebbene queste migrazioni siano da sempre oggetto di studi antropologici e archeologici, negli ultimi quarant'anni l'utilizzo della genetica di popolazioni ha permesso un notevole avanzamento delle conoscenze in materia. Oggi possiamo analizzare tramite un approccio multidisciplinare informazioni provenienti da diverse fonti, identificando relazioni tra popolazioni antiche e contemporanee (capitolo1). Tra queste, lo studio di specie che si sono coevolute con l'uomo può aiutare ad identificare gli effetti di una migrazione, contribuendo a chiarificare situazioni dove una diretta analisi del DNA porti a risultati controversi. In questo contesto il capitolo 2 contiene uno studio della variabilità genetica del batterio *Helicobacter pylori* in popolazioni siberiane. Utilizzando un approccio bayesiano approssimato ho determinato l'origine delle sottopopolazioni batteriche

individuate in questa regione, impiegandole successivamente per identificare le migrazioni umane in Siberia e nelle Americhe.

In aggiunta all'analisi genetica di organismi non umani, anche lo studio della variabilità culturale nelle diverse popolazioni in esame può fornirci indicazioni sul loro passato. L'analisi di marcatori culturali e biologici ha infatti evidenziato come la variazione di entrambi sia spesso causata da simili processi, come l'isolamento e il contatto tra popolazioni diverse. Nel capitolo 3 analizzo due casi studio focalizzati sulla coevoluzione tra diversità genetica e linguistica. Il primo studio riporta l'utilizzo di un nuovo metodo basato sulla sintassi per determinare le relazioni tra lingue in Eurasia, comparandole successivamente con quelle genetiche per determinare l'effetto concertato delle migrazioni sulla diversità biologica e culturale. Il secondo caso studio prende invece in esame migrazioni su scala locale e riguarda l'origine di diverse popolazioni nella Tailandia settentrionale, collettivamente chiamate Khon Mueang, che mostrano connessioni linguistiche con le popolazioni Tai-Kadai del Sudest asiatico.

Inferire migrazioni utilizzando unicamente dati contemporanei, siano essi biologici o culturali, può portare ad ottenere una visione parziale della storia demografica delle popolazioni analizzate. L' utilizzo di materiale genetico proveniente da resti fossili consente, invece, di osservare direttamente la variabilità degli individui che hanno preso parte a migrazioni nel passato. Nel capitolo 4 presento i risultati di uno studio basato su genomi mitocondriali completi provenienti da necropoli medioevali associate alla migrazione dei Longobardi in Europa. La diversità genetica tra individui longobardi e non-longobardi coevi è stata impiegata, attraverso un approccio simulativo, per risolvere ipotesi basate, fino ad ora, unicamente su fonti archeologiche.

Table of contents

The works presented in this thesis are published/under preparation for submission under the following titles:

Moodley, Y., **Brunelli, A**., Ghirotto, S., Chelysheva, V., Klyubin, A., Maady, A., Mominyalev, K., Linz, B., Achtman, M. The genetic diversity of *Helicobacter pylori* in Siberia (Manuscript in prep).

Brunelli, A., Kampuansai, J., Seielstad, M., Lomthaisong, K., Kangwanpong, D., Ghirotto, S., & Kutanan, W. (2017). Y chromosomal evidence on the origin of northern Thai people. PloS one, 12(7), e0181935.

Longobardi, W., Bartlett, M., Ceolin, A., Guardiano, C., Irimia, M.A., Kazakov, D., Michelioudakis, D., Radkevich, N., Sarno, S., Boattini, A., **Brunelli, A.**, Ghirotto, S., Sazzini, M., Susca, R. R., Tassi, F., Luiselli, D., Barbujani, G., Pettener, D. Who likes to travel alone? Grammars and genes in the history of Old World migrations (Manuscript under prep.)

Vai, S., **Brunelli, A.**, Modi, A., Tassi, F., Vergata, C., Pilli, E., Lari, M., Susca, R.R, Giostra, C., Pejrani Baricco, L., Bedini, E., Koncz, I., Vida, T., Mende,B.G., Winger, D., Loskotová, Z., Geary, P., Barbujani, G., Caramelli, D., & Ghirotto. S. A genetic perspective on Lombard migrations (Manuscript in prep.)

Additional works in which I have took part during my PhD are:

Kutanan, W., Kampuansai, J., **Brunelli, A**., Ghirotto, S., Pittayaporn, P., Ruangchai, S., Schroder, R., Macholdt, E., Srikummool, M., Kangwanpong, D., Hubner, A., Arias Alvis, L., Stoneking, M. (2017). New insights from Thailand into the maternal genetic history of Mainland Southeast Asia. bioRxiv, 162610. (Under review at the European Journal of Human Genetics)

Tassi, F., Vai, S., Ghirotto, S., Lari, M., Modi, A., Pilli, E., **Brunelli, A.**, Susca, R.R., Budnik, A., Labuda, D., Alberti, F., Lalueza-Fox, C., Reich, D., Caramelli, D. & Barbujani, G. Genome diversity in the Neolithic Globular Amphorae culture and the spread of Indo-European languages. (Accepted for publication in Proceedings of the Royal Society B.)

Kutanan, W., Kampuansai, J., Srikummool, M., Kangwanpong, D., Ghirotto, S., **Brunelli, A.**, Stoneking, M. (2017). Complete mitochondrial genomes of Thai and Lao populations indicate an ancient origin of Austroasiatic groups and demic diffusion in the spread of Tai–Kadai languages. Human genetics, 136(1), 85-98.

Chapter 1. INTRODUCTION

Population genetic tools in the study of human movements

The movement of people has been a feature of human populations in both ancient and more recent times. Broadly speaking, migration can be defined as the permanent movement of all or part of a population to settle and inhabit a new territory, separate from the territory in which it was previously based (Bellwood 2014). Over its history, archeology had a leading role in forming our knowledge on the past movements in which *Homo sapiens* was involved. However, several important migratory processes cannot be easily understood applying only the methods provided by this discipline. The similarity between two archaeological cultures can be the result of commercial exchanges, one or two directional cultural influences, spread of technological knowledge or, instead, the final outcome of a migration that brought into contact populations that were separated until that moment. Moreover, to understand past population movements archaeological studies rely on proxies, such as artifact and technologies, which are indirect representation of the people who have carried them. The field of population genetics has revolutionized the study of human movements, allowing inferences from the DNA of living populations or, more recently, from the genetic material retrieved directly from ancient human remains. Analyzing the migration history of different populations through their DNA means looking at patterns of gene-flow, as individuals who migrate into a different region but do not reproduce leave no genetic traces of their passage. Nowadays, studies on whole nuclear genomes are becoming increasingly common, providing an excellent source of material to analyze past events of gene-flow and isolation worldwide (Nielsen et al. 2017). However, the first studies aimed at detecting similarities between populations were not based on DNA, but rather on "classical markers" such as variants found in the blood group system ABO, the presence of different haemoglobin proteins and isoforms of the Human Leukocyte Antigen (HLA). The presence of these gene products was used as an indirect indication of the different alleles encoding them, providing the first method to quantitatively compare genetic diversity between populations (Cavalli-Sforza 1998). In one of the

first groundbreaking studies, the frequency of the alleles for 10 loci of these classical markers retrieved in living populations allowed to visualize the migration of Neolithic farmers from the Near East into Europe, with subsequent admixture with the local hunter-gatherer populations (Menozzi et al. 1978). The passage from classical markers to the direct use the DNA molecule became possible with the development of the PCR technique, which allowed to replicate high quantities of DNA from the smallest fragments of genetic material (Mullis and Faloona 1987). Molecular markers brought obvious advantages to the study of human migrations, as functional products of DNA reflect only a small fraction of the variation present in the genome. Moreover, loci encoding proteins represent candidate targets for selection processes, whereas other genomic region behave more neutrally, i.e. change in time under the effects of demographic changes, accumulating mutations over generations. Neutral markers are extremely important in population genetics as their variation, being largely free from selective pressures, can reflect past demographic processes such as contraction or expansion in population size.

In the 1980s the first studies employing molecular markers that aimed to understand the diffusion of modern human populations where based on mitochondrial DNA (mtDNA), a circular double stranded molecule of 16569 bp (Pakendorf and Stoneking 2005). Like the ones of other eukaryotes species, the cells of *Homo sapiens* are characterized by the presence of mitochondria, small organelles with a probable bacterial origin that absolve to the energetic requirements of the organism (Kivisild 2015). There are hundreds of mitochondria inside a cell, each one carrying its own copy of mtDNA, making this marker easy to be retrieved for both studies looking at living and ancient populations (Michaels et al. 1982; Pakendorf and Stoneking 2005). Moreover, mtDNA has an estimated mutation rate of $2.67x10^{-8}$ substitutions per site per year (Fu et al. 2013), which is considerably higher than the one found for nuclear genes and result in high levels of mtDNA polymorphism among different individuals. Finally, mtDNA lacks recombination and is transmitted exclusively via maternal inheritance, enabling the study of female migrations via phylogenies (Pakendorf and Stoneking 2005). These analysis led to the confirmation of an African origin for our species in 1987, consistent

with the proposed out-of-Africa model, as phylogenetic trees of human mtDNA had their deepest branches in Africa (Cann et al. 1987). Subsequent mtDNA studies tackled different anthropological questions from the matrilineal point of view, from the presence of ancient population structure in southern Africa (Barbieri et al. 2013) to the conquest of the Andes by South America populations (Fuselli 2003) and the extent of the Austronesian expansion in Oceania (Kayser et al. 2006). However, the majority of modern human societies practice patrilocality (70%) where, if a man and a woman marry but are not from the same locality, it's the woman who moves rather than the man (Jobling and Tyler-Smith 2003). This process likely resulted in the homogenization of mtDNA ancestries in an area, reducing the information that this marker can highlight on past population processes (Seielstad et al. 1998). Complementary to mtDNA studies are works based on the non recombining portion of the Y chromosome (NRY), a considerable portion of the male sex chromosome (95%) that allows to track population histories from the paternal line. As is the case for mtDNA, NRY haplotypes are not subjected to recombination and, therefore, are usually passed from generation to generation intact bar mutation events. An analysis of a series of single nucleotide polymorphisms (SNPs) on more than 1,500 individuals from Africa, Asia, Europe and the Americas confirmed the out-of-Africa migration model for modern humans (Hammer et al. 1998). More recently, a study on worldwide Y chromosome SNPs and short tandem repeats (STRs) located patterns of demographic expansions in relation to human technological advances (Poznik et al. 2016).

Before the advent of next generation sequencing (NGS) it was more cost-effective to analyze uniparental markers such as mtDNA and NRY in different populations with PCR and Sanger sequencing (Kivisild 2015). Today, even if new sequencing methods are allowing researchers to easily obtain complete genomic data, there are several reasons for which mtDNA and NRY are still extremely useful in studies on human migrations (for examples of applications see case studies reported in **Chapter 3 and 4**).

With the publication of the reference sequence for the human genome (Lander et al. 2001) more than 1.4 million autosomal SNPs were identified (Sachidanandam et al. 2001) enabling geneticists to

determine the history of populations employing known sets of polymorphic markers and developing SNP arrays for population genetics. The use of SNP arrays was initiated in medical genetics, where specific SNPs where investigated for the association with pathogenic traits (Novembre and Ramachandran 2011). The revolution brought with SNP arrays was caused by the fact that now, using a small set of predetermined SNPs, it was possible to infer the genetic variation in surrounding DNA regions in linkage disequilibrium (LD) with them. The discovery of patterns of LD and the characterization of sequence variants was the main task of the HapMap project, which resulted in the discovery of methods to develop more accurate and cheaper SNPs array (Gibbs et al. 2003). Over the last 15 years population genetic works investigating human migrations based on SNPs flourished, testing hypothesis that ranged from the relationship between genes and geography (e.g. Novembre et al. 2008) to the actual routes of colonization taken by our ancestors (e.g.Tassi et al. 2015). As for every partial representation of the entire genome, SNPs array present problems concerning the way that the included markers are selected. This ascertainment bias, if not taken into account during an analysis, can cause a systematic distortion of genetic variation favoring populations in which one determined SNP was first found (Lachance and Tishkoff 2013). However, recent efforts have been made to develop arrays for human population genetics with a known level of ascertainment bias (Lazaridis et al. 2014).

After an initial draft published in 2001, the Human Genome project was finally completed in 2004, after 14 years and 3 billion dollars spent (Human Genome Sequencing Consortium 2004). While it was a monumental achievement, it also highlighted the need to move beyond Sanger sequencing into faster and cheaper methodologies to obtain complete genomic information. The development of NGS technologies was an answer to the problem and today we are able to obtain the complete genome of an individual in days. While differing in some details, NGS methods all share three main characteristics: they don’t need bacterial cloning of DNA fragments, they sequence the DNA library in thousands of parallel reactions and are able to obtain the sequencing output without the need for an electrophoresis (van Dijk et al. 2014). Being characterized by an output of short reads, NGS

methods relies heavily on bioinformatics tools for task such as alignment and variant calling, while significant investment has to be made in order to obtain sufficient computational power to analyze whole genome sequencing (WGS) data. Thanks to the advent of NGS techniques in October 2015, less than 11 years from the end of the Human Genome project, the 1000 Genomes Project was able to efficiently reconstruct the low-coverage genomes of 2,504 individuals from 26 populations, discovering more than 88 million new genomic variants (Auton et al. 2015). Even with the computational challenges that they represent, the availability of numerus high quality genomes has opened the way to study of the effects of human migration at the genomics level (e.g. Mallick et al. 2016), providing a high resolution window in the past of our species.

Biological and cultural proxies in the study of human migrations

As we have seen the use of common genetic markers such as SNPs, mitochondrial DNA, Y chromosome and STRs has greatly improved our ability to unravel past population movements (for a review see Cavalli-Sforza & Feldman, 2003). However, the use of these markers may sometimes fail to reconstruct recent demographic processes, such as individuals presenting low levels of genetic diversity as a consequence of a bottleneck followed by a rapid migration (Wirth et al. 2005). In addition to the direct analysis of human DNA, different cultural and biological sources can help to elucidate the history of human movements. One of the most efficient methods that has been coupled with traditional population genetics is the analysis of other species which took part in human migrations, also referred as bioproxies. While having different characteristics, bioproxies can be considered as "living-artifacts", meaning organisms taken to a place by people moving between locations (Jones et al. 2013). Over the last 20 years studies on organisms moving with humans have identified three main categories of bioproxies:

1. Domesticated species
2. Commensals
3. Parasites and pathogens

The study of species falling in the first category have helped understanding their spread from domestication centres to their current location and have been carried out on both animals (e.g. Ajmone-Marsan et al. 2010; Murray et al. 2010) and plants (e.g. van Heerwaarden et al. 2011; Myles et al. 2011). More importantly for understanding human migrations, tracking these organism can provide insight into population movements. The study of mitochondrial DNA retrieved from domestic pigs (*Sus scrofa*) in Southeast Asia and New Guinea, for example, was able to identify different clades associated with subsequent waves of human colonization (Larson et al. 2007).

Commensal species are organisms that make use of resources, such as food and shelter, provided by our species and can be transported unintentionally during movement of people (Jones et al. 2013). Different commensals have been proven helpful in elucidating human migrations, both invertebrates (e.g. Keller 2007; Jesse et al. 2011) and vertebrates (e.g. Matisoo-Smith and Robins 2004). A good example of commensal often employed in these kind of studies is the house mouse (*Mus musculus*). Some mitochondrial haplotypes found in this species have been associated with the spread of Norwegian Vikings from their homeland during the 8^{th}-10^{th} century CE (Searle et al. 2009). Moreover, a study on the colonization of Iceland by both mice and Vikings showed how demographic changes in human populations, such as reduction of effective population size due to founder effect, were mirrored in the genetic diversity of the co-migrating population of *Mus musculus* (Jones et al. 2012).

The analysis of the numerous pathogens and parasites that have travelled with humans for thousands of years has provided some of the most striking examples of coevolution (Wirth et al. 2005). A study on mitochondrial DNA from a global samples of human follicle mites, for example, showed how genetic diversity in these symbionts reflected ancient divergences between human populations, supporting an "Out of Africa" model for our species (Palopoli et al. 2015). Pathogens, such as different species of bacteria, also presents high mutations and recombination rates, allowing for the identification of recent human movements and zones of secondary contact between different populations (Falush et al. 2003). Humans are the only known reservoir of *Mycobacterium leprae*, a

bacteria whose infection lead to a chronic dermatological and neurological disease commonly known as leprosy. An analysis of seven strains of this bacteria was able to trace the origin of leprosy in East Africa, with subsequent spread of this disease following human migrations (Monot et al. 2005). More recently, the phylogeographic study of 400 sequences of *M.leprae* from different regions of the world enabled the authors to locate a connection between European and Asian strains, possibly caused by movement of people on the Silk Road (Monot et al. 2009).

Different unresolved issues in the migration history of our species can benefit from the uses of bioproxies. In this context I have used the genetic diversity of *Helicobacter pylori*, a common bacteria living in our stomachs, to determine human population movements in Siberia over the Last Glacial Maximum and the timing of the arrival of our species in the Americas (**Chapter 2**).

The study of closely related cultural complexes, characterized by features such as shared material culture or related languages, has helped to clarify relationships between different populations (Bellwood 2014). Culture is defined as information capable of affecting individual's phenotypes, which they acquire from other conspecifics by teaching or imitation (Boyd and Richerson 1985).

The coevolution between genes and culture began to be formally investigated during the '80, when mathematical models where applied to understand the development of human cultural traits and behaviours (Cavalli-Sforza and Feldman 1981; Boyd and Richerson 1985). However, the first idea of a parallelism between cultural change and genetic evolution was not new, as it was advanced by Charles Darwin in his Origin of the Species (Darwin 1859, **Chapter 3**). In his Descent of Man, he also stated 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. . ."

Researchers in this field view inheritance of genes and culture as a common contribution, by previous generation, into the new ones. However, cultural traits can be acquired not only from parents, but also from other members of the population, possibly leading to conflictual histories with genetics (Cavalli-

Sforza and Feldman 1981). Even accounting for this, different cultural markers appear to show geographic signatures of human demographic history (e.g. Ranciaro et al. 2014; Creanza and Feldman 2016). The techniques employed to study these dynamics are typically developed from the field of population genetics; a typical analysis of gene/culture coevolution in a human group usually involves a study of variation in allele/genotype frequency, together with change of cultural traits in the same populations (Laland et al. 2010). Following works have identified that cultural variation, like biological variation, is influenced by demic processes such as migration, changes in population size and adaptation to different ecological context (for a review see Creanza and Feldman 2016).

One of the most striking analogy between cultural and biological transmission is the loss of variants due to chance. This process, commonly known as genetic drift in population genetics, can occur in cultural traits when the number of practitioners is small and the chance of transmission become low (Shennan 2001). The effects of this cultural drift will be higher in small populations than in larger ones, because such random losses are more likely. An analysis of marine foraging tools in Oceanian populations showed how larger islands, harbouring a greater number of individuals, where characterized by a larger repertoire of instruments with respect to smaller ones (Kline and Boyd 2010).

However, as is the case for gene-flow, new cultural traits can be introduced in a population following contact with another, replenishing the amount of variation lost by drift and giving the possibility of developing new adaptive traits. This was the case for the Polar Inuit of north-western Greenland that, when visited by explorers in mid-nineteen century, lacked kayaks, snow houses with long entryways and bows typical of other Inuit populations. They received this tools in 1862 after a population from Baffin island migrated in the region (Golden 2006).

The interaction between cultural practices and the genetic history of a population is also evident analysing the consequences of assortative mating, such as consanguinity and marriages within subgroups of a population. In the classical example of Hindu caste systems, women are allowed to marry into higher social caste under specific circumstances while men are traditionally not socially

mobile. This has led to female-specific gene-flow and, in some cases, to genetic stratification of the population (Bamshad et al. 1998).

Selection for specific alleles can also be triggered by cultural factors, as it is shown by the well studied example of lactose tolerance. Lactose tolerant individuals are individuals in which the ability to digest lactose persist into adulthood. This characteristic arise from one or multiple SNPs found upstream of the lactase gene that, depending on the population considered. The frequency of lactose tolerant people is high in population that have traditionally a history of dairy farming, such as Europeans and herders from the Middle East and Africa, but remain low almost everywhere else. There is, therefore, a strong correlation between the culture of milk drinking and lactose tolerance, which led to the hypothesis that dairying and milk-drinking created the selective pressure that drove alleles for lactose tolerance at high frequencies (Ranciaro et al. 2014). This conclusion is further strengthened by the absence of the allele for lactose in the genome of early Neolithic Europeans, suggesting that this characteristic was not present or, at least, present at low frequencies 7,000 – 8,000 years ago (Burger et al. 2007). Similar examples of selective pressure caused by culture were located for the salivary amylase gene in populations with a starch-rich diet (Perry et al) and in the hemoglobin S allele in Kwa-speaking agriculturalist from West Africa, which face increasing risk of malaria from their practice of cut clearings in forest to grow crops.

Since the beginning of gene-culture coevolution studies, linguistic diversity between populations has been widely investigated for patterns of concordance (or discordance) with their genetic variation. In **Chapter 3** I report the analysis of structural linguistic and genetic variation among 28 Eurasian populations. Moreover, as information from linguistic studies may help to shed light on past migrations, in the same chapter I report an analysis on the genetic diversity of Northern Thailand populations in relationship with their linguistic affinity.

Discovering ancient migrations and contacts with aDNA

Once an organism die its DNA usually degrades rapidly due to the action of endogenous nucleases and proteases, combined with the effect of exogenous organisms such as bacteria and fungi (Hofreiter et al. 2001). However, under specific environmental circumstances such as low temperatures, low humidity and high salt concentration, DNA can survive for thousands of years. In 1984 Russell Higuchi and colleagues published the first genetic data extracted from an extinct organism, a partial mitochondrial sequence of a quagga zebra (*Equus quagga quagga*) retrieved from a museum specimen prepared a century earlier (Higuchi et al. 1984). This work launched the new field of ancient DNA (aDNA) that, over the last 30 years, revolutionised the field of population genetics. Before recovering molecules from archaeological remains became a feasible task, past changes in demography could only be inferred from current patterns of genetic diversity, relying heavily on assumptions for parameters such as population growth and migration rates. With aDNA it has become possible to directly compare the genetic makeup of living populations with the one we obtain from their ancestors.

However, due to aDNA characteristics, specific laboratory protocols and ad-hoc bioinformatics pipelines are required to integrate this data in population genetics analyses (Slatkin and Racimo 2016). The main problem associated with ancient genetic material concerns the presence of contamination. The DNA extracted from a living individual will consist, for the main part, of endogenous material if standard laboratory practices are followed. On the contrary, most of the genetic data extracted from archaeological remains tends to have an exogenous origin, such as microbial and environmental DNA left by different organisms. Moreover, contamination of ancient samples after collection by modern researchers represent a problem, especially when extinct human populations are studied (Hofreiter et al. 2001). Human DNA is widespread in location where archaeological remains are handled, such as laboratories and museums, and cannot be distinguished by aDNA from *H.sapiens*. In order to deal with these issues, researchers have agreed to a set of rigorous practices when analysing aDNA that cover every step of the sample preparation, from the

treatment of remains with UV radiation to the inclusion of adapters to tag endogenous molecules in order to prevent confusion with subsequent contaminants (Slatkin and Racimo 2016).

The extreme low quantity of genetic material obtained from ancient samples and its overall quality represents another challenge in ancient DNA studies. Even if an organism die in the perfect conditions for DNA preservation, different post-mortem processes will lead to changes in the nucleotide sequence ad to its extreme fragmentation. The sugar-phosphate backbone and the nitrous bases of DNA are, for example, modified by the combined effect of oxidation and background radiation (Hofreiter et al. 2001). Moreover, hydrolytic processes such as deamination and depurination will lead to breaks in DNA molecules, resulting in average length of sequence fragments retrieved from ancient remains of around 70 bp (Green et al. 2009). This alterations to the genetic material of an organism are typical features of aDNA and, however problematic, can be useful to discriminate between real endogenous sequences and contaminants (Skoglund et al. 2014). With the advent of PCR (Mullis and Faloona 1987) and, especially, of the genomic revolution, the amplification of even the lowest fraction of aDNA became feasible.

The first study employing aDNA in the context of ancient human populations was made by Svante Pääbo in 1985 and involved an analysis of Egyptian mummies (Pääbo 1985). In the context of this work, bacterial cloning was used to amplify the DNA of a 2,400 year old mummy of a child, obtaining 3.4 kilobases of endogenous genetic material. This initial study was quickly followed by several aDNA studies that have, until recently, mainly involved the use of mitochondrial data (mtDNA). Due to their sheer abundance in the cell, mitochondria are more likely to be retrieved from ancient remains with respect to nuclear DNA. Mitochondrial DNA also present a high mutation rate and no recombination, leading to high haplotype diversity that can be studied with gene phylogenies (e.g. Ghirotto et al. 2013; Alberdi et al. 2015). Even with the genomic revolution underway, studies based on ancient mitochondrial data remain useful to provide high quality sequences on which infer past demographic processes and migrations (e.g. Modi et al. 2017). The genomic revolution impacted also aDNA studies and, nowadays, NGS technologies allows for in depth studies on past human

migrations (Bentley et al. 2008). High throughput sequencing allows for the parallel sequencing of millions of DNA molecules, exponentially reducing the cost of sequencing and increasing the amount of data generated. Moreover, this technologies are particularly well suited to analyse shorter DNA fragments, such as the one we normally found in aDNA, as they do not rely on target PCR amplification of the molecule using primers (Green et al. 2010). Ancient DNA sequences are also short enough to be sequenced completely from both ends, further reducing sequencing errors (Stoneking and Krause 2011).

The ability to obtain aDNA was invaluable to determine the interaction of ancient humans with extinct archaic hominids. The first draft of the Neanderthal genome (~1.3× coverage) was obtained from three individuals found in the Vindija Cave in Croatia (Green et al. 2010). When compared to different genomes from several modern populations, the Neanderthal data showed higher similarity with present-day non African individuals, a pattern that was shared by all Neanderthal genomes sequenced until today (Prüfer et al. 2013; Castellano et al. 2014). The hypothesis of an admixture event between Neanderthals and anatomically modern humans outside Africa seems to receive final confirmation from an analysis of 37,000–42,000-year-old human remains from Pestera cu Oase, in Romania. The aDNA retrieved from the individual presented 6-9% of Neanderthal contribution, suggesting that he might have had a Neanderthal ancestor in the least 4-6 generations (Fu et al. 2015). The use of aDNA techniques was pivotal in determining the relationship between *Homo sapiens* and the Denisovan archaic hominids. The 1.9x genome from a small finger bone, one of the few anatomical remains found in the Denisova cave, and the following high coverage one allowed to characterized this population as a sister group of Neanderthals (Reich et al. 2010; Meyer et al. 2012). Moreover, when compared to the DNA of modern human populations, the Denisovan genome showed higher affinity with native Australians and Melanesians, suggesting another ancient admixture event (Sawyer et al. 2015).

Even within anatomically modern humans, aDNA revealed the incredible complexity of human colonization of different world regions. The New World Artic present optimal conditions for DNA

preservation and, in 2010, the 20x genome of 4,000 years old individual identified as Paleo-Eskimo was obtained (Rasmussen et al. 2010). This single ancient genome was sufficient to suggest that ancient inhabitants of Greenland came into the Americas with a different migration with respect to the one that gave rise to present-day Native Americans. More details on the colonization of the Americas were obtained analysing the genome of a 24,000 Siberian individual named Mal'ta (Raghavan et al. 2014). While the Mal'ta genome had low affinity to present-day East Asian populations, it showed marked similarities with both Western Eurasian and Native Americans, suggesting and admixture event between the ancestors of the latter population and ancient Siberians before the first crossing of the Bering Strait. Recent work with aDNA retrieved from ancient Americans, including data from the 9,000 years old "Kennewick Man", highlighted genetic continuity between the first migrants into this region and present-day Native populations (Raghavan et al. 2015; Rasmussen et al. 2015). These studies, together with the numerous ones focused on ancient Western Eurasians (for more details see **Chapter 4**), show the importance of cold and dry environmental conditions for the survivability of genomic data. Indeed, nuclear genomic sequences from remains found in other world regions were, until recently, few and far between (Slatkin and Racimo 2016). In 2016 genome wide data were retrieved from three individuals that lived on the islands of Vanuatu and one from the island of Tonga, spanning a time period from 3,100 to 2,300 years before present (Skoglund et al. 2016). A comparison of these aDNAs with modern data from East Asians and Oceanians found little Papuan ancestry in them, in stark contrast with what is found in current inhabitant of Vanuatu and Tonga, locating a population movement that happened in the last 2,000 years. In 2015 the first African ancient genome was obtained from a 4,500 years old individual found in the Mota cave in southeastern Ethiopia (Llorente et al. 2015). As it was the case for the Kennewick man in North America, the Mota DNA showed general continuity with present inhabitant of Ethiopia and revealed the influence of European Neolithic farmers on sub-Saharan populations. This single ancient genome from Africa was recently joined by 16 more retrieved in different regions and

spanning a period from 10,000 to 1,300 YBP, allowing the analysis of Western Africa population structure and the discovery of several episodes of gene-flow (Skoglund et al. 2017).

Even accounting for the potential problems, human aDNA studies have proven invaluable to discover the genetic diversity of past populations, infer selection processes and determine past demographic events. In **Chapter 4** I report the results of a study focused on reconstructing an ancient migration in Western Europe using aDNA. Specifically, the work involved an analysis of complete mitochondrial genomes from individuals associated with the Lombard culture, in order to unravel their colonization route and the potential contacts with local populations.

Chapter 2. TRACING HUMAN MIGRATIONS WITH *HELICOBACTER PYLORI*

Helicobacter pylori: a brief introduction

In 2005 Marshal & Warren were awarded the Nobel prize in medicine and physiology for the discovery of the cause of gastritis, the bacteria *Helicobacter pylori* (Marshall and Warren 1984). *H.pylori* is a gram-negative, spiral shaped bacterium characterized by the presence of 5-7 flagella at the distal end. The pathological importance of this bacterium goes beyond gastritis, as the presence of *H.pylori* has been associated with MALT (mucosa-associated lymphatic tissue lymphomas), peptic and duodenal ulcers. However, the offset of actual diseases occur in a limited number of cases and the bacteria behaves as a gastric commensal in the majority of infected individuals (Wirth et al. 2005). As a consequence of this reduced pathogenicity, the distribution of *H.pylori* in human populations is nearly ubiquitous, with more than half of the world population infected (Feldman 2001). The rates of infection varies according to socioeconomic conditions, with developing countries harboring up to 80% of adult infected, while industrialized populations show reduced figures, hovering around 40% or even 20% (Malaty and Graham 1994). *H.pylori* is usually acquired early in life, during childhood, from ingestion of the bacteria transmitted by family members. While this vertical transmission seems to be the prevalent way to acquire an infection, horizontal transmission between unrelated individuals is also present (Schwarz et al. 2008). Infection by *Helicobacter pylori* is usually chronic and lasting for the entire life of the host (Suerbaum and Michetti 2002). *H.pylori* colonize the gastric mucosa, mucus layer of the stomach, an environment that protects this bacteria from the high acidity of the gastric tract. However, the gastric mucosa still present an harsh and acidic environment, forcing selective pressure on *H.pylori* to develop specific adaptations, such as the ability to produce Urease in order to hydrolyze the urea into carbon dioxide and ammonia, raising the PH to a level compatible with its growth (Achtman and Suerbaum 2001). The genome of *Helicobacter pylori* is quite small, around 1.65 million bases annotated for 1500 genes, resulting in it being the first bacteria for which

two strains were completely sequences, 26695 (Tomb et al. 1997) and J99(Alm et al. 1999). However, genetic diversity between different *H.pylori* strains is greater than the one found in most of the other species of bacteria (Achtman et al. 1999). Identical sequences are indeed uncommon in *H.pylori*, being found only within families or close communities: an analysis of 3,850 nucleotides in 370 strains from different world regions obtained 1,418 polymorphic positions (Falush et al. 2003). This high diversity is caused by the fast mutation rate found in *H.pylori*, which has been estimated to be between 8.47×10^{-7} and 9.73×10^{-4} per nucleotide per generation (Montano et al. 2015). It has been suggested that this mutation rate could be caused by the lack of genes for DNA repair in *H.pylori* (Stearns and Koella 2008). Recombination is also more frequent in this bacteria than in others (Achtman et al. 1999) and has been proven to have an higher importance in diversifying strains than mutation (Falush et al. 2001). The characteristics of extreme diffusion, high mutation and recombination rate, small genome size and familiar transmission make *H.pylori* an ideal candidate to be used as a bioproxy to investigate human migrations.

The association between *Helicobacter pylori* and anatomically modern humans

The first studies looking at the genetic variability of *H.pylori* were not based on complete genomes but rather on MLST gene fragments. Multi locus sequence typing (MLST) is a technique that concerns the sequencing of several housekeeping gene portions, typically seven, that are sufficiently distanced from one another to be affected by a single recombination event (Didelot and Falush 2007). These studies highlighted the clustering of *H.pylori* sequences according to the continent in which they were sampled, highlighting a profound phylogeographic structure for this bacteria (Achtman et al. 1999). Subsequent studies involving a higher number of MLST sequences confirmed this pattern and were able to name the found clusters based on the region in which they were found more commonly, such as hpAfrica1 in Western Africa or hpEurope in Europe (Falush et al. 2003). This groups were, in turn, divided into different regional subpopulations called “hsp” that can shed light on the spread of human populations carrying them. For example, hspMaori is a subpopulation of

hpEastAsia that was isolated from Oceanic populations and from the indigenous people of Taiwan, supporting the "out-of-Taiwan" model of colonization for remote Oceania (Moodley et al. 2009). This division into geographical clusters also highlighted the legacy of recent population movements, such as the European colonial expansion, which brought along with them strains of bacteria not thought to be originally present in the region (Falush et al. 2003). Nowadays we can find *H.pylori* strains classified as hpEurope in the Americas, in Africa and in Australia (Linz et al. 2007; Nell et al. 2013; Thorell et al. 2017) while the results of the transatlantic slave trade could be detected by the presence of hspWestAfrica in different North American populations.

All of the different populations and subpopulations of *Helicobacter pylori* are the results of events of admixture and splits between proposed ancestral sources, an analysis of which can shed further light on past human movements (Falush et al. 2003). The hpEurope population, for example, has shown influences of both an ancient Eurasian source (AncientEurope1) and a group more closely related to middle eastern/north African strains (AncientEurope2). The tight association between *Homo sapiens* and *Helicobacter pylori* was further strengthened by a combined analysis of 532 bacterial isolates and 1048 human individuals typed for 783 autosomic microsatellites, which showed a similar pattern of decreasing genetic diversity with distance from east Africa in both species (Linz et al. 2007). This result hints to a probable coevolution between the bacteria and its host, with an association of the two that goes back before humans left Africa to colonize the rest of the world. Recent analyses employing the complete genome of this bacteria, while extremely important to infer the effect of selection and to estimate demographic changes in *H.pylori* populations, did not change the patterns inferred by MLST studies, highlighting a high degree of diversity inside African groups, especially in the Southern African hpAfrica2, and a general similarity between strains found in the other regions of the world (Montano et al. 2015). The increasing availability of new strains, such as the one belonging to the 5,300 years old *H.pylori* found in an European Copper age mummy (Maixner et al. 2016), allows researchers the ability to test previous hypothesis on human migrations (i.e. Moodley et al.

2009; Linz et al. 2014), as well as the possibility to discover previously undetected patterns (i.e.Thorell et al. 2017).

Case study: The genetic diversity of *Helicobacter pylori* in Siberia

Outline of the research

Siberia is a vast area that comprise all territories between the Ural Mountains to the west, the Pacific Ocean to the east and the steppes of Kazakhstan and Mongolia to the south. While this region is characterized by challenging environmental conditions, anatomically modern humans were able to colonize western Siberia as early as 45,000 years ago (Fu et al. 2014), before venturing above the arctic circle at least 27,000 years ago (Pitulko et al. 2004). Today, this region is inhabited by a relatively limited number of people, often living in small communities, which are however characterized by languages belonging to different families (i.e. Turkic, Chukotko-Kamchatkan, Mongolic) and contrasting subsistence techniques (i.e. pastoralism, hunter-gathering, herding) (Pugach et al. 2016). The patterns of genetic diversity between these populations is still largely understudied and even massive projects, such as the 1,000 genome (Auton et al. 2015), often include only a reduced representation of the individuals inhabiting this region. In recent years, thanks to the findings of genomes belonging to ancient Siberian individuals (e.g.Raghavan et al. 2014), the complex role of Siberia as a gateway for human migrations to Western Eurasia and Northern America is beginning to be unraveled.

Anatomically modern humans used the Beringia land bridge in the eastern part of Siberia to colonize North America (Brigham-grette et al. 2004; Goebel et al. 2008). However, the time of colonization, the size of the founder population and the relationships of these migrants with other human groups are still unclear (for a review see Skoglund and Reich 2016). In addition to the colonization of the Americas, the most controversial topic in Siberian archaeology regards internal migrations and, specifically, how humans coped with the increasing cold of the last glacial maximum (LGM 20 –

18,000 years ago). Ideas on the topic can be summarized by two contrasting theories: the first one see a uninterrupted occupation of Siberia by anatomically modern human during LGM, albeit in sheltered locations with favorable environment (e.g. Kuzmin 2008; Kuzmin and Keates 2013), while the second one postulate a complete depopulation of central Siberia, with a recent repopulation coming from the south(e.g. Graf 2009; Pugach et al. 2016).

While the first global studies on the genetic variability of *Helicobacter pylori* discovered the existence of hspAmerind, a subpopulation of hpEastAsia found only in Native Americans (Falush et al. 2003), sequences from Siberia have yet to be reported. In order to reconstruct both movements inside Siberia and the timing of the colonization of North America we sampled 16 human populations with different ethnic background for presence of *Helicobacter pylori*. We then explored the genetic variability of this bacteria inside the region, constructing demographic models aimed at testing both previously thought hypotheses and specific diffusion scenarios.

Material and Methods

Samples of *H.pylori* where obtained by Gastroduodenal endoscopy and were cultured as described in previous publications (Breurec et al. 2012). DNA was extracted with the QIAmp ™ (Qiagen, Courtaboeuf, France) while PCR amplification and sequencing of the MLST genes *atpA*, *efp*, *mutY*, *ppa*, *trpC*, *ureI* and *yphC* was performed as described in Linz et al. 2007. This procedure resulted in 395 new strains belonging to 16 Siberian populations (Fig.2.1).

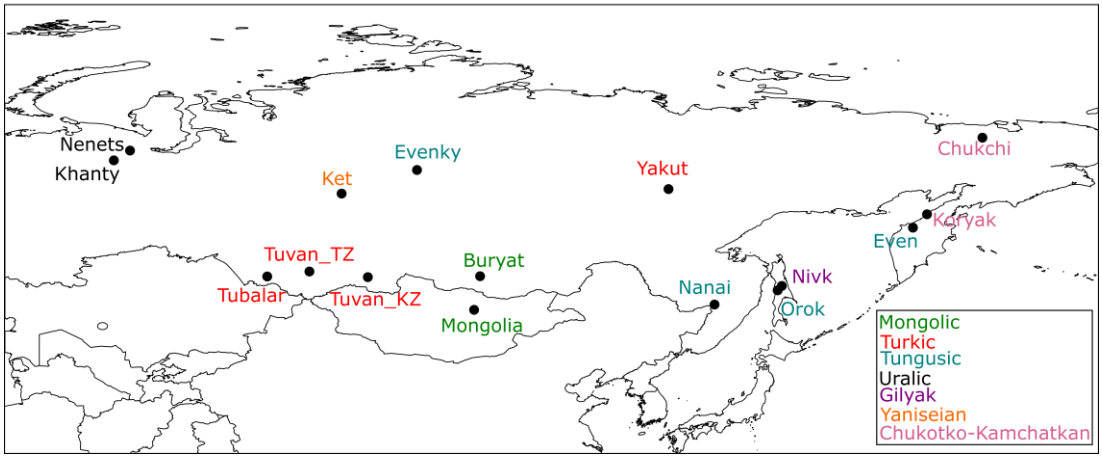


Fig 2.1. Populations sampled for Helicobacter pylori colour coded based on their linguistic family

In order to determine relationship between these new strains and the previously discovered variability inside Asia of *H.pylori* we merged our new sequences with 620 previously published MLST strains (https://pubmlst.org/ helicobacter/ Jolley and Maiden 2010). We aligned or sequences using *mafft* (Katoh and Standley 2013) and removed missing positions using a custom made R script. Our final dataset included 1,015 sequences from Asia, Oceania and the Americas, which shared 3,345 positions.

Our first need was to determine the population structure inside our dataset and, in order to do it, we employed a multivariate approach conducting a discriminant analysis of principal components (DAPC) (Jombart et al. 2010). This method assess the presence of clusters optimizing the variation of allele frequencies between- and within-groups, returning the most supported subdivision through the Bayesian information criteria (BIC). The DAPC is also characterized by a lack of assumptions about Hardy-Weinberg equilibrium or linkage equilibrium. We assessed the most supported number of clusters for our *H.pylori* data employing the *find.clusters* function in *adegenet* 1.3–1 (Jombart 2008), comparing the result of 10 independent runs using a custom made R script. We then ran the DAPC analysis with 1,000,000 iterations checking the consistency of the inferred groups over 10 different runs. As a confirmation of the pattern of similarity between the groups outlined by the DAPC we also conducted a Bayesian analysis of population structure using the software STRUCTURE

(Prithchard et al. 2000). We ran 200,000 iterations discarding the first 50,000 as burn-in and testing a number of partitions (K) that went from 2 to 10 under the linkage model (Falush et al. 2003), replicating 10 run for each tested K.

To determine the origin pattern of the newly identified Siberian groups and the timing of human migration to the Americas we employed an Approximate Bayesian Computation (ABC) approach (Beaumont et al. 2002; reviewed in Bertorelle et al. 2010). As the problematic required different tactics we employed two set of models.

The first set regarded the origin of the new Siberian subpopulations of *H.pylori* and was constituted of two basic demographic models, called tree-like and admixture, which we applied on different bacterial groups depending on the ancestry that we were investigating (Fig. 2.4A, 2.5A). The tree-like model postulated an origin of the target subpopulation by split from a previously established one, without additional gene flow from other *H.pylori* groups. The admixture model, instead, hypothesized the development of the target subpopulation from the genetic contribution of two existing groups.

The second set was focused on reconstructing human migration into the Americas and, to do so, we employed only the sequences clustering within the hspAmerind subpopulation. We first subdivided our hspAmerind dataset based on the sampling location: north Siberia (NS; Khanty, Nent, Tuvan and Evenk), east Siberia (ES; Nivk, Orok, Hokkaido and Nanai), Kamchatka (KC; Chuchi, Koryak and Even) and Americas (AM; Eskimo, Athabaskan, Venezuela, Huitoto and Peru). We then tested 8 models including always one population of hspEastAsia (HongKong) as outgroup (Fig. 2.6A). MOD1-4 depicted simple tree-like scenarios, two of which postulate migration events before (MOD3) or after (MOD4) the flooding of the Bering Strait. MOD5 was designed to test an event of gene-flow from western Eurasia (represented here by NS) in the ancestry of Native Americans (Raghavan et al. 2014). Finally, MOD6-8 simulate a bottleneck after the founder event that led to the colonization of North America, which could have happened without following events of migration (MOD6) or, instead, also have included bidirectional migration from AM to KC (MOD7) and from ES to KC (MOD8) to account for the similarities in strains retrieved from these regions.

The ABC methodology allows to compare statistics computed on the genetic variability of the observed *H.pylori* subpopulations with the ones calculated on simulated datasets by means of the coalescent theory. These simulated dataset are, in turn, generated taking into account the prior distribution of several demographic parameters, which characterize each of the competing models developed to explain the genetic variability of *H.pylori*. The demographic model producing the simulations with the statistic closer to the observed data is regarded as the best representation, between the proposed ones, of the evolutionary history of the studied *H.pylori* populations. To generate the simulated datasets, we used the software package *ABCtoolbox* (Wegmann et al. 2010), running 500,000 simulations for each tested model. To summarize the genetic information contained in each model we calculated an array of summary statistics within and between populations: we considered the number of haplotypes, the number of private polymorphic sites, Tajima's D, the mean number of pairwise differences for each population, the mean number of pairwise differences between populations and pairwise F_{st}. The simulations generating the summary statistics most similar to the observed ones, measured by means of Euclidean distance, are chosen to compute the posterior probability of each model using a weighted multinomial logistic regression (LR, Beaumont 2008). In the LR methodology, the model is considered as the categorically dependent variable in the simulations and the summary statistics as the predictive variables. The regression is local around the vector of observed summary statistics and the probability of each model is finally evaluated at the point corresponding to the observed vector of summary statistics. Maximum likelihood is used to estimate the β coefficients of the regression model. To evaluate the stability of the models' posterior probabilities, we considered different thresholds by considering different number of retained simulations for LR (50,000, 125,000, 250,000 best simulations). As the time of formation of target populations is one of the central parameters we required to validate hypothesis on the peopling of Siberia and the colonization of North America, we estimated the parameters of the chosen model using a locally weighted multivariate regression (Beaumont et al. 2002) on the 5,000 best-fitting simulations after a *logtan* transformation (Hamilton et al. 2005).

Results

The k-means analysis highlighted 10 as the most supported number of clusters in which to divide our *H.pylori* dataset (Fig 2.2A) while a scatterplot of the DAPC (Fig. 2.2B) highlighted the pattern of similarity between them.

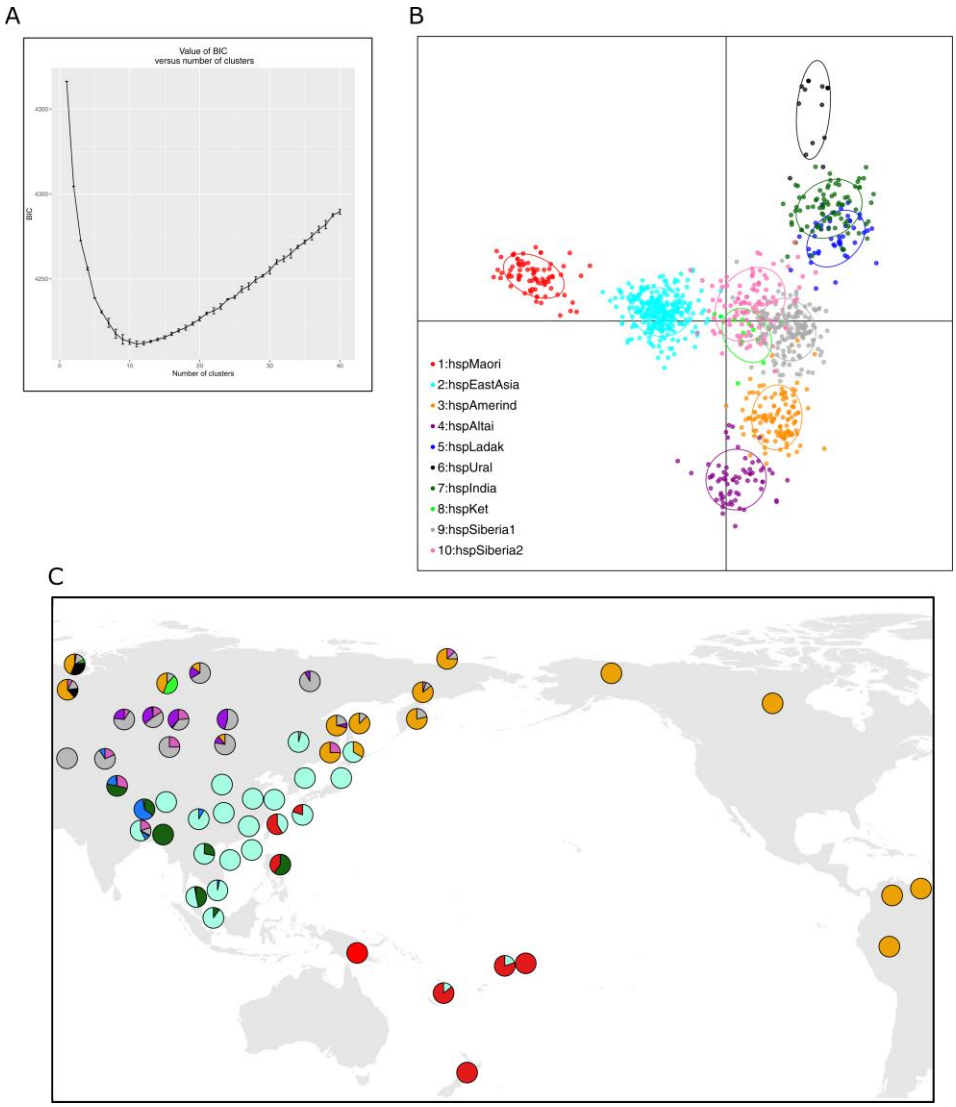


Fig.2.2. A)Mean BIC over 10 runs of k-means analysis, B) Scatterplot of the most supported number of groups for the H.pylori dataset, C) Frequency of the groups resulting from the DAPC in the studied populations

Analyzing the spatial distribution of the identified groups and the affiliation of the previously identified strains contained in them (Fig. 2.2C), we determined that five of the located clusters represented already established subpopulations of *H.pylori*. Group 1 (red) was typically found in Oceanic populations and was well differentiated from its closest relative, group 2 (light-blue),

allowing us to assign the former to hspMaori and the latter to hspEastAsia (Falush et al. 2003; Moodley et al. 2009). The intersecting group 7 (dark green) and group 5 (blue) presented, respectively, high frequencies in south/southeast Asia and northern India, indicating that they represented the two main subpopulation of hpAsia2; hspIndia and hspLadak. While the 4 already discussed groups did not present any newly sequenced *H.pylori* strain from Siberia clustering within them, they provided additional confirmation on the validity of the DAPC method in assigning sequences to subpopulations, as it was able to reconstruct identical hsp to the one previously determined with STRUCTURE. Group 3 (orange) contained all the published sequences attributed to the hspAmerind subpopulation and recovered from Native Americans (Linz et al. 2007) but, in addition, also included 108 newly sequenced strains from Siberia. The distribution pattern of this subpopulation seemed peculiar as, with the exception of 11 sequences in the Kets and 4 in the Evenky, hspAmerind was absent from the majority of the central Siberian populations and appeared to be relegated to the eastern/western fringes of the region.

Amongst the five newly identified groups containing Siberian sequences, three clustered together between hspAmerica, hspEastAsia and hpAsia2: group 8 (light green), group 9 (gray) and group 10 (pink). Group 9 was constituted by 177 new sequences and represented the most diffuse cluster in central Siberian populations, reaching frequencies of 91% in Yakuts and 66% in Evenks, while also being present, albeit at lower frequencies, in eastern and western populations. Compared to group 9, group 10 showed a reduced diffusion and was mainly present in southern Siberian populations, such as in Tuvans (23% in Todzha and 16% in Kyzil) and Nanai (25%). Due to their geographic spread and of their prevalence in the region we provisory named these two groups as hspSiberia1 (group 9) and hspSiberia2 (group 10). While similar to the previous clusters in the DAPC scatterplot, group 8 presented the most limited number of strains included (13) and distribution, being found at high frequencies exclusively in the Ket population (44%). As a consequence of this, we labeled this cluster hspKet. The two remaining DAPC groups made up entirely of newly sequenced Siberian strains departed from the pattern of general similarity shown by hspSiberia1-2 and hspKet. Group 6 (black,

16 sequences) presented only relative similarities with hpAsia2 in the scatterplot and the strains clustering within it belonged only to the Nenet and Khanty populations. We provisionally name this possible subpopulation of hpAsia2 as hspUral, due to its presence in this limited geographical region. As for hspUral, group4 (purple, 67 sequences) was restricted to a specific geographic region, the Altai Mountains, and was mainly found in populations speaking Turkic languages such as the Yakuts, the Tubalars and the Tuvans. While dissimilar from the majority of the remaining clusters located by the DAPC, this newly found hspAltai showed affinity with hspAmerind.

The pattern showed by the STRUCTURE analysis at K=10 confirm the overall pattern found by the DAPC (Fig. 2.3).

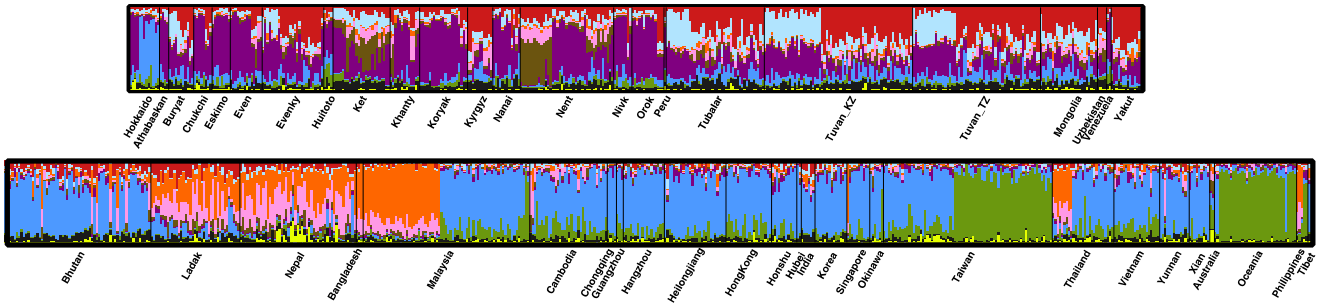


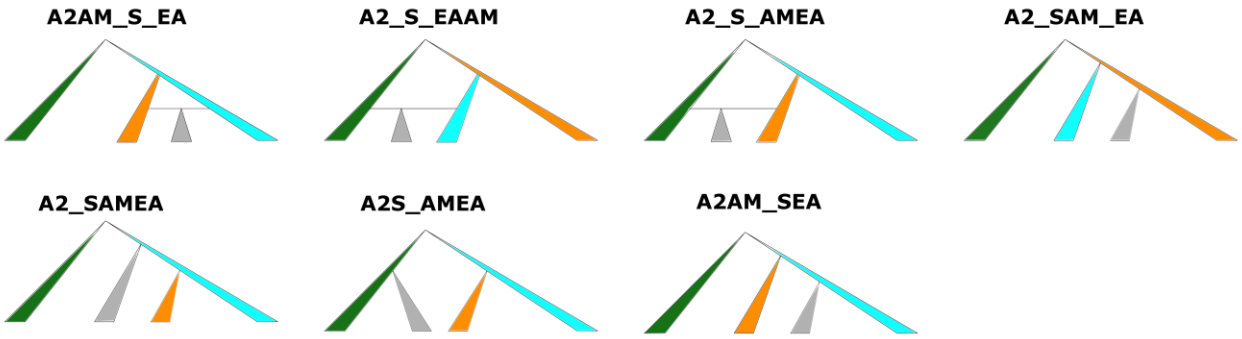
Fig.2.3. STRUCTURE analisys employing the Linkage Model at K=10

The blue component can be associated with hspEastAsia, and was mainly found at high frequencies Chinese and Southeast Asian samples. The light green component was particularly present in Oceanic strains, which suggest it can represent hspMaori. The Nepalese, Ladak and a portion of Malaysian samples were characterized by an orange component identifiable as hspIndia. hspLadak, instead, can be associated with the pink component found at high frequencies in Nepal and Ladak. The samples from the Americas where characterized by a purple component, which was also found at high frequencies in different eastern/western Siberian populations, thus representing hspAmerind. The strains from the Altai region, such as the ones found in Tuvan and Tubalar, showed high percentages of a cyan component, confirming the proposed hspAltai subpopulation. Through the entire Siberian region we could detect the presence of a red component, which could be associated with hspSiberia1, while the black component fitted the distribution observed for hpSiberia2/hspKet, here considered as

a single entity probably as a consequence of the high similarity already found by the DAPC. The dark green component instead can be localized in samples from Nents and Khanty, suggesting it could define hspUral. However, differently from the DAPC, we observe an additional subdivision, the yellow component, typical of the Nepalese samples and found at low frequencies also in the ones from Southeast Asia. We speculate that these could indicate some other level of substructure inside hspEastAsia.

Given the pattern of similarity showed by the DAPC, we developed seven ABC models to test the origin of the two most diffused new groups in Siberia, hspSiberia1 and hspSiberia2, including admixture and simple split scenarios involving the subpopulations most similar to them (Fig. 4A). The posterior probabilities obtained via logistic regression highlighted an admixed origin for both subpopulations, as hspSiberia1 resulted to have originate by a contact between hpAsia2 and hspAmerind (Fig. 2.4B) while hspSiberia2 stemmed from a mixing between hpAsia2 and hspEastAsia (Fig. 2.4C). The estimated time for the two admixture events was also very close, having happened around 2,800 generations ago for the event that gave rise to hspSiberia1 and around 3,000 generations ago for the one resulting in hspSiberia2. These inferences were supported by our ability to correctly re-estimate previously obtained divergence times, such as the one between hspAmerind and hspEastAsia (about 28-21,000 generations ago; Moodley et al. 2009), and mutation rate values comparable to the ones already published ($3.02\text{-}3.36 \times 10^{-7}$; Montano et al. 2015).

A



B

	A2AM_S_EA	A2_S_EAAM	A2_S_AMEA	A2_SAM_EA	A2_SAMEA	A2S_AMEA	A2AM_SEA
50,000	0.299	0.006	**0.652**	0.043	0	0	0
125,000	0.318	0.013	**0.629**	0.039	0	0	0
250,000	0.314	0.032	**0.606**	0.047	0	0	0

C

	A2AM_S_EA	A2_S_EAAM	A2_S_AMEA	A2_SAM_EA	A2_SAMEA	A2S_AMEA	A2AM_SEA
50,000	0.258	**0.719**	0.02	0	0	0	0.001
125,000	0.250	**0.727**	0.02	0	0	0	0.002
250,000	0.217	**0.757**	0.02	0	0	0	0.003

Fig.2.4. A) Models tested for determining the origin of hspSiberia1-2. Colors represent subpopulations included in the models: hpAsia2 (green), hspAmerica (orange), hspEastAsia (light blue) and the tested hspSiberia 1/2 (gray). Posterior probabilities for each model performed by ABC analysis under weighted multinomial logistic regression for hspSiberia1 (B) and hspSiberia2 (C)

We then extended our ABC analysis to hspKet, designing five models based on the similarities shown by this subpopulation in the DAPC and including also the previously inferred demographic patterns presented by the other groups of *H.pylori* in the dataset (Fig. 2.5A). The model postulating an admixture event between hspSiberia1 and hspAmerind received the highest support based on the obtained posterior probabilities (Fig. 2.5B) and the estimated time for this contact has been around 2,212 generations in the past. We also noted that the second best model, which obtained lower but comparable posterior probabilities, hypothesized the mixing of hspAmerind and hspSiberia2, instead of hspSiberia1, further highlighting the similarities between the two subpopulations.

A

B

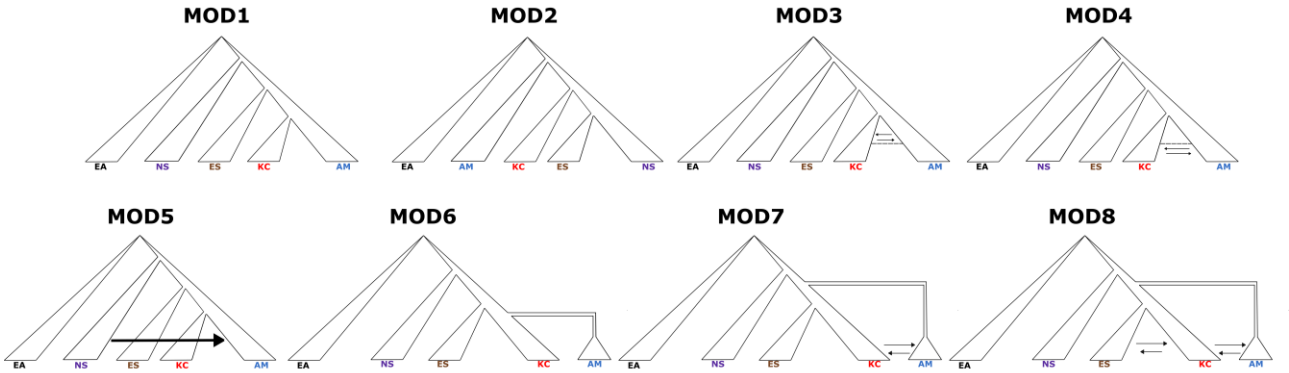
	A2S1AS2_K_EA	A2S1A_K_S2EA	A2S1_K_AS2EA	A2S1_K_EAS2A	A2S1_K_S2AEA
50,000	0.020	0.282	**0.456**	0.035	0.206
125,000	0.036	0.298	**0.441**	0.052	0.172
250,000	0.064	0.354	**0.326**	0.092	0.165

Fig.2.5. A)Models tested for determining the origin of hspKet. Colors represent subpopulations included in the models: hpAsia2 (green), hspAmerica (orange), hspEastAsia (light blue), hspSiberia 1 (gray), hspSiberia2 (pink) and the tested hspKet (light green). B) Posterior probabilities for each model performed by ABC analysis under weighted multinomial logistic regression

We also determined the time of origin of hspAltai using a demographic model involving its split from hspAmerind and considering hspEastAsia as outgroup. We estimated the median time of 3,455 generations ago for the formation of this subpopulation.

Finally, we attempted to model the human migration into the Americas using the retrieved *H.pylori* strains belonging to hspAmerind and, in order to do it, we developed eight models (Fig. 2.6A). The bottleneck model postulating a first separation of the American strains from the eastern/northeastern Siberian ones, with subsequent migration involving the formers and the ones present in Kamchatkan populations, received the strongest support (Fig. 2.6B). The median time estimated for this bottleneck to start was of 12,870 generations in the past, with a recovery of the American populations starting around 2,800 generations later. The effective population size (Ne) of the American populations at the end of the bottleneck was estimated to be extremely small (Ne =140) that, over the course of 10,000 generations, underwent an exponential growing phase resulting in the current estimate Ne of around 83,863. We also detected low levels of migrations between Kamchatka and the Americas (0.001 immigrants per generation), while no appreciable back migration was found.

A



B

	MOD1	MOD2	MOD3	MOD4	MOD5	MOD6	MOD7	MOD8
50,000	0.286	0.020	0.001	0.044	0	0	**0.57**	0.079
125,000	0.292	0.018	0.002	0.046	0	0	**0.571**	0.071
250,000	0.417	0.027	0.007	0.071	0.001	0	**0.421**	0.055

Fig. 2.6. A) Models tested for the colonization of the Americas. Acronyms represent the geographical subdivisions of hspAmerind: north Siberia (NS), east Siberia (ES), Kamchatka (KC), and the Americas (AM). All models included hspEastAsia (EA) as outgroup. B) Posterior probabilities for each model performed by ABC analysis under weighted multinomial logistic regression

Discussion and conclusion

Migrations in Siberia and the Americas have always been a contentious topic in human population history (For a review see Skoglund and Reich 2016). In this work we studied population movements inside these regions employing the bacterial bioproxy *Helicobacter pylori*, which has previously proven useful in similar settings (for a review see Moodley and Linz 2009). When we compared the newly obtained 395 bacterial sequences from 16 Siberian human populations with the previously published Eurasian strains, we located five unrecognized clusters that could be assigned to new *H.pylori* subpopulations. The distribution of these groups highlighted a peculiar geographical pattern inside Siberia, where the human populations from the central region showed high frequencies of three new bacterial subpopulations: hspSiberia1, hspSiberia2 and hspAltai. The eastern and western regions were instead characterized by high frequencies of hspAmerind, a subpopulation that was, until now, been identified only in Native Americans. When the demographic history of hspSiberia1

and hspSiberia2 was investigated by ABC coalescent modelling, an admixed model between either hspAmerind (for hspSiberia1) or hspEastasia (for Siberia2) and hpAsia2 obtained the highest posterior probabilities. A simpler tree-like model separating hspAltai from hspAmerind provided, instead, a good fit for the modality of origin of this third group. All of these events were estimated to have happened less than 3,500 generations ago. It is important to note that the generation time of *H.pylori* should not be interpreted as the replication of the single bacteria, but in terms of the number of secondary infections each primary infection produces (Stearns and Koella 2008). In light of this assumption the proposed generation time of 1 generation per year for *H.pylori*, which account for the dual modality of transmission present in this bacteria, has received support in previous publications (e.g. Montano et al. 2015). This generation time allows the translation of our ABC estimates directly into years, obtaining that the most diffused *H.pylori* subpopulations in central Siberia share a recent origin no further than 3,500 years ago, and later spread to their current locations. This result support the hypothesis of a depopulation of central Siberia during the LGM, with a recent repopulation of this region by human groups coming from the south (Pugach et al. 2016). This hypothesis is further strengthened by the presence of a specific subpopulation, hspAltai, found at high frequencies in Turkic speaking populations, which migrated in their current location from a "Inner Asian Homeland" situated across south Siberia and northern Mongolia (Yunusbayev et al. 2015). The presence of hspAmerind in human populations located in the eastern and western fringes of Siberia, together with the estimate obtained for the time of its origin (> 20,000 years ago) indicate that this subpopulation could represent the oldest group of *H.pylori* characteristic of the Siberian region, which was carried by human populations during their migration to locations with favorable conditions at the LGM.

We also discovered the existence of a subpopulation similar to the previously described hspSiberia1 and hspSiberia2, called hspKet, which was specific of a known linguistic isolate. The ABC modelling procedure highlighted an admixed origin for this subpopulation, involving a contact between hspSiberia1 and hspAmerind about 2,200 years ago, possibly as a consequence of the repopulation of central Siberia. The Ket population today number around 1,200 individuals and represented, until

their forced settlement in 1930, the last nomadic hunter-gatherers of northern Asia (Vajda 2009). The existence of hspKet can be associated with the strict endogamy practiced by this population until the 17th century that, together with the vertical transmission of *H.pylori* in family members, could have promoted the differentiation of this subpopulation.

Finally, we used the newly sequenced strains belonging to hspAmerind to understand human migrations into the Americas. The highest support was obtained by a model in which the American populations definitely diverged from the ancestors of eastern/northeastern populations ~12,800 years ago, undergoing a following bottleneck that lasted around 2,800 years. This pattern can be associated with the final flooding of the Beringia land-bridge ~11,500 years ago (Hoffecker and Elias 2007), which definitively isolated human populations in the Americas from the ones in Siberia. Moreover, the existence of a population bottleneck in the American strains of *H.pylori* mirrors the results obtained by studies on human mitochondrial DNA (e.g. Tackney et al. 2015) and genomic data (e.g. Reich et al. 2012). We do not find traces of subsequent migrations into the Americas, such as the one involving the Paleoeskimo culture that happened ~4,500 years ago (Rasmussen et al. 2010). However, we note that this limitation was probably caused by the restricted availability of Native American strains, not the main focus of this study, and further works involving extensive sampling may help to shed light on more recent population movements in this region.

In conclusion, *Helicobacter pylori* proved to be an invaluable tool to reconstruct both recent and ancient human migrations, providing a resolution comparable to the one obtainable with large, and more expensive, human genomic studies.

Chapter 3. MOVEMENT OF LANGUAGES AND GENES

Coevolution of genes and languages

Amongst the multiple cultural tools employed to understand past human movements, the analysis of language-relatedness has been one of the most commonly associated with population genetics (e.g. Luca Cavalli-Sforza et al. 1988; Sokal 1988; Barbujani and Pilastro 1993; Creanza et al. 2015) . The first idea of a possible parallelism between genes and languages was formulated by Charles Darwin when, in his *On the origin of species by means of natural selection*, he proposed that the relationship between human populations, in the form of a phylogenetic tree, could provide the best way to understand the affiliation of different languages (Darwin 1859). The first attempts at a complete worldwide analysis of Darwin's intuition was performed more than a century later by population geneticists, when obtaining allele frequency data from several world populations became a feasible task. In these first works it was advocated a large scale correspondence between the distribution of classical markers (e.g. blood groups, serum proteins) and proposed linguistic family classification, such as the one provided by Ruhlen (1987) (Luca Cavalli-Sforza et al. 1988; Sokal 1988). The supporters of this coevolution theory argues that genes and languages are transmitted in comparable ways: genes are obligatory passed on from parents to their offspring while languages tend to be inherited in a vertical line as well (Pakendorf 2014). Moreover, common population processes such as isolation can promote the diversification of both languages and genes with respect to the neighbouring populations. Conversely, secondary contact between two or more populations can promote the homogenization of their genetic makeup by means of gene-flow, while also allowing linguistic contact and the consequent loaning of words between languages.

Different works have proven that, indeed, linguistic and genetic diversity correlate, at least when analysed at a local/regional scale (Belle and Barbujani 2007; Longobardi et al. 2015). However, when studies focused on broader patterns, they obtained contrasting results. A recent analysis of 2,082 phonemes in 246 worldwide populations found that both genetic and phonemic diversity correlated

with geographical distances between sampling regions, meaning that languages that were spatially closer were also phonetically similar, independently on the linguistic family to which they belonged (Creanza et al. 2015). Moreover, there can also be differences in linguistic and genetic correlations depending on the analysed ancestry. When uniparentally inherited markers where analysed in Burkina Faso, only the paternally inherited Y-chromosome diversity showed a structure that reflected the language families present in the area, while mtDNA presented no differentiation between different linguistic groups (Barbieri et al. 2012). Finally, as outlined in the following case study, works on the concordance between linguistic and genetic diversity are also hampered by a lack of a universally recognized linguistic markers, making Darwin's century old question still an open challenge.

The following case studies focus on the analysis of genetic and linguistic diversity. Additional works in which I have participated on the subject are:

Kutanan, W., Kampuansai, J., **Brunelli, A**., Ghirotto, S., Pittayaporn, P., Ruangchai, S., Schroder, R., Macholdt, E., Srikummool, M., Kangwanpong, D., Hubner, A., Arias Alvis, L., Stoneking, M. (2017). New insights from Thailand into the maternal genetic history of Mainland Southeast Asia. bioRxiv, 162610. (Under review at the European Journal of Human Genetics)

Kutanan, W., Kampuansai, J., Srikummool, M., Kangwanpong, D., Ghirotto, S., **Brunelli, A.**, Stoneking, M. (2017). Complete mitochondrial genomes of Thai and Lao populations indicate an ancient origin of Austroasiatic groups and demic diffusion in the spread of Tai–Kadai languages. Human genetics, 136(1), 85-98.

Case study: Grammars and genes in the history of Old World migrations

Outline of the research

Comparing genetic and linguistic diversity has the ability of shed light on both the movement history and cultural transmission of human populations. A common practice to compare these two dissimilarities involve the application of distance matrix and correlation analyses such as Mantel's test (Mantel 1967). However, while genetic distances are built on empirically assessed markers, linguistic ones are often constructed using controversial phylogenetic classifications. Different methods have been proposed and applied to circumvent this issue. The use of linguistic cognates – set of words that can be traced to a common ancestor – has provided empirically built distances to be compared with genetic ones (e.g. Balanovsky et al. 2011; de Filippo et al. 2012). While being an advancement with respect to arbitrarily built classifications, this method has been shown to being reliable only when short time frames are considered, as sound changes and replacement of words reduce the number of cognates with time (Pakendorf 2014). Moreover, accidental similarities tend to emerge, due to the combination of arbitrariness of lexical variation with general constraints on possible phonological systems. Therefore, the reliable inference of distant (across linguistic families) relationships has, to this day, proved to be difficult.

In this work we re-evaluate the gene-language correlation issue in Eurasia through a new type of linguistic characters, i.e. universally definable and discrete grammatical differences (syntactic parameters) recently provided by the Parametric Comparison Method (PCM; Longobardi and Guardiano 2009). We built a database of such polymorphic grammatical loci, which deductively derive several thousands of syntactic phenomena, and obtained reliable linguistic distances between 28 Eurasian populations speaking languages belonging to different linguistic families. These distances were compared against those defined through ~300k whole-genome autosomal SNPs from 1,303 individuals belonging to the same populations.

In principle, history may distribute languages and genes either in close correlation (through the same demic movements, or just along the same geographical routes) or arbitrarily: in the latter case, grammars could be transmitted with unremarkable gene movements, or conversely not carried along by populations expanding into different territories. Therefore, in order to make the gene/language congruence question adequately testable, we first tested for evidence of a gene-language parallelism inside Eurasia. We then quantifying how much of it depended on common demic processes, jointly shaping genetic and linguistic diversity.

Materials and Methods

We built a database of polymorphic grammatical loci (2100 parametric states, 75 each for 28 languages from 9 traditionally irreducible Old World linguistic phyla), which deductively derive several thousands of syntactic phenomena, and computed distances between the sampled languages employing the PCM. This method formalizes the availability or unavailability of certain syntactic properties as binary syntactic characters, which in principle enable us to precisely calculate the syntactic distance between any two languages, and to generate and validate evolutionary representations for such language sets (Longobardi and Guardiano 2009). Properties compatible with a positive or negative value of a certain parameter (represented as + or – in Table A below, figure 2), but which are deducible from the states of one or more other parameters, are not represented as a potential comparandum, because such a property is not the result of an independent historical process, i.e. a distinct parametric change. Thus, certain parameters are considered to be relevant for a language X, and thus "active" in X, only when a number of other parameters are in a certain state: otherwise the state of such parameters is 0, and they will not be taken into consideration when calculating the syntactic distance of X from other languages. Therefore, out of the total number of parameters used, any two languages may only be compared with respect to the subset of parameters for which neither language has a 0 setting. The syntactic distance of two languages was taken to be the ratio of the number of their differences to the cardinality of such a subset (i.e. to the sum of identities and

differences) and was defined as the "Normalized Hamming (or Jaccard) distance" of the corresponding strings of parameter values: $D_{SYN} = d/(i+d)$.

Genome-wide high-density SNPs data for populations matching the available languages were collected from the literature (Tab. 3.1.1). Populations were selected with the aim at maximizing the correspondence between genetic points and collected languages. When multiple matches were found for the same language (e.g. Jordanians, Palestinians, Saudis and Syrians for Arab), genetic samples from these populations were pooled together, after having checked for their genetic homogeneity. For those cases in which a genetic sampling point did not exactly overlap with the area of spoken language, the best proxy available was used (e.g. Eastern Siberia Eskimos for Inuit, Gujaratis for Marathi, Southern Han Chinese for Cantonese, Mandenka for Wolof. Genomic data were merged and filtered using the software PLINK 1.07 (Purcell et al. 2007) to include (i) only single-nucleotide polymorphisms (SNPs) with genotyping success rate higher than 98% as well as minor allele frequency of at least 1%, and (ii) only individuals showing less than 1% of missing genotypes. In addition, outlier individuals and closely related samples (kinship coefficient PiHat $\geq$ 25%) were excluded to avoid possible biases due to biological relatedness. The merging and filtering procedure resulted in a final dataset of 1303 individuals genotyped on different Illumina bead arrays for a common set of 284,677 loci. Genetic distances (dGEN) based on Weir and Cockerham Fst (Weir and Cockerham 1984) values were calculated with the 4P software (Benazzo et al. 2015). In order to check for possible biases due to differences in sample sizes, Fst values were calculated for both the whole dataset and by randomly subsampling 20 individuals per population.

Language	Population Sample	N(Tot)	N(Rand20)	Reference	Geo. Point	Latitute	Longitude
Arabic (Ar)	Jordanians, Palestinians, Saudis, Syrians	102	20	Behar et al 2010, Li et al 2008	Riyadh	46.724100	24.711660
Basque Central (cB	French Basque	24	20	Li et al 2008	Bilbao	-2.933334	43.266667
Bulgarian (Blg)	Bulgarians	13	13	Yunusbayev et al 2011	Sofia	23.323638	42.69756
Buryat (Bur)	Buryats	42	20	Rasmussen et al 2010, Cardona et al 2014	Ulan-Ude	107.600000	51.83333
Cantonese (Can)	Han Chinese South (CHS) *	92	20	HapMap	Hong Kong	114.133333	22.18333
German (D)	Germans	13	13	Yunusbayev et al 2015	Berlin	13.408056	52.51861
English (E)	British individuals from England and Scotland (GBI	94	20	HapMap	London	-0.127500	51.50722
Inuktikut (Inu)	Eskimo *	13	13	Cardona et al 2014	Cape Dezhnev	-169.651944	66.07917
Estonian (Est)	Estonians	15	15	Raghavan et al 2014	Tallinn	24.745278	59.43722
Farsi (Far)	Iranians	19	19	Behar et al 2010	Tehran	51.416667	35.68333
Finnish (Fin)	HapMap Finnish individuals from Finland (FIN)	99	20	HapMap	Helsinki	24.937500	60.17083
French (Fr)	French	28	20	Li et al 2008	Paris	2.351944	48.85667
Greek (Grk)	Greeks	20	20	Behar et al 2013	Athens	23.716667	37.96667
Hebrew (Heb)	Middle Eastern Jews	14	14	Behar et al 2010	Jerusalem	35.216667	31.78333
Hindi (Hi)	Upper Caste Uttar Pradesh (Brahmins and Kshatriy	15	15	Metspalu et al 2011	New Delhi	77.200000	28.600000
Hungarian (Hu)	Hungarians	19	19	Behar et al 2010	Budapest	19.050278	47.47194
Italians (It)	North Italians, Toscan individuals (TSI)	120	20	Li et al 2008, HapMap	Rome	12.482778	41.89306
Japanese (Jap)	Japanese, Japanese individuals (JPT)	129	20	Li et al 2008, HapMap	Tokyo	139.691700	35.68951
Marathi (Ma)	Gujarati India individuals from Texas (GIH) *	97	20	HapMap	Mumbai	72.825833	18.96472
Mandarin (Man)	Han Chinese, Han Chinese in Beijing (CHB)	147	20	Li et al 2008, HapMap	Beijing	116.391389	39.90556
Pashto (Pas)	Pathans	23	20	Li et al 2008	Khyber Pass	71.203940	34.075700
Polish (Po)	Poles	16	16	Behar et al 2013	Warsaw	21.008433	52.23230
Romanian (Rm)	Romanians	16	16	Behar et al 2010	Bucarest	26.096111	44.43556
Russian (Rus)	Russias	25	20	Li et al 2008	Moscow	37.617778	55.75167
Serbo-Croat (SC)	Croats	20	20	Behar et al 2013	Zagreb	15.966667	45.800000
Spanish (Sp)	Iberian populations in Spain (IBS)	46	20	HapMap	Madrid	-3.683333	40.43333
Turkish (Tur)	Turks	19	19	Behar et al 2010	Ankara	32.849000	39.953000
Wolof (Wo)	Mandenkas *	22	20	Li et al 2008	Dakar	-17.450000	14.68333
		1303	**513**				

* Non-perfect matches between linguistic sampling location and genetic proxy available

Tab.3.1.1 Languages and populations sampled in the study

Principal Components (PCA) and Admixture analyses were used to explore the genetic structuring pattern and to test relationships among populations. The PCA was carried out with the smartpca program of the EIGENSOFT package (Patterson et al. 2006). For admixture analysis the unsupervised ancestry-inference method of the ADMIXTURE software (Alexander et al. 2009) was used. Since ADMIXTURE algorithm requires unlinked SNPs, the combined set of markers was thinned by excluding SNPs in strong LD ($r2 > 0.1$) within a sliding window of 50 SNPs, advanced by 10 SNPs at a time. The pruned dataset consisted of 57,772 autosomal markers. We ran series of admixture analyses, assuming an increasing number of "ancestral populations" from K=2 through 10. The predictive accuracy of each model at a given K was evaluated by means of the cross-validation (CV) index returned by the ADMIXTURE software.

Previously, the potential for geographical distance to act as a confounding factor in the correlation between genetic and linguistic distances was accounted for by controlling for the great circle distance (GCD) between the population centres involved (Longobardi et al. 2015). In order to account for the

likely migration routes between the main geographic regions, here the GCDs between population pairs were constrained using 4 obligatory waypoints (Ramachandran et al. 2005). However, for points spread across three continents, GCDs may not be suitable; large obstacles such as oceans and mountain ranges mean that the distance populations would have to travel from one point to another may far exceed the direct great circle distance between those points. These obstacles can be interpreted as resistances to the free spread of individuals during migrations between two points in a landscape. Least cost path (LCP) distances measure the optimal route between two locations, defined here as the one that passes between points with the minimum accumulation of resistances or 'costs' (Howey 2007). However LCP modeling assumes that the traveling individual has complete knowledge of the landscape he is moving through thus being able to select the route offering the lowest resistance based on this knowledge (McRae and Beier 2007). Circuit modeling attempts to account for this limitation by employing a concept of resistance distance which incorporates both the minimum movement distance (or cost) and the availability of alternative pathways (Mcrae 2006). In order to best characterize the geographical distances between the populations in our dataset we constructed a cost-surface raster for LCP and resistance (circuit-modelling conceived, CO) distances. We obtained values for elevation from WorldClim database and vector files containing major rivers and coast outlines from NaturalEarth. We merged those data with the function mosaic of ArcMap 10.1 to create a cost-surface raster that was later reclassified following the guidelines provided in Tassi et al. 2015 (Tassi et al. 2015). Computations of LCP were done using the function *costDistance* from the package *gdistance v.1.1-9* (van Etten 2012). To compute Resistance (CO) distances we utilized Circuitscape v4.0 (Mcrae 2006). We used the coordinates of the 28 worldwide populations analysed to compute a pairwise comparison between all populations, allowing the program to connect eight neighbouring cells.

The correlations between genetic, geographic and linguistic distance matrices were estimated by performing standard Mantel and Partial Mantel permutation tests, according to the *mantel* and

mantel.partial functions implemented in the R software package *vegan* (Mantel 1967; Oksanen et al. 2008). The statistical significance of the results was empirically tested over 10,000 permutations.

Results

We first ascertained that the PCM also classifies the non-IndoEuropean languages of the sample correctly, to the extent these families are established (Fig. 3.1.1).

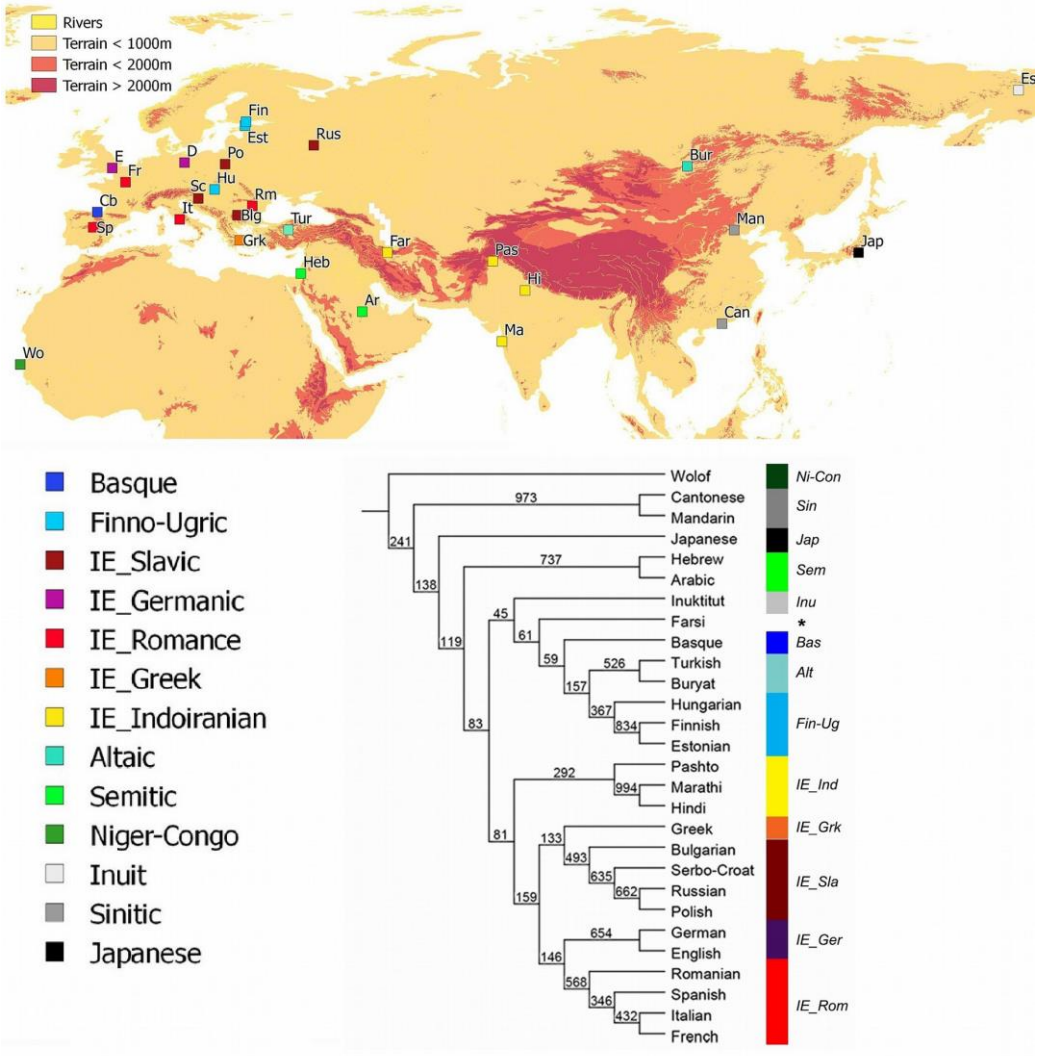


Fig. 3.1.1. Geographic distribution of the 28 analyzed languages/populations displaying the resistance raster used for computations of least-cost path/commuting distances based on the parameters reported in the legend at the top-left and phylogenetic tree of corresponding languages based on the 75 syntactic parameters. Color codes identify population affiliation to corresponding language families.

Then, we compared syntactic and genetic distances of corresponding populations (Tab. 3.1.2). The extent and significance of the gene-language correlation obtained (r = 0.53) constitutes *prima facie* evidence for a strong association between genomic and linguistic diversity. In order to control for the likely effect of geography on the gene-language relationship, we computed five geographical distance matrices: the straightforward Great Circle Distance (GCD), and metrics accounting for territory complexity and possible migration routes: Roads Map, GCD with waypoints, Least Cost Path, Resistance (RE). A partial Mantel test yielded a correlation between genetic and linguistic diversity stable and significant with every geographic distance considered (r ranging between 0.26 and 0.35). At this density of analysis, languages correlate with geography only as a byproduct of their congruence with genetics, while the gene/language congruence is largely independent of geography.

Distances	**Mantel Corr (r)**	**P-value**
d_{GEN}-d_{SYN}	0.5286	0.0001
d_{GEN}-d_{GEO}	0.6882 - 0.9117	0.0001
d_{SYN}-d_{GEO}	0.4352 - 0.4751	0.0001
d_{GEN}-$d_{SYN(GEO)}$	0.2641 - 0.3508	0.0044
d_{GEN}-$d_{GEO(SYN)}$	0.5995 - 0.8844	0.0002
d_{SYN}-$d_{GEO(GEN)}$	-0.0197 - 0.1158	0.4259

Tab.3.1.2. Mantel and partial Mantel correlations for genetic and linguistic similarity matrices

Next, we examined analytically the correlation of syntactic and genetic distances for each of the 28 populations (Fig. 3.1.2). In 21 cases (75%), the gene-language congruence is clearly confirmed; populations speaking similar languages resemble each other genetically. In 7 cases (Turkish, Farsi, Basque, Japanese, Hungarian, Inuktitut and Wolof; 25%) the gene-language correlation is statistically non-significant, with r-values on either side of zero.

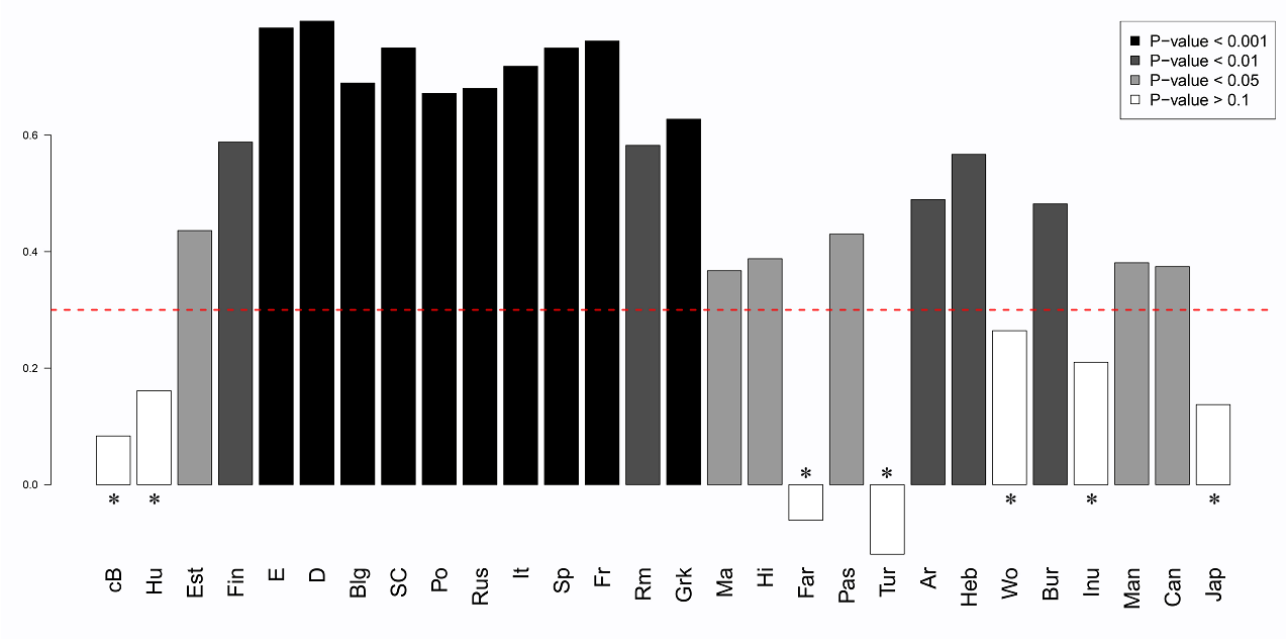


Fig.3.1.2. Gene/language correlation values (Spearman rank test) for each of the 28 populations. Bar heights account for the correlation extent, while color codes represent the corresponding significance degree as detailed in the legend at the top right. Statistically non-significant correlations are marked with an asterisk (*).

In order to further scrutinize this discrepancy, we conducted several population genomics analysis on our populations. A PCA on the whole genomic dataset distributes the considered populations according to geographical patterns, in agreement with the common expectations of population genetics (Fig.3.1.3A and B). In particular, PC1 (7.10%) differentiates the East Asians vs. West Eurasians, while PC2 (1.65%) separates the African group (Wolof) from the rest of populations (Fig. 3.3A). Once one of the two outliers, i.e. Wolof, is removed (Fig. 3.1.3B), the West Eurasian variability is stretched along a gradient of genomic variation ranging from India to Europe, through the Middle East. In this context, present-day Iranians are separated from their IE linguistic neighbours (Pashto, Pas; Hindi, Hi; Marathi, Ma), and group with (Altaic and Semitic) speakers from the Middle East (Turkish, Tur; Arabic, Ar; Hebrew, Heb). Mixing proportions estimated by the ADMIXTURE software (Fig. 3.1.3C) confirm the results described above, identifying ancestry groups that largely correspond to those observed in previous analyses: Wolof (dark-green), Inuit (light-grey), East Asian (grey) and three West Eurasian components, which in turn identify particular geographic domains,

i.e. India (yellow), Middle East (green) and Europe (blue-variants). For classification criteria from K=6 through 10 (that are the best predictors of population genetic structure according to the distribution of cross-validation errors, the considered Iranian (Farsi-speaking) population appears as a 'blend' whose most important 'ingredients' are Indian (~22%) and Middle Eastern (~55%) genetic ancestries. In particular, Indian genetic component is highly present in Pas (~55%), Hi (~74%) and Ma (~90%), while the Middle Eastern one reaches average frequency of 76% in Ar, 73% in Heb and 50% in Tur.

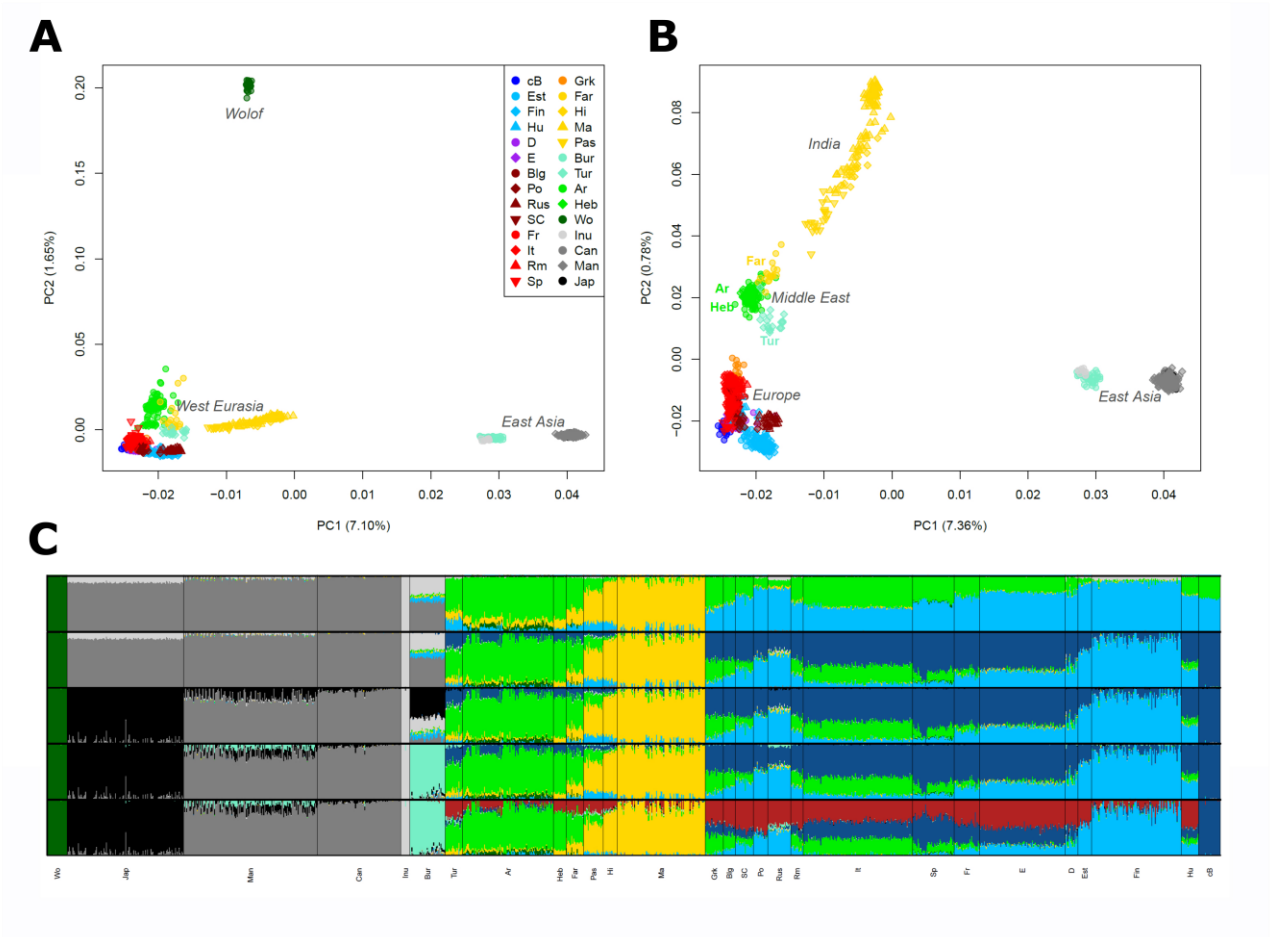


Fig. 3.1.3. Genetic structure of analysed populations. Projection of the first two principal components (PCA) including (A) and excluding (B) Wolof. Each point (corresponding to one individual) is represented and color-coded according to its linguistic affiliation as reported in the legend of the plot. (C) Population structure inferred by ADMIXTURE analysis of autosomal data for the most significant values of K from 6 (top) to 10 (bottom). At any K, each individual is represented by a vertical bar of genetic components, color-coded according to the K reconstructed ancestral populations. Individuals are grouped and labeled at population/language level and ordered by corresponding language families.

Discussion and Conclusion

Our experiments revealed a high language/gene correlation in Eurasia, which, furthermore, once controlled for geography, turned out 5 to 7 times higher (and statistically more significant) than previously discovered using other linguistic variables (Creanza et al. 2015). Thus, we corroborate Darwin's hypothesis that languages and genes across the Old World have diversified together, i.e. through robust demic movements: we only found only few exceptions, in all of which language features have travelled without massive gene displacement, but never the opposite (Fig. 3.2). We analyzed these cases individually. Wolof's syntactic distances from all other sampled languages are high, but the genetic ones closest to the available populations (Mandenkas) are proportionally even higher. High genetic distances for Sub-Saharan populations are indeed expected (Tishkoff and Williams 2002; Gurdasani et al. 2015); the disproportion between genetic and syntactic distances interestingly suggests that grammatical variation is more constrained than the biological one, which theoretically may be due to strict universal conditions on possible grammars (Chomsky 1993), on broader cognitive capacities (Chomsky 2005), on their chance diversification (Anti-Babelic Principle) (Guardiano and Longobardi 2005), or to their combination. The Inuktitut case goes in the opposite direction: gene-language mismatch is probably due to the greater geographical discrepancy between the language most accessible for sampling (Inuktitut, Eastern Canada) and the nearest genetic proxy available (North-Eastern Asia), thus overstretching the linguistic distances with respect to the genetic ones. Crucially, the remaining five cases of weak correlation can all be reduced to the same pattern, so called 'elite dominance': minor genetic inputs with relevant cultural changes (Renfrew 1992). Two salient cases of this pattern, Hungary (Longobardi et al. 2015) and Turkey (Di Benedetto et al. 2001) underwent a complete replacement of autochthonous languages by small groups of conquerors from distant territories (9 and 11-15 5centuries, respectively). In our sample Turks, like Hungarians in Europe, are genetically indistinguishable from the neighbouring populations (Fig. 3.1.3) and do not bear significant resemblance to the other Altaic-speaking population examined (Buryats). The Basque case exhibits a reversal of the same elite dominance

model: anciently, genetically similar populations surrounding Basques spoke languages related to Basque (e.g. ancient Aquitanian), but then adopted newcomers' languages (Celtic, Latin/Romance), while their genetics was more marginally affected by the latter populations (Günther et al. 2015). The Basques, instead, retained their language, whereas their neighbouring populations drifted away from them linguistically more than genetically. Next, in the case of Farsi, we detected plausible traces of Arabic (about 24%, 8/33 of parameters susceptible to change) and Altaic (~21%, 7/33) syntactic interference. Altaic linguistic interference was not paralleled by significant introgression of genes, the present-day Iranian gene-pool being instead similar to core Near Eastern components (Fig. 3.1.3). Given that significant genetic influence specifically attributable to Altaic populations was not observed in Turkey, the gene/language mismatch in Persia resembles the elite dominance in Anatolia, though with only partial replacement of linguistic traits (the language remains IE in other properties), sufficient, however, to place Farsi outside the IE subtree (Fig. 3.1). Furthermore, the Japanese are genetically closer to our other East Asian populations (Chinese and Buryats) than expected from syntactic distances, which reflect the uncertain/isolate status of Japanese in standard linguistic taxonomies. The genetic make-up of the modern Japanese apparently stems from the admixture of two different components, the Pleistocenic populations that reached Japan 38,000–37,000 BP, later giving rise to the Jomon culture (12,000 BP), and Yayoi farmers, who arrived around 2,500 years ago (Günther et al. 2015). The situation remains compatible with our prediction that one of the migrations must have had more linguistic than genetic effects, pending further investigation of East Asian languages/populations. Finally, removing the cases (Turks, Basques, Persians and Hungarians) independently justified through elite dominance, we recalculated the correlations for the remaining 24 populations, obtaining expectedly higher and significant values (d_{GEN}-d_{SYN}: 0.66, p-value: 0.0001; d_{GEN}-d_{SYN}(GEO): 0.32-0.50, p-value: 0.0065-0.0001).

Thus, syntax-based experiments provide a positive answer to both questions derived from Darwin's congruence hypothesis: across Eurasia, genes and grammars have 1) prevailingly diffused along similar routes and 2) mostly through the same demographic events. The parallelism is only disrupted

by few cases where grammars have spread without a conspicuous corresponding gene-flow, with no evidence for the reverse trend (massive genetic introgressions without relevant grammatical changes). The scenario advanced here proved robust enough to withstand possible complications concerning the outlying populations/languages (Wolof and Inuktitut) or geographically non-homogeneous coverage. Characters provided by a formal theory of syntactic variation, can probe our past more deeply than other language traits and promise to complement genomic approaches to demographic history with previously unavailable linguistic insights

Case study: Y chromosomal evidence on the origin of northern Thai people

Outline of the research

The area of northern Thailand situated in proximity to southern China, northern Myanmar and northern Laos hosts several ethnicities who can be linguistically classified in four groups: Austroasiatic, Tai-Kadai (TK), Hmong-Mien and Sino-Tibetan. The languages belonging to the Austroasiatic subfamily Mon-Khmer (MK) are spoken today by populations, e.g. Lawa (LW) and Mon (MO), historically and archaeologically recorded as the native inhabitants of this area before the arrival of Tai-Kadai people form southern China 2,000 years ago (Condominas 1990; Penth 2000; Pittayaporn 2014). Other ethnicities, such as the Hmong-Mien (e.g. Hmong) and different Sino-Tibetan groups (e.g. Karen), migrated from nearby countries to the mountainous areas of northern and western Thailand no more than 200 years ago (Grundy-Warr et al. 2003; Schliesinger 2015a). In addition, other recent migrations from southern China and/or northern Myanmar are recorded as

involving several Tai-Kadai groups such as Lue (LU), Khuen (KH), Yong (YO) and Shan (SH) (Schliesinger 2015b). Linguistic similarity between populations has often been used to disentangle patterns of relationship, under the assumption that a common language implies a common origin (Luca Cavallisforza et al. 1988; Sokal 1988). However, genetic similarities between different populations are often more complex than expected from linguistic data due to the effect of processes such as drift and migration (Kutanan et al. 2014). The mountainous area of northern Thailand consists of several river plains surrounded by mountains, which continue from the Shan Hills in bordering Myanmar to Laos. In this region, different geographical areas are often occupied by different ethnolinguistic groups. The hill tribes, such as LW and SH speaking groups, currently live on the mountain where mobility is quite limited while, on the other hand, population such as MO and different TK speaking groups inhabits the well-connected lowland regions. Geography, acting in addition to cultural/linguistic isolation, might have been an influential factor in determining the divergence by inbreeding of these populations (Kutanan, Kampuansai, Fuselli, et al. 2011). Due to the combined effect of these different processes, northern Thailand is an extremely interesting site for studies on human population genetics. Among the multi-ethnic groups in northern Thailand, the Khon Mueang (KM) are the most represented, with a total number of individuals reaching 6 million (Lewis et al. 2009). KM is the name with which local northern Thai people, possibly the Yuan (YU), call themselves, and refers more to a past social and political category rather than a distinct population (Bhumisak 1990; Charoenmuang 2001). Linguistically speaking, the KM's language is similar to that of the YU, which is classified as belonging to the TK family.

It is widely accepted that genetic markers can be efficiently used to reconstruct past populations' history and interactions. Even after the rise of whole-genome techniques, uniparental markers continue to be widely used in population genetics. Due to their specific patterns of inheritance, the information provided by the Y-chromosome and by the mitochondrial DNA (mtDNA) allows for in depth analyses of sex-specific patterns of population history and demography. These data can be used to locate asymmetric contributions of male and female individuals to a migration, an event of

admixture and generally to the pattern of gene-flow in a certain area. A previous study, based on comparisons among autosomal Short Tandem Repeat (STR) loci, suggested an admixed origin for KM, with a higher contribution from the TK than from the MK groups (Kutanan, Kampuansai, Colonna, et al. 2011). On the other hand, a coalescent modelling using mtDNA genome data indicated southern China as the most probable origin of the KM, without admixture with LW groups in northern Thailand (Kutanan, Kampuansai, Brunelli, et al. 2017). Y chromosomal data of KM and of their linguistic and geographic neighbors in northern Thailand have been reported, but they have been limited to STR markers (Kutanan, Kampuansai, Nakbunlung, et al. 2011; Kutanan, Kampuansai, Fuselli, et al. 2011). Here, we investigated newly generated data of single nucleotide polymorphism (SNP) on Y chromosome along with previously published Y-STRs and mtDNA data. We used a combination of classical statistical analyses and model based simulations to shed light on the past population dynamics linked to the origin of KM of northern Thailand.

Material and Methods

A sample of 519 males belonging to 24 populations from northern Thailand was subdivided into three groups: Khon Mueang (KM), Mon-Khmer (MK) and Tai-Kadai (TK) (Table 3.2.1). The DNA samples were obtained from our previous studies (Kutanan, Kampuansai, Nakbunlung, et al. 2011; Kutanan, Kampuansai, Colonna, et al. 2011). We genotyped a total of 104 binary polymorphisms on the Y chromosome according to iPLEX Assay (Jurinke et al. 2005) using a Sequenom Mass ARRAY iPLEX Platform (Sequenom, Hamburg, Germany). To assign specific Y chromosomal haplogroup or Y lineage to each individual, we employed a phylogenetic hierarchical approach based on Y-DNA Haplogroup Tree YSOGG 2016 (International Society of Genetic Genealogy 2016). The use of human subjects for this study was ethically approved by Chiang Mai University, Thailand.

In order to investigate the origin of the KM populations from both maternal and paternal perspectives using a simulations based analysis, we assembled two datasets in which we collected genetic information for 17 Y-STRs loci: DYS19, DYS388, DYS389a, DYS389b, DYS390, DYS391,

DYS392, DYS393, DYS426, DYS434, DYS435, DYS436, DYS437, DYS439, DYS460, DYS461, Y-GATA-A10 and mtDNA HVR-I sequence (Table 3.2.1). A total of 536 genotypes for Y-STR and 1,109 for mtDNA-HVRI sequences were retrieved from literature (Kutanan, Kampuansai, Fuselli, et al. 2011; Kutanan, Kampuansai, Nakbunlung, et al. 2011).

			HVR-I				Y-STR				Y-SNP	
Population	Code	Group	N	*h*	sd	No. of haplotypes	N	*h*	sd	No. of haplotypes	N	No. of haplogroups
Khon Muang 1	KM1	KM	50	0.967	0.0121	31	21	1	0.0147	21	21	7
Khon Muang 2	KM2	KM	41	0.974	0.0103	25	16	0.9917	0.0254	15	16	5
Khon Muang 3	KM3	KM	36	0.967	0.0141	22	15	1	0.0243	15	15	7
Khon Muang 4	KM4	KM	52	0.98	0.0085	36	29	1	0.0091	29	29	9
Khon Muang 5	KM5	KM	43	0.933	0.0222	22	20	1	0.0158	20	21	8
Khon Muang 6	KM6	KM	45	0.954	0.0193	29	22	1	0.0137	22	22	12
Khon Muang 7	KM7	KM	46	0.934	0.0201	21	23	0.9921	0.0154	21	23	6
Khon Muang 8	KM8	KM	45	0.961	0.0143	26	22	0.9913	0.0165	20	22	11
Khon Muang 9	KM9	KM	45	0.932	0.028	25	22	0.987	0.0201	20	22	7
Khon Muang 10	KM10	KM	30	0.922	0.023	12	14	0.989	0.0314	13	14	5
Mon	MO	MK	41	0.921	0.0216	16	15	0.981	0.0308	13	18	6
Lawa 1	LW1	MK	46	0.959	0.0134	25	25	0.95	0.0237	15	25	4
Lawa 2	LW2	MK	50	0.913	0.0178	15	25	0.9533	0.0296	18	25	4
Khuen	KH	TK	60	0.967	0.0096	31	29	0.9877	0.0133	25	24	7
Lue 1	LU1	TK	51	0.915	0.0274	23	25	0.99	0.0142	22	24	9
Lue 2	LU2	TK	44	0.878	0.0257	14	21	0.981	0.0197	17	22	6
Lue 3	LU3	TK	50	0.988	0.0072	39	26	0.9969	0.0117	25	26	7
Lue 4	LU4	TK	46	0.932	0.0197	19	24	0.9783	0.0205	20	19	4
Yuan 1	YU1	TK	39	0.969	0.0145	26	20	0.9895	0.0193	18	19	6
Yuan 2	YU2	TK	50	0.974	0.0094	30	25	0.9833	0.0171	21	23	7
Yuan 3	YU3	TK	50	0.966	0.0116	28	26	0.9692	0.022	20	24	11
Yuan 4	YU4	TK	44	0.948	0.0147	21	21	0.9952	0.0165	20	19	7
Yong	YO	TK	62	0.965	0.0088	31	31	0.9892	0.0108	26	26	9
Shan	SH	TK	43	0.972	0.0106	26	19	0.9942	0.0193	18	20	7

Tab. 3.2.1 Samples included in this study and basic indices of genetic diversity: N = number of samples; h= haplotype diversity: sd = standard deviation. The linguistic affiliation of the populations is coded as: TK = Tai-Kadai; MK = Mon-Khmer; KM = Khon Mueang

We employed a discriminant analysis of principal components (DAPC) (Jombart et al. 2010) on both uniparental datasets to define genetic relationship among KM, MK and TK groups. The DAPC analysis can be used to investigate the relationship between populations optimizing the variation between- and within-groups while being free from assumptions about Hardy-Weinberg equilibrium or linkage equilibrium. We first assessed the best number of clusters in the HVR-I and Y-STR datasets using the *find.clusters* function in *adegenet 1.3-1* (Jombart 2008) and compared the results of 5

independent runs using a custom made R script. We then ran the DAPC analysis with 100,000 iterations and checked for the consistency of the groups founded.

An exploratory analysis such as the DAPC however do not account for geographic location of the samples, thus precluding the visualization of geographic locations where gene flow between populations is either hindered or facilitated. Estimated Effective Migration Surfaces (EEMS) which employed individual based migration rates can be used to visualize zones with higher/lower migration with respect to the overall rate (Petkova et al. 2015). The region under study is first divided in a grid of demes and the individuals are assigned to the deme closest to their sampling location. The matrix of effective migration rates is then computed by EEMS based on the stepping-stone model (Kimura and Weiss 1964) and on resistance distances (Mcrae 2006). We applied EEMS to a matrix of pairwise Φst distances constructed on the mtDNA dataset using Arlequin 3.5 (Excoffier and Lischer 2010) and on the 17 Y-STRs using the script available from Github at https://github.com/dipetkov/eems. We averaged three runs each with 200, 300, 400 and 500 demes to produce the final EEMS surface, as the number of demes simulated during the grid construction phase can influence the scale of the deviation from overall migration detected (Petkova et al. 2015). Each single run consisted of 200,000 burn--in steps followed by 500,000 MCMC iterations sampled every 1000 steps. We plotted the averaged EEMS and checked for MCMC convergence using the rEEMSplots package in R v 3.2.2.

In order to unravel the origin of the KM population, we employed an Approximate Bayesian Computation (ABC) approach on both the HVR-I and Y-STR datasets. We first constructed two competing models (Fig. 3.2.1), which are admixture and tree-like models based on our previous results (Kutanan, Kampuansai, Colonna, et al. 2011; Kutanan, Kampuansai, Srikummool, et al. 2017).

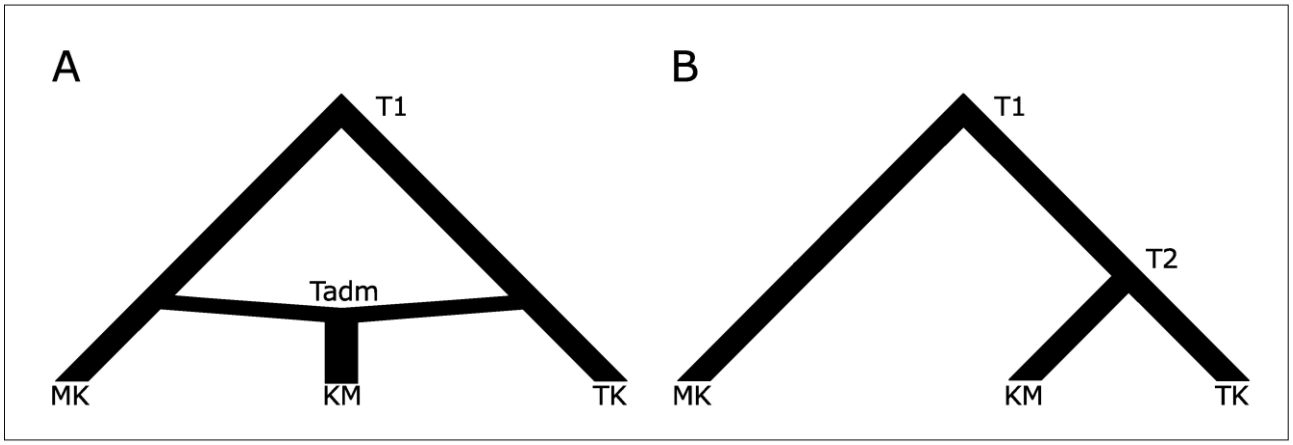


Fig. 3.2.1. The models tested in the ABC analysis on Mon-Khmer (MK), Khon Mueang (KM) and Tai-Kadai (TK) populations: admixture (A) and tree-like (B). Times of populations split and admixture are indicated as T1, T2 and Tadm.

In the admixture model, the KM population originated as a consequence of an admixture event from the parental populations, the MK and TK groups. The tree-like model postulates instead a recent separation of the KM and TK populations and a split of this combined population from the MK ones further back in time. For both the admixture and the tree-like models, we assumed constant effective population sizes based on historical records and that the prior distributions were all uniform (For a comprehensive description of the ABC procedure refer to the **case study in Chapter 2**). To evaluate the stability of the models' posterior probabilities, we considered different thresholds by considering different number of retained simulations for LR (25 000, 50 000, 75 000 and 100 000 best simulations). To generate the simulated datasets, we used the software package ABCtoolbox, running 500 000 simulations for each model. To calculate the models' posterior probabilities, we used R scripts from http:// code.google.com/p/popabc/source/browse/#svn%2Ftrunk%2Fscripts, modified by SG. To summarize the genetic information contained in both the HVR-I and Y-STR datasets we calculated two arrays of statistics within and between populations. For the mitochondrial dataset, we considered the number of haplotypes, the number of private polymorphic sites, Tajima's D, the mean number of pairwise differences for each population, the mean number of pairwise differences between populations and pairwise Fst. When we analysed Y-STR, we used as summary statistics the mean and the s.d. over loci in each population of four parameters: the number of alleles, haplotype diversity,

modified Garza–Williamson index and the allelic range. Finally, as the genetic heterogeneity was observed in KM populations (Kutanan, Kampuansai, Nakbunlung, et al. 2011), we repeated the ABC approach outlined in Fig 1 for each KM population (KM1 to KM10).

Results

The haplogroup frequencies in each studied population are listed in Table 3.2.2 and represented geographically in Fig 3.2.2. Haplogroup O-PK4 (or O1b1a1) is the most diffuse, being present in all populations with frequencies ranging from 7.7 % (YO) to 72.0 % (LW1).

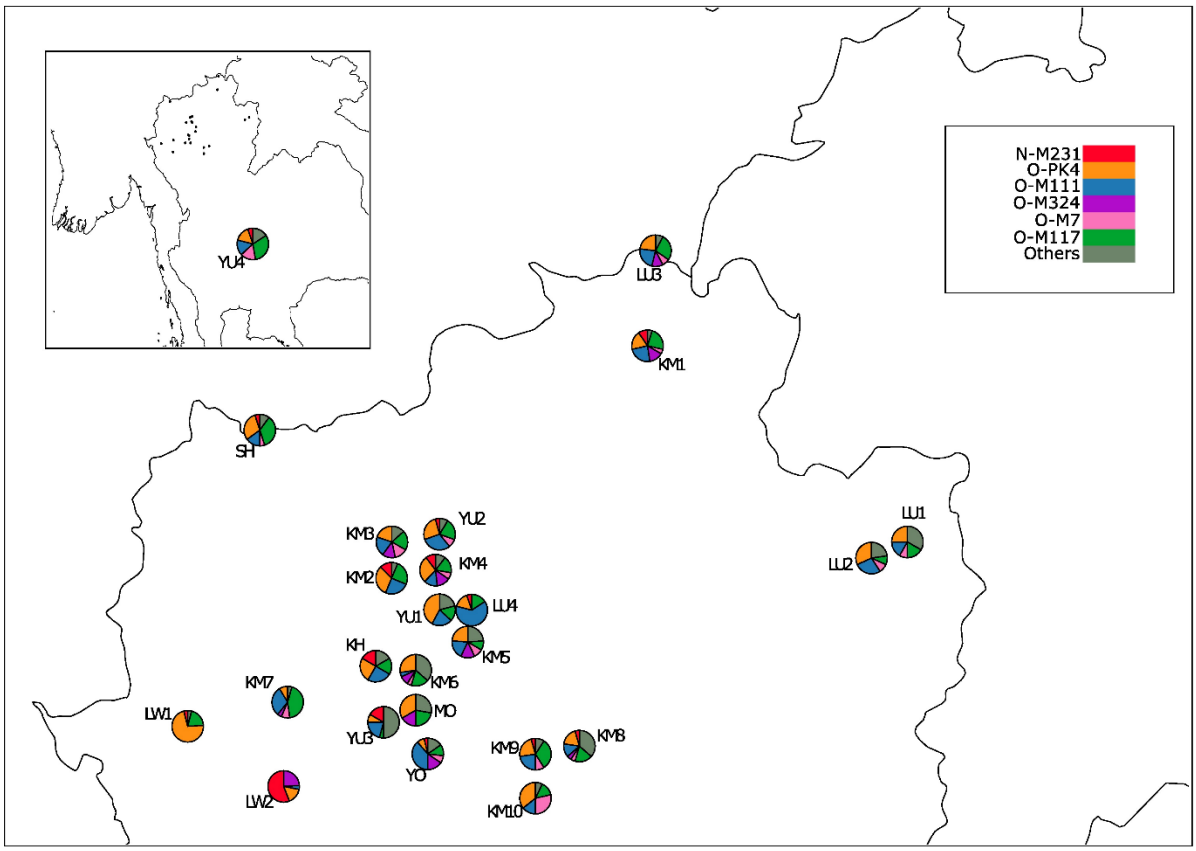


Fig. 2.2.2. Location of the studied populations and frequencies of the 7 major haplogroups obtained in Northern Thailand.

The O-M117 (O2a2b1a1) is a haplogroup which is also commonly found (4.3% - 43.5%) in almost all the populations, except LW2. The differentiation of LW2 is also evident from the elevated frequency of haplogroup N-M231 (56.0%), which occurred only at minor frequencies in the other

populations (ranging from 4.0% to 16.6%). However, LW2 is the only MK population harbouring O-M111 (O1b1a1a1a1a), a haplogroup otherwise widely distributed in both KM and TK. The similarity between KM and TK groups is further shown by the shared presence of haplogroups C-M217 (C2), D-M15 (D1a1), O-P203 (O1a1a) and O-M7 (O2a2a1a2). This similarity is enhanced when considering the populations inhabiting the central part of northern Thailand. Some other haplogroups commonly present in KM, such as C-M130 and O- M324 (O2a), seem to be shared with some MK (MO, LW2) and some TK (LU3, YO, YU3) groups. Interestingly, while being geographically removed from the other sampled locations, YU4 showed similar haplogroup distribution related to the other TK populations.

		Haplogroups and diagnostic SNPs																						
Population	Size	C	C1b1a1	C2	CF	E1	D1a1	J2	K	H1a1	L1a1	N	NO	O1a1a	O1a2	O1b1a1	O1b1a1a1a1a	O2	O2a	O2a2a1a2	O2a2b1	O2a2b1a1	R1	R2a
		M130	M356	M217	P143	P173	M15	M172	P128	M52	M76	M231	P195	P203	M50	PK4	M111	M122	M324	M7	M134	M117	M173	P249
KM1	21											9.5		4.8		19.1	23.8		14.3	4.7		23.8		
KM2	16			6.2								12.5				31.3	25					25		
KM3	15	6.7												6.7		20	20		13.3	13.3		20		
KM4	29			3.5		3.5				3.5		10.3				27.6	13.8		13.8	6.8		17.2		
KM5	21			4.8			4.8							14.3		23.8	19		14.3	9.5		9.5		
KM6	22	4.6	4.6	4.6			4.6	4.6								27	4.6	4.6	9	4.6		18.2	9	
KM7	23						4.4									8.7	30.3		4.4	8.7		43.5		
KM8	22	4.6		9							9	4.6		4.6	9	18.2	13.6		4.6	4.6		18.2		
KM9	22						4.6		4.6			4.6				22.7	22.7			9		31.8		
KM10	14	7.1														35.7	14.3			28.6		14.3		
MO	18							5.6						5.6		33.2			16.7			22.2		16.7
LW1	25						4					4				72						20		
LW2	25											56				16	4		24					
KH	24						8.3					16.7		4.2		25	25				4.1	16.7		
LU1	24			8.3			4.2						4.2	12.4		25	16.7	4.2		8.3		16.7		
LU2	22													9.1	13.6	31.8	27.3			9.1		9.1		
LU3	26												4	4		23	23		11.5	7.7		26.8		
LU4	19											5.3				15.8	63.2					15.8		
YU1	19			5.3												42.1	21				10.5	15.8	5.3	
YU2	23			4.4								4.4				26.1	30.4	4.3		8.7		21.7		
YU3	24	4.2		12.5			4.2					16.6		4.2		8.3	20.8	16.6			4.2	4.2	4.2	
YU4	19									5.3		5.3		10.5		15.8	15.8			15.8		31.5		
YO	26						3.9					3.9		7.7	3.9	7.7	38.4		15.4	7.7		11.4		
SH	20				5							5		5		30	15			5		35		

Tab.3.2.2. Frequencies of Y chromosomal SNP haplogroups in KM, MK and TK populations

The DAPC analysis failed to find a clear most supported number of *K* in both uniparental datasets. However, once we assigned each sample to either its language group (Figs 3.2.3A and B) or its original population (Figs 3.2.3C and D), the resulting scatterplots highlighted several patterns. The DAPC analysis based on the Y-STR dataset revealed a general overlapping of the Tai-Kadai and Khon Mueang clusters while the Mon-Khmer populations seemed clearly distinct, especially LW1

and LW2. When grouped separately, some the Khon Mueang populations inhabiting central northern Thailand such as KM3, KM4, KM5 and KM1 slightly departed from the general trend of similarity with the majority of Tai-Kadai. These populations appeared closer to the Mon-Khmer populations than the rest of KM. The Tai-Kadai population clusters were largely overlapping, with the notable exception of the SH that fell closer to LW1 and MO. The DAPC based on the mtDNA-HVR1 dataset presented less separation among both linguistic and population-based clusters, showing the overall similarity between maternal lineages in northern Thailand.

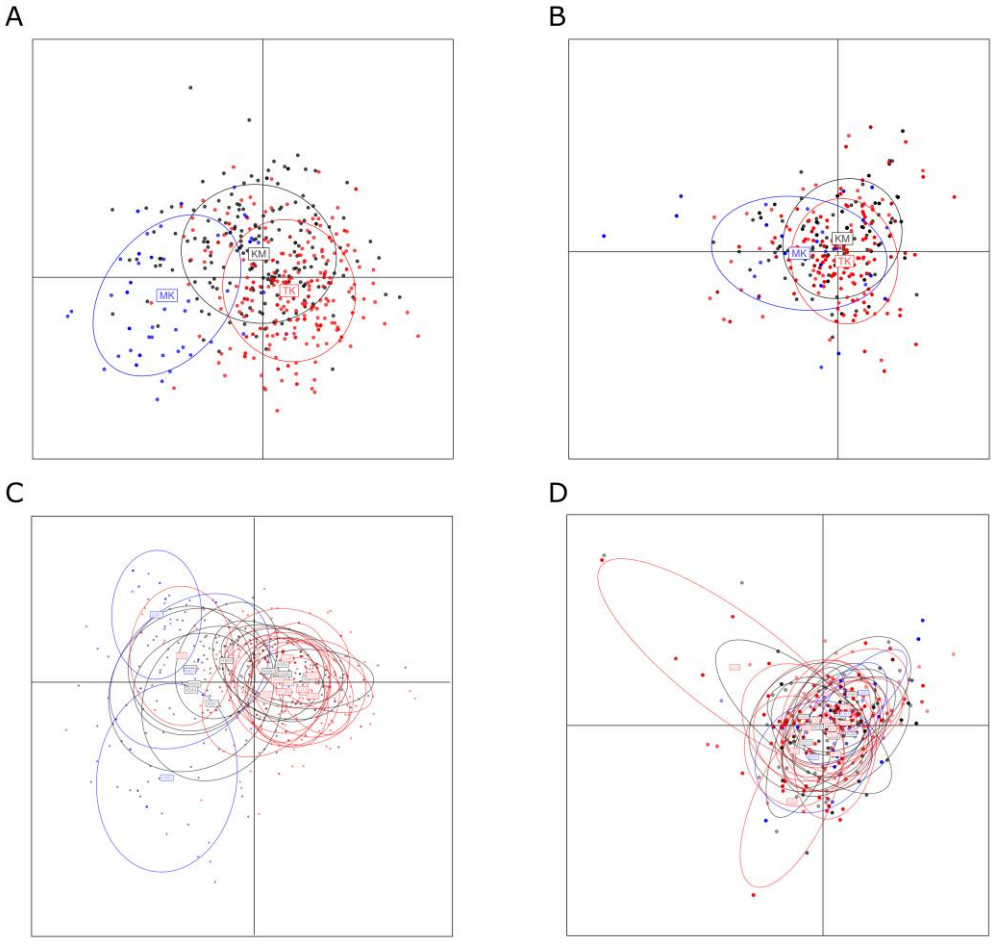


Fig. 3.2.3. DAPC analysis on the Y-STR (A,C) and mtDNA-HVR1 (B,D) datasets. Clusters are colored based on linguistic groups: TK (red), KM (black) and KM (blue). Individuals were grouped based on linguistic affiliation (A,B) or populations (C,D).

The EEMS surfaces showed an overall pattern of good connectivity between neighbouring populations in northern Thailand with only moderate reductions/increment of migration rates (Figs 3.2.4A and B). Especially evident in the Y-STR dataset, the geographic outlier population, YU4, was connected with TK and KM populations residing in the central part of northern Thailand by a corridor of high effective migration (Fig 3.2.4A). These northern Thai populations were in turn well connected with each other and with the eastern Lue (LU1 and LU2) populations. The strongest barrier in the Y-STR dataset was, not surprisingly, the one separating LW1 and LW2 populations, leading to lower migration rates with surrounding KM and TK populations (KM7 ,YU3 and YO). The MO did not conform to the isolation pattern presented by the Lawa, showing higher than expected migration rates with TK and KM populations. The EEMS surface based on the mtDNA-HVR1 dataset (Fig 3.2.4B) highlighted a weak spatial structure compared with the Y-STR dataset. We observed lower than expected migration rates between LW1, LW2 and the KM populations from the southern part of northern Thailand (KM8, KM9, KM10), as well as a feeble barrier between LU1 and LU2. However, higher migration rates indicate the connection between the Shan and populations residing in the central part of northern Thailand.

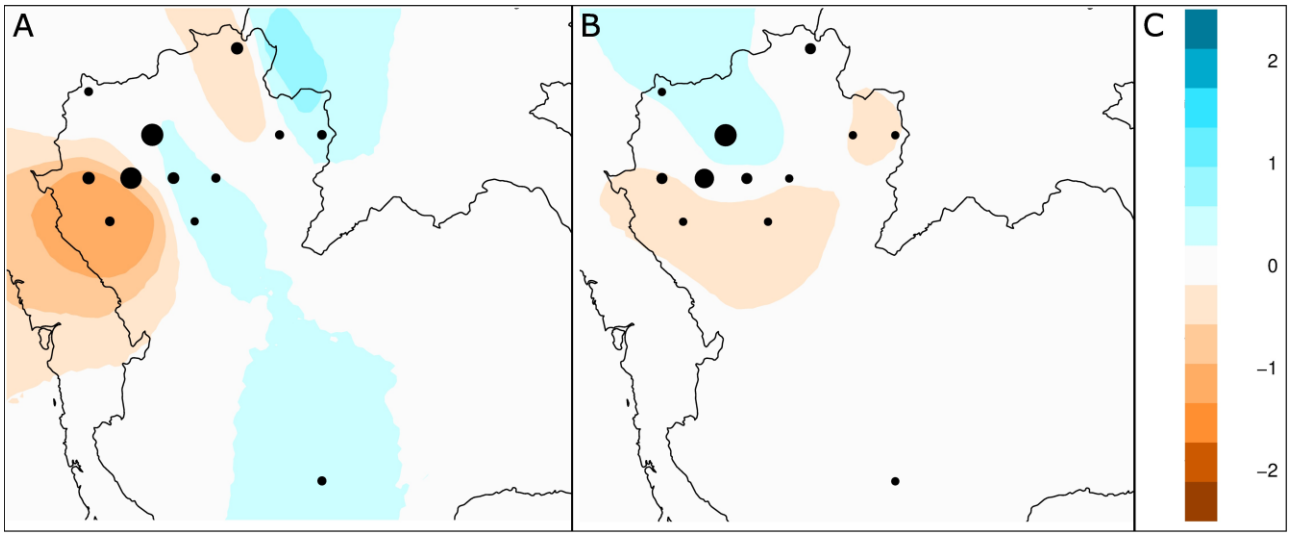


Fig. 3.2.4. EEMS analysis of effective migration rates (m) on the Y-STRs (A) and mtDNA-HVR1 (B) datasets. The effective migration rate is represented on a log10 scale represented on the right. Areas showing negative values (orange) represent possible barriers to gene-flow while zones with

positive values (blue) correspond to places of increased gene-flow with respect to normal IBD (white).

The posterior probabilities from ABC analysis of the two considered evolutionary models are presented in Tab. 3.2.3. For the Y-STR dataset, we found that the tree-like model postulating a recent split between the ancestors of modern KM and TK populations provided a better explanation for the KM origin than an admixture model. The high and stable posterior probabilities over a different number of retained simulations confirmed the chosen tree-like model. Once we repeated the ABC analysis for each separated KM population, the results weakly supported tree-like model in most of the KM populations. The ABC analysis failed to support the tree-like model in KM9. The results based on mtDNA HVR-I were also less indicative of supporting the tree-like model than the admixture, however, the results obtained from simulations conducted on separated KM populations supported the tree-like model in almost all of the KM populations (e.g. KM2, KM3, KM4, KM6, KM7 and KM10). There was only one instance of KM5, in which the admixture model was preferred for the mtDNA-HVR1 dataset.

	HVR-I		**Y-STR**	
Threshold	Admixture model	Tree-like model	Admixture model	Tree-like model
25000	0.566	0.434	0.294	0.706
50000	0.479	0.521	0.297	0.703
75000	0.468	0.532	0.299	0.701
100000	0.487	0.513	0.302	0.698

Tab. 3.2.3. Posterior probabilities of each model performed by ABC analysis under weighted multinomial logistic regression.

Discussion and Conclusion

The investigation of Y chromosomal lineages in northern Thai populations revealed that the majority of the sampled individuals could be assigned to one of three common haplogroups: O-PK4 (O1b1a1),

O-M117 (O2a2b1a1) and O-M111 (O1b1a1a1a1a). These lineages are also prevalent in Chinese and other Southeast Asian populations (Li et al. 2008; Cai et al. 2011; He et al. 2012; Ning et al. 2016). The overall pattern of haplogroups distribution was also generally homogeneous in our studied populations, with subhaplogroups of O1 and O2 reaching the highest frequencies amongst the studied individuals. This was especially true for TK and KM populations. Interestingly, the MO from northern Thailand show the presence of haplogroups usually found in South, Central and West Asia: R-P249 (R2a) and J-M172 (J2) (Sengupta et al. 2006). Connection between the ethnic Mon and populations from South and Central Asia was already proposed from previous identification of mtDNA lineage W3a1b (Kampuansai et al. 2017). Both groups of the MK speaking LW groups showed high differences between each other and from other populations, presenting low levels of haplogroup diversity (Tab. 3.2.1) with high frequencies of O-PK4 (O1b1a1) (72% in LW1) and N-M231 (56% in LW2). Haplogroup N-M231 is prevalent in today's TK and Hmong-Mien speaking populations, as well as in Han of southern China (Shi et al. 2013). We also detected haplogroup C-M130 and its sublineages (C-M356 and C-M217) in almost all KM and YU populations, as well as in one Lue group (LU1). Haplogroup C-M130 has been found mainly in Mongolia and in Korea, while in Southeast Asia it reaches high frequencies in the eastern part of Indonesia. C-M217 is, instead, typically present at high frequencies across northeast Asia (Stoneking and Delfin 2010). It is worth to note that we observed haplogroup D1-M15 in several TK and KM populations, although at low frequencies. High frequency of this lineage was reported in China especially in Tibet, Quiang and Yao (Wang et al. 2014). To account for the presence of C and D lineages in our TK groups, we speculate that paternal admixture among several ethnolinguistic groups in the area of southern China were heavily influenced by Han and Mongol expansion from the north (LaPolla 2001; Blum 2002; Wen et al. 2004; Zhong et al. 2010). This would have happened before the southward migration of the TK ancestor to northern Thailand, which could contribute to the presence of C and D lineages in these populations. In agreement with earlier studies of mtDNA and autosomal STRs variation (Kampuansai et al. 2017; Kutanan, Kampuansai, Brunelli, et al. 2017) that reported genetic

differentiation of MK speaking groups in northern Thailand, the two LW groups appeared to differ from other populations and from each other, as indicated by the DAPC results (Fig. 3.2.3C). A lower level of gene flow caused by the presence of geographic barriers may be the driven factor of this differentiation, as suggested by the EEMS surfaces (Fig 3.2.4). On the other hand, the Mon is less genetically differentiated from the KM and TK groups. Although the genetic origin of the Mon is related to South Asia, recent admixture with the Tai sources could have been the main force that shaped the genetic variation of the northern Thai Mon. To investigate the origin of the KM, we proposed two demographic models, which summarized previous hypothesis on their origin. The first scenario depicted an admixture event involving local MK populations and migrating TK groups, while a second tree-like model suggested that the KM originated following a recent split from the ancestors of modern TK populations (Fig 3.2.1). We then proceeded to test these hypotheses with ABC simulations on both maternally and paternally inherited data. When all the KM individuals were pooled together, the most supported model was the one postulating a close relationship between these populations and TK (Tab. 3.2.3). This demographic pattern was clearer when Y-STRs were employed instead of mtDNA, possibly suggesting significant maternal contribution from local MK speaking on current KM populations. This conclusion was reinforced when the single KM populations were considered separately. In some specific cases (KM5), the admixture scenario obtained higher posterior probabilities than the tree-like one, highlighting the importance of considering small scale local processes when investigating the history of a population. In conclusion, we observe contrasting pattern of paternal and maternal genetic variation with a clearer genetic structure in the Y-STRs than in the mtDNA-HVR1 sequences. Our different approaches suggest nonetheless a common origin between KM and TK populations, previously proposed to be located in southern China (Kutanan, Kampuansai, Srikummool, et al. 2017). After an initial migration southward, fragmentation process in separate villages and contact with local MK speaking people could have promoted secondary events of admixture, especially in the matrilineal lineage. Although some models and populations compared are different from our previous mtDNA genome study (Kutanan, Kampuansai,

Srikummool, et al. 2017), the results are confirmed with the additional inclusion of post-settlement contacts with MK populations. Future work employing complete Y sequences and/or in depth autosomal information from northern Thai and surrounding populations will be crucial to determine the migration origin and to evaluate the full impact of secondary admixture.

Chapter 4. ANCIENT DNA AND RECENT MIGRATIONS IN EUROPE

Neolithic and post-Neolithic migrations in Europe

The Neolithic transition saw an overall substitution of a hunter-gatherer life style in favour of sedentary farming techniques. This revolution occurred independently in different regions of the world (Diamond and Bellwood 2003) and, for Europe, it originated in the Near East around 12,000-11,000 BP (Barker and Goucher 2015). Since the first archaeological studies, two different demographic scenarios were developed to explain the modalities of this transition: the demic diffusion model and the cultural diffusion model (Bellwood 2001). The first one postulate that the spread of farming was mediated by a migration of people, implying a significant genetic contribution of these Neolithic farmers into the European gene pool (Ammerman and Cavalli-Sforza 1984). The cultural diffusion model instead hypothesize that the transition to agriculture was mainly a cultural phenomenon, involving the sharing of the new techniques without a population migration and the consequent gene-flow (Renfrew and Boyle 2000). While uniparental studies initially provided confirmations of the demic model (e.g. Chikhi et al. 2002), it was with the advent of complete ancient genomes that the diversity between current Europeans and Mesolithic hunter-gatherers became clearly visible (Lazaridis et al. 2014; Haak et al. 2015; Lazaridis et al. 2016). Today we know that early farmers from Anatolia were the source of the Neolithic migration into Europe and, while ancient samples present differences with the current inhabitants of the region (Omrak et al. 2016), their ancestry can be detected in modern-day Southern Europeans and, albeit at lower frequencies, in northern populations too (Lazaridis et al. 2014). It is different to understate the impact of the Neolithic migration, as the genetic makeup of entire regions was subjected to a monumental change with respect to the Mesolithic period (e.g. Modi et al. 2017). However, these newly arrived farmers did not isolate themselves from the resident hunter-gatherer populations, and events of admixture were evident from the analysis of Middle-Neolithic individuals (Günther et al. 2015).

Even accounting for their importance, the Neolithic migrations were not the only big movement of populations happened in Europe during last few millennia. Differently from earlier samples, individuals from central Europe dated around 5,000 years ago present a genetic affinity with the ancient Siberian genome of Mal'ta (Raghavan et al. 2014) suggesting an event of gene-flow with a source originated in central Eurasia. This connection was probably caused by the massive migration of the Yamnaya herders coming from the Pontic-Caspian steppes, which have been suggested to be descendent from different hunter-gatherer populations from Russia and Caucasus (Allentoft et al. 2015; Haak et al. 2015). This Yamnaya migration was associated also with several new cultures that arose in central Europe around 4,800 years ago, collectively known as Corded Ware, Battle Axe or Single Grave (Vandkilde 2007). This hypothesis based on archaeological findings was recently supported by studies on ancient genomes, which highlighted how 79% of the ancestry found in Corded Ware samples from Germany could represent ancestry derived from the Yamnaya herders (Haak et al. 2015). This later culture would expand as far as the British Isles, shaping the diversity that can be found even today in European populations.

After the migrations of the Bronze Age, European populations were interested by additional events of gene-flow, albeit at a much reduced scale. For example, southern European populations show evidence of gene-flow from Northern Africa and the Near East dated to have happened during the first millennium CE, probably as a consequence of the Arabic conquest. Moreover, an influx of Northern Europe was also detected in Southern European populations around the medieval period, while the formation of the Slavic people (around 1,000 years CE) had effects on both Northern and Eastern European populations (Busby et al. 2015). Even if regional in focus, these later contributions had the power to significantly change the genetic makeup of a location. The Etruscan culture, for example, is documented as being present in Central Italy between the eight and the first century BCE. A study on ancient mtDNA retrieved from Etruscan remains highlighted how their ancestry could be detected only partially in modern human populations inhabiting the same regions, hinting at one or

multiple event of population replacements happened during the last five centuries (Ghirotto et al. 2013).

While the Neolithic and post-Neolithic periods represent only a small fraction of the total human history, the migrations located within them have permanently shaped the ancestry of modern European populations.

The following case study focuses on medieval migrations in Europe. Additional works in which I have participated on the subject of Post-Neolithic migrations are:

Tassi, F., Vai, S., Ghirotto, S., Lari, M., Modi, A., Pilli, E., **Brunelli, A.**, Susca, R.R., Budnik, A., Labuda, D., Alberti, F., Lalueza-Fox, C., Reich, D., Caramelli, D. & Barbujani, G. Genome diversity in the Neolithic Globular Amphorae culture and the spread of Indo-European languages. (Accepted for publication in Proceedings of the Royal Society B.)

Case study: A genetic perspective on Lombard migrations

Outline of the research

From the first century AD, Europe has been interested by intense population movements, also known as Barbarian migrations. According to historical records dated back to the first century CE, the Lombard were Germanic people that inhabited the northern Elbe region (Johne 2008). Around 500 CE the term “Lombard” reoccurs in the region north of the middle Danube, and three generations later, in 568 CE, the Lombard kingdom has been funded in Italy (Geary 2002). There is a clear connection between the material culture of Pannonia and Italy at the end of the 6th century that suggests strong interaction and communication – possibly as a result of a migration known from the written sources – between these two regions (Giostra 2011). Archaeological and written sources, however, are open to different interpretations, and unable to answer questions with regard to the biological nature of migrations, as well as to their impact upon previously-settled populations. In particular, it is still highly debated whether and to what extent attributes such as grave goods and

burial traditions are indicators of Lombard social identity, and whether the spread of these material markers across Europe is actually linked to population movements rather than to horizontal cultural transmission. In this light, the analysis of ancient genetic data is fundamental to obtain a better understanding of past population dynamics and interactions. The only Lombard ancient genetic data ever published so far were sequences of the mtDNA control region from a single cemetery in Italy, Collegno, that showed evidence of genealogical ties between Lombard and a specific modern population from the same region (Vai et al. 2015) and sequences of mtDNA control region and of informative positions in the coding region from the cemetery of Szólád in Hungary (Alt et al. 2014). A broader analysis, based on larger assemblages of samples and genetic markers, is clearly necessary to provide a finer resolution of medieval population movements and interactions in Europe. In this study we sequenced complete mitochondrial genomes from nine early-medieval cemeteries located in the Czech Republic, Hungary and Italy, for a total of 88 individuals. Based on archaeological remains, some of these burial communities presented at least a portion of individuals who can be associated to Lombard/Germanic culture (hereby we refer to these cemeteries as LC), as opposed to burials in which no artifacts related to Lombard culture has been found in any graves (hereby referred as NLC). We used these sequences to explicitly test the Lombard migratory route and the possible contacts of this moving populations with previous inhabitant of the region.

Materials and Methods

We extracted complete mitochondrial DNA from teeth, long bones and petrous bones in 136 individuals from nine early-medieval cemeteries located in the Czech Republic, Hungary and Italy (Fig. 4.1A). Archeologically, 5 of these cemeteries have been partly or entirely associated with the Lombard culture: Mušov in the Czech Republic (LRCMUS), Szólád (LHSZ) and Hegykő (LHHEG) in Hungary, Collegno (LICOL) and Fara Olivana (LIFAR) in Italy. The other 4 sites -Fonyód (NLHFON), Hács-Béndekpuszta (NLHHACS) and Balantoszemes-Szemesi Berek (NLHBAL) in

Hungary, Torino-Giardini Reali (NLIGR) in Italy- while being coeval and geographically close to the other cemeteries, did not show any cultural link with the Lombards. Double-stranded NGS libraries, constructed from each extract, were enriched for mitochondrial DNA and sequenced on an Illumina MiSeq. Sequences were demultiplexed and sorted according to the sample, then raw sequence data were analyzed using the pipeline described in (Modi et al. 2017). Merged reads were mapped on the revised Cambridge Reference Sequence, rCRS (GenBank Accession Number NC_012920). Reads with mapping quality below 30 were discarded. Consensus sequence for each sample was obtained considering positions covered at least 3 fold, and base calling was performed with at least 70% of concordance between reads. Misincorporation pattern was analysed using MapDamage 2.0 (Jónsson et al. 2013) and contamination estimate was performed by contamMix (Fu et al. 2013). Only samples with at least 92% of the mitochondrial genome covered at least 3 fold, with CtoT values higher than 17% and MAP authentic values higher than 92% according to contamMix were considered for population genetics analysis. This resulted in a total of 79 suitable sequences from cemeteries archeologically associated to the Lombard culture (7 from LRCMUS, 40 from LHSZ, 8 from LHHEG, 23 from LICOL, 1 from LIFAR) and 9 suitable sequences from cemeteries not associated to the Lombard culture (3 from NLHFON, 3 from NLHHACS, 2 from NLHBAL and 1 from NLIAR) (Table S2). The 88 medieval mitogenomes were sequenced to an average coverage depth of 86.10x (from 6.66x to 201.89x).

The mitochondrial haplogroups were assigned according to PhyloTree Build 16 on Haplogrep (van Oven and Kayser 2009; Kloss-Brandstätter et al. 2011). Phylogenetic networks, based on nucleotide variation in the new 88 mtDNA sequences, were constructed using the Median Joining algorithm (Bandelt et al. 1995) implemented in Network 5.0 program (http://www.fluxus-technology.com). The ε value was set to 0 and the transversions were weighted 3x the weight of transitions. Networks were subjected to maximum parsimony post-analysis. Haplogroup frequencies for modern and medieval populations were retrieved from previously published data as outlined in (Csákyová et al. 2016).

Since the vast majority of studies provided haplogroup frequencies inferred from HVRI, we reassigned the newly reported LC and NLC to haplogroups employing only this region. PCA based on haplogroup frequencies was conducted employing the function *fviz_pca_biplot* from the library *factoextra* (Kassambara and Mundt 2016) in R 3.4.0. We compared LC and NLC complete mitochondrial genomes with a PCA using the function *dudi.pca* from the package *adegenet* (Jombart 2008). We also computed pairwise differences between sequences with Arlequin v. 3.5. (Excoffier and Lischer 2010) and visualized them employing an MDS computed with the function *cmdscale*. To locate possible population structure inside our dataset we assessed the best number of clusters inside it using the *find.clusters* function in *adegenet* (Jombart 2008), comparing the output of 10 independent runs using a custom-made R script. We then applied a DAPC analysis (Jombart et al. 2010) on the dataset with 100,000 iterations.

To explicitly test the importance of NLC populations on the genetic makeup of migrating LC individuals we conducted demographic simulations under an Approximate Bayesian Computation model framework. In order to provide an unbiased representation of genetic diversity inside our dataset, we first removed related samples based on the kinship analysis presented in (Krishna et al. in prep.). We also excluded from our demographic dataset 2 individuals from Szólád (LHSZ27A1/2) as the dating of their burial suggest a more recent origin with respect to surrounding graves. This process resulted in a reduced dataset of 79 unrelated sequences. We hypothesized two models called admixture and continuity (Figure 3). The first recreate a scenario where a migrant population, LC from northern Europe, receive gene flow from local NLC individuals before moving to colonize other regions. The continuity model instead postulates no contact between LC and NLC populations, mimicking only the proposed Lombard migration from Czech Republic to Italy. The simulated datasets for each scenario were obtained using the *fastsimcoal2* simulator within the software package ABCtoolbox (Wegmann et al. 2010). We performed the model selection procedure making use of the novel approach developed by Pudlo etl al. 2016, called *ABC-rf* (Pudlo et al. 2016), which rely on the

"random-forest" machine learning algorithm (Breiman 2001). Random forest uses the simulated datasets for each model in a reference table to predict the best suited model at each possible value of a set of covariates. After selecting it, another Random Forest obtained from regressing the probability of error of the same covariates determine the posterior probability. This procedure allows to overcame the difficulties traditionally associated with the choice of summary statistics, while gaining a larger discriminative power among the competing models (Pudlo et al. 2016). We compared our models with ABC-rf considering 20,000 simulations per model and 1,000 trees in the forest, using the functions provided in the *abcrf* R package. To summarize the genetic information contained in our sequences we considered the number of haplotypes, the number of private polymorphic sites, Tajima's D, the mean number of pairwise differences for each population, the mean number of pairwise differences between populations and pairwise F_{st}. These summary statistics were obtained with arlsumstat (Excoffier and Lischer 2010). We validated the model selection procedure calculating the classification error through the *abcrf* function of the *abcrf* R package. To do this, we used as pseudo-observed datasets each dataset of our reference table. To verify whether the selected models are able to generate the observed data, we performed a linear discriminant analysis (LDA) and verify whether the observed value fall within the variation generated by the tested scenarios. The LDA plot have been generated using the function of the *abcrf* R package. In order to estimate the parameters for the model chosen by the ABC-rf procedure (the admixture one) we ran further simulations reaching 1 million datasets. We applied a locally weighted multivariate regression (Beaumont et al. 2002) after a logtan transformation (Hamilton et al. 2005) of the 3,000 best-fitting simulations to estimate the admixture model's parameter using an R scripts from http://code.google.com/p/popabc/source/browse/#svn%2Ftrunk%2Fscripts, modified by SG.

Results

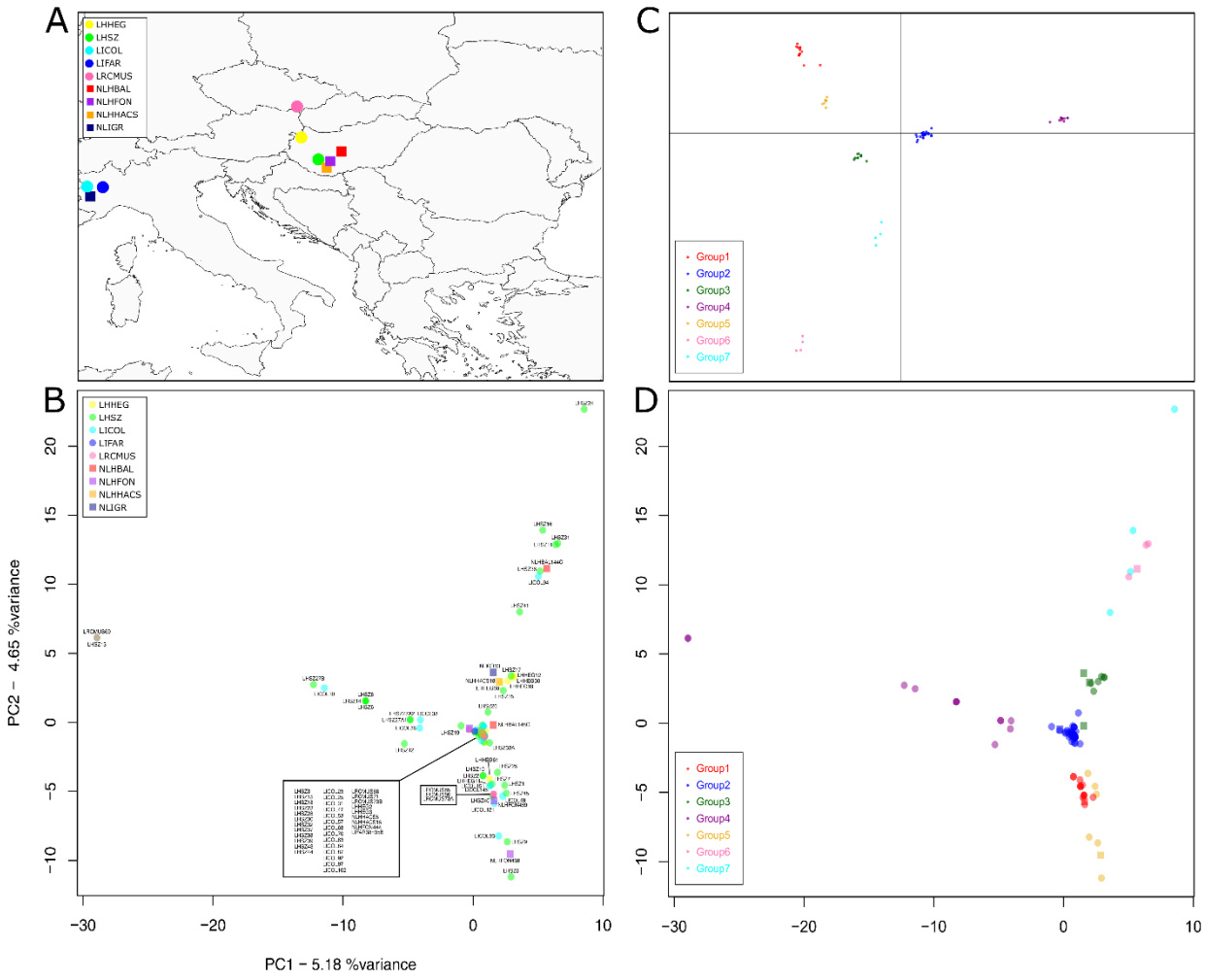


Fig. 4.1. A) Locations of the samples included in the study, B) PCA of the 88 LC and NLC sequences, C) Scatterplot of the DAPC analysis at K=7 and D) PCA colored according to the group scheme highlighted by the DAPC

Sequences were assigned to 71 distinct haplotypes, falling within the expected overall mitochondrial diversity of western Eurasian mtDNA. The haplogroup assignment, indeed, showed that the majority of the individuals belong to the H, T2 and J lineages (respectively occurring in 33, 11 and 7 samples, Figure 4.2B), all of them commonly observed in Europe (Kivisild 2015). Mušov and Szólád LC cemeteries show similar frequencies of the H haplogroup (28% and 32% respectively), while in Collegno this frequency is doubled (60%) (Figure 4.2B). In Szólád and Hegykő (LC groups in Hungary) the haplogroup distribution showed a substantial differentiation; as an instance, we found

the U8 haplogroup uniquely present in Hegykő, with a frequency of 25%. Phylogenetic links between haplotypes and their distribution among the archaeological sites are shown in the Median Joining Network (Figure 2A).

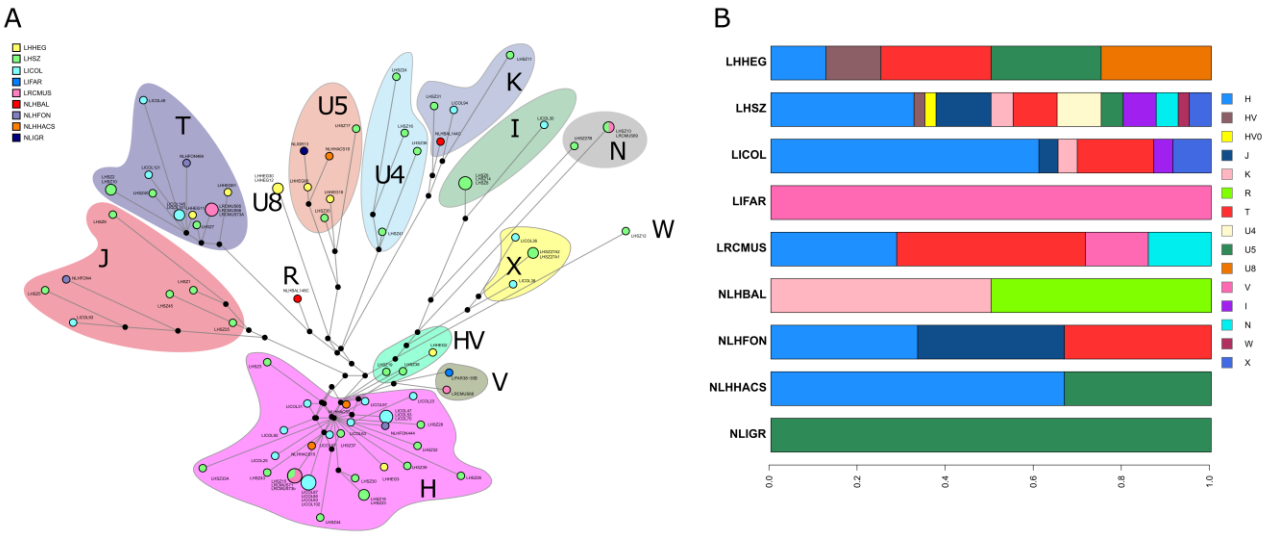


Fig. 4.2 A) Network of the 88 newly sequenced samples. Colors are associated with the cemeteries from which the samples come from, the size of the circles is proportional to the number of samples carrying a specific haplotype, and the background shading indicates the affiliation of the lineages to the major haplogroups. B) Barplot representing haplogroup frequencies in LC and NLC populations

To elucidate the affinities between our new LC/NLC samples and coeval individuals we retrieved data on mtDNA haplogroup frequencies for 10 European medieval populations. A principal component analysis (PCA) on this new dataset highlighted the similarity between the LC grave of Szólád and medieval populations from Central Europe (Slovakia 800-1100 CE, Poland 1000-1400 CE) (Fig 4.3). The same pattern of relatedness, albeit less evident, emerged when looking at the LC populations of Mušov and Collegno. The latter clustered midway between Slovakian and medieval samples from Southern Europe (Spain 500-600 CE, Italy 900-1400 CE). The striking difference presented by the Hungarian LC population of Hegykő can probably be attributed to the nearly unique presence of haplogroup U8 in this population. The Hungarian NCL populations of Hács-

Béndekpuszta and Balantoszemes-Szemesi Berek clustered close to the Italian Piedmont (500 – 700 CE) and Spanish samples, while Fonyód fell close to our LC samples from Collegno.

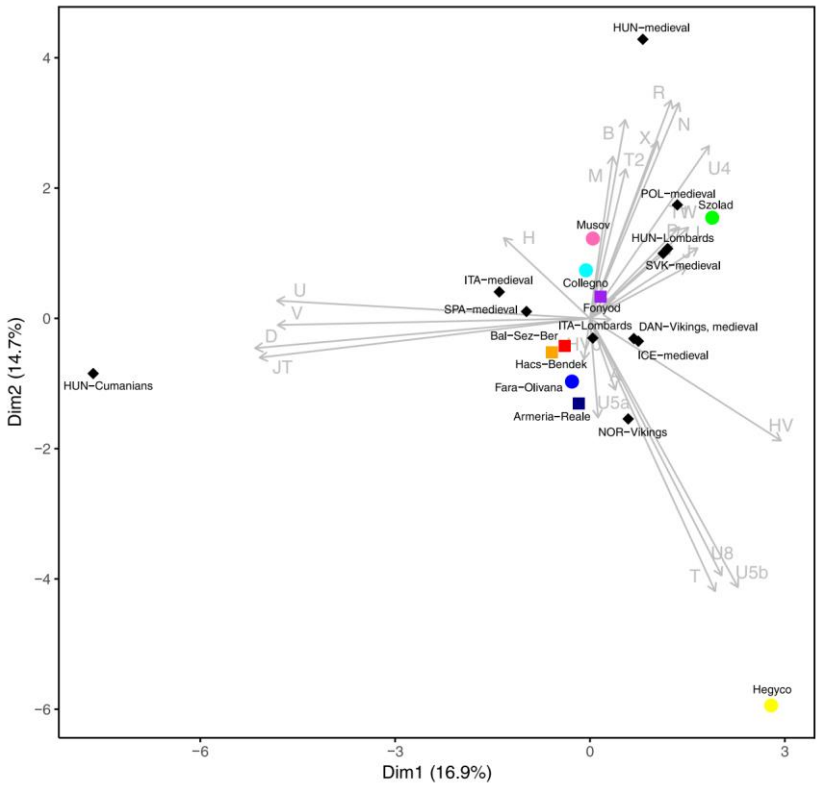


Fig. 4.3. PCA based on haplogroup frequencies of LC and NLC samples and 10 medieval populations

We explored the relationships between LC and NLC individuals through a principal component analysis (PCA) (Figure 4.1B). The first two axes of the PCA suggest a degree of similarity between groups, as NLC individuals are found across all of the range of genetic variation shown by the LC samples. There is also no clear geographical structure between samples in our dataset, with individuals from Italy, Hungary and Czech Republic clustering together. However, the first PC clearly separates a group of 12 LC individuals found at Szólád, Collegno and Mušov from a group composed by both LC and NLC individuals. To further shed light on this peculiar genetic structure we performed a k-means analysis and a discriminant analysis of principal components (DAPC) on the whole dataset. At K = 4 the 12 LC individuals located by the first PC form a cluster together, remaining undivided even at K = 7, the most supported number of clusters (Figure 4.1C and D). The presence in this group

of LC sequences belonging to macrohaplogroups I and W, commonly found at high frequencies in northern Europe (e.g. Finland, Headman et al. 2007), suggests (although certainly does not prove) the existence of a link between these 12 LC individuals and the original homeland of Lombards. These individuals would have maintained typical Northern features in their new territories in Hungary and Italy. This possibility is strengthened by archaeological information from the Szólád cemetery, where 8 of the 12 individuals in this group originate, indicating that all these samples were found buried with typical Lombard artefacts such as swords and brooches. We do not find the same tight association for the 3 samples from Collegno. The 3 graves are indeed devoid of evident Germanic cultural markers; however they are not placed in a separate and marginal location—as for the tombs without grave goods found in Szólád —but among graves with wooden chambers and weapons that are distinctive markers of Germanic culture. In this light, the so buried individuals may have been members of the Lombard community as well, but belonging to the lowest social level (identified by the written sources as semi-free or slaves). This social condition could explain the absence of weapons, which were reserved to free individuals, and could be related to mixed marriages, whose offspring occupied a lower social rank. Finally, this group also includes an individual from the Mušov graveyard. This finding is particularly interesting in light of the fact that the Mušov necropolis has been only tentatively associated with Lombard occupation (see Star Methods for details), based on the presence of few archaeological markers.

We tested hypotheses about the local demographic impact of Lombard migrations by an Approximate Bayesian Computation framework, comparing a model of admixture between LC and local NLC populations and a migration model with no contact between the two (Fig. 4.4).

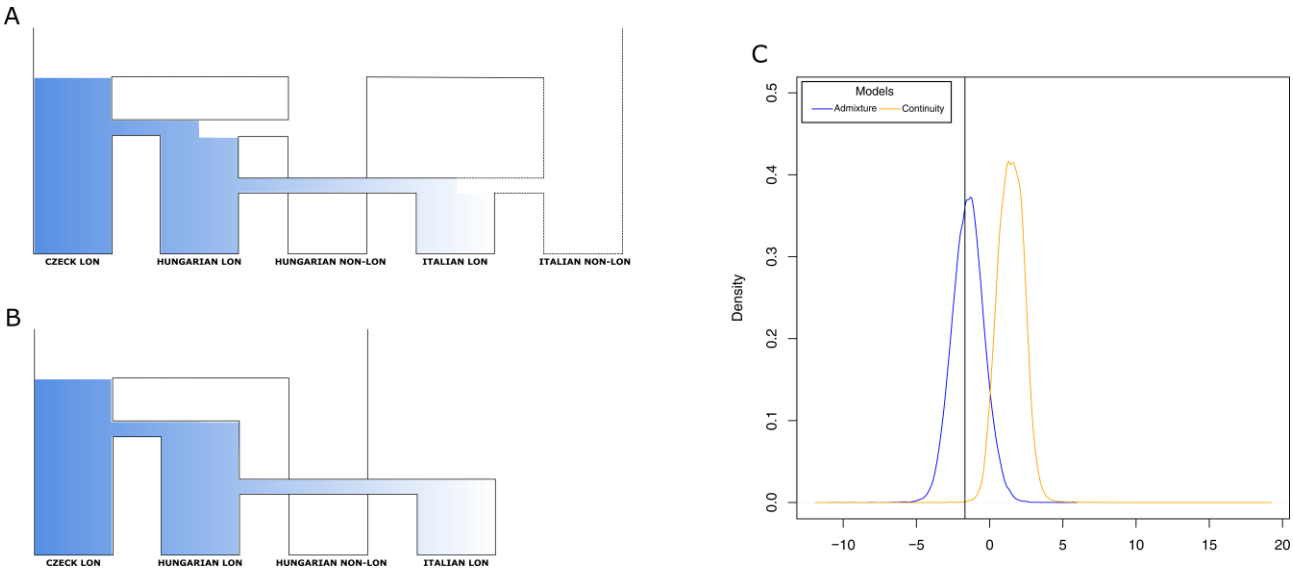


Fig. 4.4. A) Admixture and B) continuity models tested with ABC simulations. C) LDA plot of the simulations, the black vertical line represent the observed data.

The admixture model received strong support, with probability of 99% versus a 1% attributed to the alternative model. The two models were well recognized from the ABC procedure we applied, as indicated by the LDA plot (Fig. 4.4C). This result indicates the important effect that gene-flow from previous inhabitants of the region had into the migrating LC group. Indeed, when we estimated the extent of these admixture events, we observed that more than 80% of the genetic makeup of Hungarian LC population could be traced to NLC people inhabiting the region, while Czech LC contributed around 18%. This could either indicate a reduced contribution of Czech LC on the genetic make-up of Hungarian LC, or that the Mušov cemetery, while showing archaeological signs of Lombard occupation, is not a good proxy of the LC population that moved from the Czech Republic to Hungary. The Collegno individuals can instead trace more than 70% of their genetic makeup to LC populations migrating from Hungary, confirming both the high degree of similarity shown by the exploratory analyses and the hypothesis based on archaeological data.

Discussion and Conclusion

In this work, we extracted and analysed complete mitochondrial genomes from 88 Early Medieval individuals sampled in 9 necropolis. Based on archaeological information we have classified these cemeteries as putatively occupied by Lombards or by different medieval communities. These genetic data have been used to explore, for the first time, the genetic variation and structure of these groups, so as to understand whether, and to what extent, different communities found along the route of the Lombard migration may closely resemble each other. Our explorative analyses highlighted a degree of genetic similarity between the LC and NLC communities, which was absolutely expected, given the well-known overall genetic similarity among all European populations. However, a peculiar set of samples, including only LC individuals, appeared always well separated from the rest of the samples. In most cases, these individuals were also associated with burials with Lombard grave goods. This particular association, together with the presence in this cluster of haplogroups that reach high frequency in Northern European populations, suggest a possible link between this core group of individuals and the proposed homeland of the Lombards. In short, people found in burials associated with the Lombard material culture show a mixture of Northern European and Central European genetic features, whereas the Northern component is rare or absent altogether in the burials not associated with the Lombard material culture. The most interesting cases emerge when genetics can help to provide a better understanding in regard to social articulation of the Lombard groups, when the archaeological data are open to different interpretations. This happened, for instance, in Collegno, where the graves of these LC individuals were devoid of Lombard cultural markers, but placed among other burials rich of Lombard material culture. In this case, this pattern may suggest the presence of individuals from the lowest social level bur still members of the Lombard community, rather than local/non Germanic people. We also estimated the relative contribution of each LC sampled group along the Lombard migration route from Pannonia to Italy. The most interesting result, besides the strong support for a model accounting for admixture between LC and local NLC populations, was the

high degree of genetic resemblance between the LC cemeteries in Hungary and Collegno, in Italy. We hence estimated that about the 70% of the lineages found in Collegno actually derived from the Hungarian LC groups, in agreement with previous archaeological and historical hypotheses. This supports the idea that the spread of Lombards I Italy actually involved movements of fairly large numbers of people, who gave a substantial contribution to the gene pool of the populations. This is even more remarkable thinking that, in many studied cases, military invasions are movements of males, and hence do not have consequences at the mtDNA level. Here, instead, we have evidence of changes in the composition of the mtDNA pool of an Italian population, supporting the view that immigration from Central Europe involved females as well as males. While nuclear data would help to elucidate these past populations' dynamics, the resolution provided by the mitochondrial genomes presented in this study, combined with detailed archaeological information, has turned out to be able to significantly improve our knowledge about different aspects of the complex pattern of migrations involving Lombard populations.

BIBLIOGRAPHY

Achtman M, Azuma T, Berg DE, Ito Y, Morelli G, Pan Z-J, Suerbaum S, Thompson SA, van der Ende A, van Doorn L-J. 1999. Recombination and clonal groupings within Helicobacter pylori from different geographical regions. Mol. Microbiol. 32:459–470.

Achtman M, Suerbaum S. 2001. Helicobacter pylori: molecular and cellular biology. Horizon Scientific Press

Ajmone-Marsan P, Garcia JF, Lenstra JA. 2010. On the origin of cattle: How aurochs became cattle and colonized the world. Evol. Anthropol. Issues, News, Rev. 19:148–157.

Alberdi A, Gilbert MTP, Razgour O, Aizpurua O, Aihartza J, Garin I. 2015. Contrasting population-level responses to Pleistocene climatic oscillations in an alpine bat revealed by complete mitochondrial genomes and evolutionary history inference. J. Biogeogr. 42:1689–1700.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19:1655–1664.

Allentoft ME, Sikora M, Sjögren K-G, Rasmussen S, Rasmussen M, Stenderup J, Damgaard PB, Schroeder H, Ahlström T, Vinner L, et al. 2015. Population genomics of Bronze Age Eurasia. Nature 522:167–172.

Alm RA, Ling L-SL, Moir DT, King BL, others. 1999. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen Helicobacter pylori. Nature 397:176.

Alt KW, Knipper C, Peters D, Müller W, Maurer A-F, Kollig I, Nicklisch N, Müller C, Karimnia S, Brandt G, et al. 2014. Lombards on the Move – An Integrative Study of the Migration Period Cemetery at Szólád, Hungary.Bondioli L, editor. PLoS One 9:e110793.

Ammerman AJ, Cavalli-Sforza LL. 1984. The Neolithic transition and the genetics of populations in Europe. Princeton University Press

Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, et al. 2015. A global reference for human genetic variation. Nature 526:68–74.

Balanovsky O, Dibirova K, Dybo A, Mudrak O, Frolova S, Pocheshkhova E, Haber M, Platt D, Schurr T, Haak W, et al. 2011. Parallel evolution of genes and languages in the Caucasus region. Mol. Biol. Evol. 28:2905–2920.

Bamshad MJ, Watkins WS, Dixon ME, Jorde LB, Rao BB, Naidu JM, Prasad B V, Rasanayagam A, Hammer MF. 1998. Female gene flow stratifies Hindu castes. Nature 395:651–652.

Bandelt HJ, Forster P, Sykes BC, Richards MB. 1995. Mitochondrial portraits of human populations using median networks. Genetics 141:743–753.

Barbieri C, Vicente M, Rocha J, Mpoloka SW, Stoneking M, Pakendorf B. 2013. Ancient Substructure in Early mtDNA Lineages of Southern Africa. Am. J. Hum. Genet. 92:285–292.

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

Barbujani G, Pilastro a. 1993. Genetic evidence on origin and dispersal of human populations speaking languages of the Nostratic macrofamily. Proc. Natl. Acad. Sci. U. S. A. 90:4670–4673.

Barker G, Goucher C. 2015. The Cambridge World History: Volume 2, A World with Agriculture, 12,000 BCE--500 CE. Cambridge University Press

Beaumont MA. 2008. Joint determination of topology, divergence time, and immigration in population trees.

Beaumont MA, Zhang W, Balding DJ. 2002. Approximate Bayesian computation in population genetics. Genetics 162:2025–2035.

Belle EMS, Barbujani G. 2007. Worldwide analysis of multiple microsatellites: Language diversity has a detectable influence on DNA diversity. Am. J. Phys. Anthropol. 133:1137–1146.

Bellwood P. 2001. Early agriculturalist population diasporas? Farming, languages, and genes. Annu. Rev. Anthropol. 30:181–207.

Bellwood P. 2014. First migrants: ancient migration in global perspective. John Wiley & Sons

Benazzo A, Panziera A, Bertorelle G. 2015. 4P: Fast computing of population genetics statistics from large DNA polymorphism panels. Ecol. Evol. 5:172–175.
Di Benedetto G, Ergüven A, Stenico M, Castrì L, Bertorelle G, Togan I, Barbujani G. 2001. DNA diversity and population admixture in Anatolia. Am. J. Phys. Anthropol. 115:144–156.
Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59.
Bertorelle G, Benazzo A, Mona S. 2010. ABC as a flexible framework to estimate demography over space and time: Some cons, many pros. Mol. Ecol. 19:2609–2625.
Bhumisak J. 1990. The origin of Siam, Tai, Laos and Khmer. Bangkok Textb. Proj. Found.
Blum SD. 2002. Margins and centers: A decade of publishing on China's ethnic minorities. J. Asian Stud. 61:1287–1310.
Boyd R, Richerson PJ. 1985. Culture and the evolutionary process. University of Chicago press
Breiman L. 2001. Random forests. Mach. Learn. 45:5–32.
Breurec S, Michel R, Seck A, Brisse S, Côme D, Dieye FB, Garin B, Huerre M, Mbengue M, Fall C, et al. 2012. Clinical relevance of cagA and vacA gene polymorphisms in Helicobacter pylori isolates from Senegalese patients. Clin. Microbiol. Infect. 18:153–159.
Brigham-grette J, Anderson PM, Lozhkin A V., Glushkova OY. 2004. Paleoenvironmental Conditions in Western Beringia before and during the Last Glacial Maximum.
Burger J, Kirchner M, Bramanti B, Haak W, Thomas MG. 2007. Absence of the lactase-persistence-associated allele in early Neolithic Europeans. Proc. Natl. Acad. Sci. 104:3736–3741.
Busby GBJ, Hellenthal G, Montinaro F, Tofanelli S, Bulayeva K, Rudan I, Zemunik T, Hayward C, Toncheva D, Karachanak-Yankova S, et al. 2015. The Role of Recent Admixture in Forming the Contemporary West Eurasian Genomic Landscape. Curr. Biol. 25:2518–2526.
Cai X, Qin Z, Wen B, Xu S, Wang Y, Lu Y, Wei L, Wang C, Li S, Huang X, et al. 2011. Human migration through bottlenecks from Southeast Asia into East Asia during last glacial maximum revealed by Y chromosomes. PLoS One 6.
Cann RL, Stoneking M, Wilson AC. 1987. Mitochondrial DNA and human evolution. Nature 325:31–36.
Castellano S, Parra G, Sanchez-Quinto FA, Racimo F, Kuhlwilm M, Kircher M, Sawyer S, Fu Q, Heinze A, Nickel B, et al. 2014. Patterns of coding variation in the complete exomes of three Neandertals. Proc. Natl. Acad. Sci. 111:6666–6671.
Cavalli-Sforza LL. 1998. The DNA revolution in population genetics. Trends Genet. 14:60–65.
Cavalli-Sforza LL, Feldman MW. 1981. Cultural transmission and evolution: a quantitative approach. Princeton University Press
Cavalli-Sforza LL, Feldman MW. 2003. The application of molecular genetic approaches to the study of human evolution. Nat. Genet. 33:266–275.
Charoenmuang T. 2001. Khon Mueang. Local Gov. Stud. Proj. Fac. Soc. Sci. Chiang Mai Univ. Chiang Mai.
Chikhi L, Nichols R a, Barbujani G, Beaumont M a. 2002. Y genetic data support the Neolithic demic diffusion model. Proc. Natl. Acad. Sci. U. S. A. 99:11008–11013.
Chomsky N. 1993. Lectures on government and binding: The Pisa lectures. Walter de Gruyter
Chomsky N. 2005. Three factors in language design. Linguist. Inq. 36:1–22.
Condominas G. 1990. From Lawa to Mon, from Saa'to Thai. Canberra Aust. Natl. Univ.
Creanza N, Feldman MW. 2016. Worldwide genetic and cultural change in human evolution. Curr. Opin. Genet. Dev. 41:85–92.
Creanza N, Ruhlen M, Pemberton TJ, Rosenberg NA, Feldman MW, Ramachandran S, Atkinson QD, Hunley K. 2015. A comparison of worldwide phonemic and genetic variation in human populations. Proc. Natl. Acad. Sci. U. S. A. 112:1265–1272.
Csákyová V, Szécsényi-Nagy A, Csősz A, Nagy M, Fusek G, Langó P, Bauer M, Mende BG, Makovický P, Bauerová M. 2016. Maternal Genetic Composition of a Medieval Population from

a Hungarian-Slavic Contact Zone in Central Europe.Hofreiter M, editor. PLoS One 11:e0151206.
Darwin C. 1859. On the Origin of Species by Means of Natural Selection. London: Murray
Darwin C. 1968. On the origin of species by means of natural selection. 1859. London Murray Google Sch.
Diamond J, Bellwood P. 2003. Farmers and their languages: the first expansions. Science (80-.). 300:597–603.
Didelot X, Falush D. 2007. Inference of bacterial microevolution using multilocus sequence data. Genetics 175:1251–1266.
van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. 2014. Ten years of next-generation sequencing technology. Trends Genet. 30:418–426.
van Etten J. 2012. gdistance: Distances and routes on geographical grids. URL http//CRAN. R-project. org/package= gdistance. R Packag. version:1.
Excoffier L, Lischer HEL. 2010. Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. Mol. Ecol. Resour. 10:564–567.
Falush D, Kraft C, Taylor NS, Correa P, Fox JG, Achtman M, Suerbaum S. 2001. Recombination and mutation during long-term gastric colonization by Helicobacter pylori: Estimates of clock rates, recombination size, and minimal age. Proc. Natl. Acad. Sci. 98:15056–15061.
Falush D, Wirth T, Linz B, Pritchard JK, Stephens M, Kidd M, Blaser MJ, Graham DY, Vacher S, Perez-Perez GI, et al. 2003. Traces of Human Migrations in Helicobacter pylori Populations. Science (80-.). 299:1582–1585.
Feldman RA. 2001. Epidemiologic observations and open questions about disease and infection caused by Helicobacter pylori. Achtman M, Serbaum S. Helicobacter pylori Mol. Cell. Biol. Wymondham Horiz. Sci.:29–51.
de Filippo C, Bostoen K, Stoneking M, Pakendorf B. 2012. Bringing together linguistic and genetic evidence to test the Bantu expansion. Proc. R. Soc. B Biol. Sci. 279:3256–3263.
Fu Q, Hajdinjak M, Moldovan OT, Constantin S, Mallick S, Skoglund P, Patterson N, Rohland N, Lazaridis I, Nickel B, et al. 2015. An early modern human from Romania with a recent Neanderthal ancestor. Nature 524:216–219.
Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, Johnson PLF, Aximu-Petri A, Prüfer K, de Filippo C, et al. 2014. Genome sequence of a 45,000-year-old modern human from western Siberia. Nature 514:445–449.
Fu Q, Mittnik A, Johnson PLF, Bos K, Lari M, Bollongino R, Sun C, Giemsch L, Schmitz R, Burger J, et al. 2013. A revised timescale for human evolution based on ancient mitochondrial genomes. Curr. Biol. 23:553–559.
Fuselli S. 2003. Mitochondrial DNA Diversity in South America and the Genetic History of Andean Highlanders. Mol. Biol. Evol. 20:1682–1691.
Geary PJ. 2002. The myth of nations: the medieval origins of Europe. Princeton University Press Princeton
Ghirotto S, Tassi F, Fumagalli E, Colonna V, Sandionigi A, Lari M, Vai S, Petiti E, Corti G, Rizzi E, et al. 2013. Origins and Evolution of the Etruscans' mtDNA.Hawks J, editor. PLoS One 8:e55519.
Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H, Ch'ang L-Y, Huang W, Liu B, Shen Y, et al. 2003. The International HapMap Project. Nature 426:789.
Giostra C. 2011. Goths and Lombards in Italy: the potential of archaeology with respect to ethnocultural identification. Post-Classical Archaeol. 1:7–36.
Goebel T, Waters MR, O'Rourke DH. 2008. The late Pleistocene dispersal of modern humans in the Americas. Science 319:1497–1502.
Golden H. 2006. Kayaks of Greenland: The History and Development of the Greenlandic Hunting Kayak, 1600-2000. White House Grocery Press
Graf KE. 2009. "The Good, the Bad, and the Ugly": evaluating the radiocarbon chronology of the

middle and late Upper Paleolithic in the Enisei River valley, south-central Siberia. J. Archaeol. Sci. 36:694–707.
Green RE, Briggs AW, Krause J, Prüfer K, Burbano HA, Siebauer M, Lachmann M, Pääbo S. 2009. The Neandertal genome and ancient DNA authenticity. EMBO J. 28:2494–2502.
Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MHY, et al. 2010. A Draft Sequence of the Neandertal Genome. Science (80-.). 328:710–722.
Grundy-Warr C, Huang S, Wong PP. 2003. Tropical geography: Research and reflections. Singap. J. Trop. Geogr. 24:1–5.
Guardiano C, Longobardi G. 2005. Parametric comparison and language taxonomy. Gramm. Parametr. Var.:149–174.
Günther T, Valdiosera C, Malmström H, Ureña I, Rodriguez-Varela R, Sverrisdóttir ÓO, Daskalaki EA, Skoglund P, Naidoo T, Svensson EM, et al. 2015. Ancient genomes link early farmers from Atapuerca in Spain to modern-day Basques. Proc. Natl. Acad. Sci. U. S. A. 112:11917–11922.
Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, Karthikeyan S, Iles L, Pollard MO, Choudhury A, et al. 2015. The African Genome Variation Project shapes medical genetics in Africa. Nature 517:327–332.
Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson K, et al. 2015. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature 522:207–211.
Hamilton G, Stoneking M, Excoffier L. 2005. Molecular analysis reveals tighter social regulation of immigration in patrilocal populations than in matrilocal populations. Proc. Natl. Acad. Sci. U. S. A. 102:7476–7480.
Hammer MF, Karafet T, Rasanayagam A, Wood ET, Altheide TK, Jenkins T, Griffiths RC, Templeton AR, Zegura SL. 1998. Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. Mol. Biol. Evol. 15:427–441.
He JD, Peng MS, Quang HH, Dang KP, Trieu AV, Wu SF, Jin JQ, Murphy RW, Yao YG, Zhang YP. 2012. Patrilineal perspective on the Austronesian diffusion in Mainland Southeast Asia. PLoS One 7:1–10.
van Heerwaarden J, Doebley J, Briggs WH, Glaubitz JC, Goodman MM, de Jesus Sanchez Gonzalez J, Ross-Ibarra J. 2011. Genetic signals of origin, spread, and introgression in a large sample of maize landraces. Proc. Natl. Acad. Sci. 108:1088–1092.
Higuchi R, Bowman B, Freiberger M, Ryder O a, Wilson AC. 1984. DNA sequences from the quagga, an extinct member of the horse family. Nature 312:282–284.
Hoffecker JF, Elias SA. 2007. Human ecology of Beringia. Columbia University Press
Hofreiter M, Serre D, Poinar HN, Kuch M, Pääbo S. 2001. Ancient DNA. Nat. Rev. Genet. 2:353–359.
Howey MCL. 2007. Using multi-criteria cost surface analysis to explore past regional landscapes: a case study of ritual activity and social interaction in Michigan, AD 1200-1600. J. Archaeol. Sci. 34:1830–1846.
Human Genome Sequencing Consortium I. 2004. Finishing the euchromatic sequence of the human genome. Nature 431:931–945.
Jesse R, Véla E, Pfenninger M. 2011. Phylogeography of a Land Snail suggests Trans-Mediterranean Neolithic transport. PLoS One 6.
Jobling MA, Tyler-Smith C. 2003. The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet 4:598–612.
Johne K-P. 2008. Die Langobarden in den Schriftquellen bis zu den Markomannenkriegen. Kult. Mitteleuropa Langobarden--Awaren--Slawen. Akten der Int. Tagung Bonn vom 25:43–50.
Jolley KA, Maiden MC. 2010. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics 11:595.
Jombart T. 2008. Adegenet: A R package for the multivariate analysis of genetic markers. Bioinformatics 24:1403–1405.

Jombart T, Devillard S, Balloux F. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 11:94.
Jones E, Skirnisson K, McGovern T, Gilbert M, Willerslev E, Searle J. 2012. Fellow travellers: a concordance of colonization patterns between mice and men in the North Atlantic region. BMC Evol. Biol. 12:35.
Jones EP, Eager HM, Gabriel SI, Jóhannesdóttir F, Searle JB. 2013. Genetic tracking of mice and other bioproxies to infer human history. Trends Genet. 29:298–308.
Jónsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L. 2013. MapDamage2.0: Fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics 29:1682–1684.
Jurinke C, Denissenko MF, Oeth P, Ehrich M, van den Boom D, Cantor CR. 2005. A single nucleotide polymorphism based approach for the identification and characterization of gene expression modulation using MassARRAY. Mutat. Res. Mol. Mech. Mutagen. 573:83–95.
Kampuansai J, Völgyi A, Kutanan W, Kangwanpong D, Pamjav H. 2017. Autosomal STR variations reveal genetic heterogeneity in the Mon-Khmer speaking group of Northern Thailand. Forensic Sci. Int. Genet. 27:92–99.
Kassambara A, Mundt F. 2016. Factoextra: extract and visualize the results of multivariate data analyses. R Packag. version 1.
Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol. Biol. Evol. 30:772–780.
Kayser M, Brauer S, Cordaux R, Casto A, Lao O, Zhivotovsky LA, Moyse-Faurie C, Rutledge RB, Schiefenhoevel W, Gil D, et al. 2006. Melanesian and Asian origins of Polynesians: mtDNA and Y chromosome gradients across the Pacific. Mol. Biol. Evol. 23:2234–2244.
Keller A. 2007. Drosophila melanogaster's history as a human commensal. Curr. Biol. 17:R77–R81.
Kimura M, Weiss GH. 1964. The stepping stone model of population structure and the decrease of genetic correlation with distance. Genetics 49:561.
Kivisild T. 2015. Maternal ancestry and population history from whole mitochondrial genomes. Investig. Genet. 6:3.
Kline MA, Boyd R. 2010. Population size predicts technological complexity in Oceania. Proc. R. Soc. B Biol. Sci. 277:2559–2564.
Kloss-Brandstätter A, Pacher D, Schönherr S, Weissensteiner H, Binna R, Specht G, Kronenberg F. 2011. HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. Hum. Mutat. 32:25–32.
Kutanan W, Ghirotto S, Bertorelle G, Srithawong S, Srithongdaeng K, Pontham N, Kangwanpong D. 2014. Geography has more influence than language on maternal genetic structure of various northeastern Thai ethnicities. J. Hum. Genet. 59:1–9.
Kutanan W, Kampuansai J, Brunelli A, Ghirotto S, Pittayaporn P, Ruangchai S, Schröder R, Macholdt E, Srikummool M, Kangwanpong D, et al. 2017. New insights from Thailand into the maternal genetic history of Mainland Southeast Asia.
Kutanan W, Kampuansai J, Colonna V, Nakbunlung S, Lertvicha P, Seielstad M, Bertorelle G, Kangwanpong D. 2011. Genetic affinity and admixture of northern Thai people along their migration route in northern Thailand: evidence from autosomal STR loci. J. Hum. Genet. 56:130–137.
Kutanan W, Kampuansai J, Fuselli S, Nakbunlung S, Seielstad M, Bertorelle G, Kangwanpong D. 2011. Genetic structure of the Mon-Khmer speaking groups and their affinity to the neighbouring Tai populations in Northern Thailand. BMC Genet. 12:56.
Kutanan W, Kampuansai J, Nakbunlung S, Lertvicha P, Seielstad M, Bertorelle G, Kangwanpong D. 2011. Genetic structure of khon mueang populations along a historical yuan migration route in northern thailand. Chiang Mai J. Sci. 38:295–305.
Kutanan W, Kampuansai J, Srikummool M, Kangwanpong D, Ghirotto S, Brunelli A, Stoneking M. 2017. Complete mitochondrial genomes of Thai and Lao populations indicate an ancient origin

of Austroasiatic groups and demic diffusion in the spread of Tai–Kadai languages. Hum. Genet. 136:85–98.
Kuzmin Y, Keates SG. 2013. Dynamics of Siberian Paleolithic Complexes (Based on Analysis of Radiocarbon Records): The 2012 State-of-the-Art. Radiocarbon 55:1314–1321.
Kuzmin Y V. 2008. Siberia at the Last Glacial Maximum: Environment and archaeology.
Lachance J, Tishkoff SA. 2013. SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. BioEssays 35:780–786.
Laland KN, Odling-Smee J, Myles S. 2010. How culture shaped the human genome: bringing genetics and the human sciences together. Nat. Rev. Genet. 11:137–148.
Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.
LaPolla RJ. 2001. The role of migration and language contact in the development of the Sino-Tibetan language family. Areal Diffus. Genet. Inherit. Case Stud. Lang. Chang.:225–254.
Larson G, Cucchi T, Fujita M, Matisoo-Smith E, Robins J, Anderson A, Rolett B, Spriggs M, Dolman G, Kim T-H, et al. 2007. Phylogeny and ancient DNA of Sus provides insights into neolithic expansion in Island Southeast Asia and Oceania. Proc. Natl. Acad. Sci. 104:4834–4839.
Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, Fernandes D, Novak M, Gamarra B, Sirak K, et al. 2016. Genomic insights into the origin of farming in the ancient Near East. Nature 536:419–424.
Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M, et al. 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 513:409–413.
Lewis MP, Simons GF, Fennig CD, others. 2009. Ethnologue: Languages of the world. SIL international Dallas, TX
Li H, Wen B, Chen S-J, Su B, Pramoonjago P, Liu Y, Pan S, Qin Z, Liu W, Cheng X, et al. 2008. Paternal genetic affinity between Western Austronesians and Daic populations. BMC Evol. Biol. 8:146.
Linz B, Balloux F, Moodley Y, Manica A, Liu H, Roumagnac P, Falush D, Stamer C, Prugnolle F, van der Merwe SW, et al. 2007. An African origin for the intimate association between humans and Helicobacter pylori. Nature 445:915–918.
Linz B, Vololonantenainab CRR, Seck A, Carod JF, Dia D, Garin B, Ramanampamonjy RM, Thiberge JM, Raymond J, Breurec S. 2014. Population genetic structure and isolation by distance of Helicobacter pylori in Senegal and Madagascar. PLoS One 9:1–7.
Llorente MG, Jones ER, Eriksson A, Siska V, Arthur KW, Arthur JW, Curtis MC, Stock JT, Coltorti M, Pieruccini P, et al. 2015. Ancient Ethiopian genome reveals extensive Eurasian admixture in Eastern Africa. Science (80-.). 350:820–822.
Longobardi G, Ghirotto S, Guardiano C, Tassi F, Benazzo A, Ceolin A, Barbujani G. 2015. Across language families: Genome diversity mirrors linguistic variation within Europe. Am. J. Phys. Anthropol. 157:630–640.
Longobardi G, Guardiano C. 2009. Evidence for syntax as a signal of historical relatedness. Lingua 119:1679–1706.
Luca Cavallisforza L, Piazzat A, Menozzif P, Mountain J. 1988. Reconstruction of human evolution: Bringing together genetic, archaeological, and linguistic data (origin of modern humans/phylogenetic trees/paleoanthropology). Evolution (N. Y). 85:6002–6006.
Maixner F, Krause-Kyora B, Turaev D, Herbig A, Hoopmann MR, Hallows JL, Kusebauch U, Vigl EE, Malfertheiner P, Megraud F, et al. 2016. The 5300-year-old Helicobacter pylori genome of the Iceman. Science (80-.). 351:162–165.
Malaty HM, Graham DY. 1994. Importance of childhood socioeconomic status on the current prevalence of Helicobacter pylori infection. Gut 35:742–745.
Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt

S, Tandon A, et al. 2016. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature 538:201–206.
Mantel N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Res. 27:209–220.
Marshall B, Warren JR. 1984. Unidentified Curved Bacilli In The Stomach Of Patients With Gastritis And Peptic Ulceration. Lancet 323:1311–1315.
Matisoo-Smith E, Robins JH. 2004. Origins and dispersals of Pacific peoples: evidence from mtDNA phylogenies of the Pacific rat. Proc. Natl. Acad. Sci. U. S. A. 101:9167–9172.
Mcrae BH. 2006. Isolation By Resistance. Evolution (N. Y). 60:1551–1561.
McRae BH, Beier P. 2007. Circuit theory predicts gene flow in plant and animal populations. Proc. Natl. Acad. Sci. United States Am. Sci. 104:19885–19890.
Menozzi P, Piazza A, Cavalli-Sforza LL. 1978. Synthetic Maps of Human Gene Frequencies in Europeans. Science (80-.). 201:786–792.
Meyer M, Kircher M, Gansauge M, Li H, Racimo F, Siebauer M, Green RE, Bryc K, Briggs AW, Stenzel U, et al. 2012. A High-Coverage Genome Sequence from an Archaic Denisovan Individual A High-Coverage Genome Sequence from an Archaic Denisovan Individual A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Sci. (New York, NY) 222:1–14.
Michaels GS, Hauswirth WW, Laipis PJ. 1982. Mitochondrial DNA copy number in bovine oocytes and somatic cells. Dev. Biol. 94:246–251.
Modi A, Tassi F, Susca RR, Vai S, Rizzi E, Bellis G De, Lugliè C, Gonzalez Fortes G, Lari M, Barbujani G, et al. 2017. Complete mitochondrial sequences from Mesolithic Sardinia. Sci. Rep. 7:42869.
Monot M, Honoré N, Garnier T, Araoz R, Coppée J-Y, Lacroix C, Sow S, Spencer JS, Truman RW, Williams DL, et al. 2005. On the Origin of Leprosy. Science (80-.). 308:1040–1042.
Monot M, Honoré N, Garnier T, Zidane N, Sherafi D, Paniz-Mondolfi A, Matsuoka M, Taylor GM, Donoghue HD, Bouwman A, et al. 2009. Comparative genomic and phylogeographic analysis of Mycobacterium leprae. Nat. Genet. 41:1282–1289.
Montano V, Didelot X, Foll M, Linz B, Reinhardt R, Suerbaum S, Moodley Y, Jensen JD. 2015. Worldwide population structure, long-term demography, and local adaptation of Helicobacter pylori. Genetics 200:947–963.
Moodley Y, Linz B. 2009. Helicobacter pylori sequences reflect past human migrations. Genome Dyn. 6:62–74.
Moodley Y, Linz B, Yamaoka Y, Windsor HM, Breurec S, Wu J-YY, Maady A, Bernhöft S, Thiberge J-MM, Phuanukoonnon S, et al. 2009. The peopling of the Pacific from a bacterial perspective. Science (80-.). 323:527–530.
Mullis KB, Faloona FA. 1987. Specific Synthesis of DNA in Vitro via a Polymerase-Catalyzed Chain Reaction. Methods Enzymol. 155:335–350.
Murray C, Huerta-Sanchez E, Casey F, Bradley DG. 2010. Cattle demographic history modelled from autosomal sequence variation. Philos. Trans. R. Soc. B Biol. Sci. 365:2531–2539.
Myles S, Boyko AR, Owens CL, Brown PJ, Grassi F, Aradhya MK, Prins B, Reynolds A, Chia J-M, Ware D, et al. 2011. Genetic structure and domestication history of the grape. Proc. Natl. Acad. Sci. 108:3530–3535.
Nell S, Eibach D, Montano V, Maady A, Nkwescheu A, Siri J, Elamin WF, Falush D, Linz B, Achtman M, et al. 2013. Recent Acquisition of Helicobacter pylori by Baka Pygmies.McVean G, editor. PLoS Genet. 9:e1003775.
Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E. 2017. Tracing the peopling of the world through genomics. Nature 541:302–310.
Ning C, Yan S, Hu K, Cui Y-Q, Jin L. 2016. Refined phylogenetic structure of an abundant East Asian Y-chromosomal haplogroup O*-M134. Eur. J. Hum. Genet. 24:307.
Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S,

Nelson MR, et al. 2008. Genes mirror geography within Europe. Nature 456:98–101.
Novembre J, Ramachandran S. 2011. Perspectives on Human Population Structure at the Cusp of the Sequencing Era. Annu. Rev. Genomics Hum. Genet. 12:245–274.
Oksanen J, Kindt R, Legendre P, O'Hara B, Simpson GL, Solymos P, Stevens MHH, Wagner H. 2008. The vegan Package. Community Ecology package. :1–174.
Omrak A, Günther T, Valdiosera C, Svensson EM, Malmström H, Kiesewetter H, Aylward W, Storå J, Jakobsson M, Götherström A. 2016. Genomic Evidence Establishes Anatolia as the Source of the European Neolithic Gene Pool. Curr. Biol. 26:270–275.
van Oven M, Kayser M. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum. Mutat. 30:E386–E394.
Pääbo S. 1985. Molecular cloning of Ancient Egyptian mummy DNA. Nature 314:644–645.
Pakendorf B. 2014. Coevolution of languages and genes. Curr. Opin. Genet. Dev. 29:39–44.
Pakendorf B, Stoneking M. 2005. Mitochondrial Dna and Human Evolution. Annu. Rev. Genomics Hum. Genet. 6:161–165.
Palopoli MF, Fergus DJ, Minot S, Pei DT, Simison WB, Fernandez-Silva I, Thoemmes MS, Dunn RR, Trautwein M. 2015. Global divergence of the human follicle mite Demodex folliculorum : Persistent associations between host ancestry and mite lineages. Proc. Natl. Acad. Sci. 112:15958–15963.
Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. PLoS Genet. 2:2074–2093.
Penth H. 2000. A brief history of L{\=a}n N{\=a}: civilizations of north Thailand. Silkworm Books
Petkova D, Novembre J, Stephens M. 2015. Visualizing spatial population structure with estimated effective migration surfaces. Nat. Genet. 48:94–100.
Pittayaporn P. 2014. Layers of Chinese loanwords in proto-southwestern Tai as evidence for the dating of the spread of southwestern Tai. Manusya J Humanit 20:47–68.
Pitulko V V., Nikolskiy PA, Girya EY, Basilyan AE, Tumskoy VE, Koulakov SA, Astanikv SN, Pavlova EY, Anisimov MA. 2004. The Yana RHS Site: Humans in the Arctic Before the Last Glacial Maximum. Science (80-.). 303:52–56.
Poznik GD, Xue Y, Mendez FL, Willems TF, Massaia A, Wilson Sayres MA, Ayub Q, McCarthy SA, Narechania A, Kashin S, et al. 2016. Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. Nat. Genet. 48:593–599.
Prithchard JK, Stepehens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data: dominant markers and null alleles. Genetics 155:945–959.
Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, et al. 2013. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505:43–49.
Pudlo P, Marin J-M, Estoup A, Cornuet J-M, Gautier M, Robert CP. 2016. Reliable ABC model choice via random forests. Bioinformatics 32:859–866.
Pugach I, Matveev R, Spitsyn V, Makarov S, Novgorodov I, Osakovsky V, Stoneking M, Pakendorf B. 2016. The Complex Admixture History and Recent Southern Origins of Siberian Populations. Mol. Biol. Evol. 33:msw055.
Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am. J. Hum. Genet. 81:559–575.
Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, Rasmussen S, Stafford TW, Orlando L, Metspalu E, et al. 2014. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature 505:87–91.
Raghavan M, Steinrücken M, Harris K, Schiffels S, Rasmussen S, DeGiorgio M, Albrechtsen A, Valdiosera C, Ávila-Arcos MC, Malaspinas A-S, et al. 2015. Genomic evidence for the Pleistocene and recent population history of Native Americans. Science 349:aab3884.
Ramachandran S, Deshpande O, Roseman CC, Rosenberg N a, Feldman MW, Cavalli-Sforza LL.

2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc. Natl. Acad. Sci. U. S. A. 102:15942–15947.

Ranciaro A, Campbell MC, Hirbo JB, Ko W, Froment A, Anagnostou P, Kotze MJ, Ibrahim M, Nyambo T, Omar SA, et al. 2014. Genetic Origins of Lactase Persistence and the Spread of Pastoralism in Africa. Am. J. Hum. Genet. 94:496–510.

Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, Metspalu M, Metspalu E, Kivisild T, Gupta R, et al. 2010. Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature 463:757–762.

Rasmussen M, Sikora M, Albrechtsen A, Korneliussen TS, Moreno-Mayar JV, Poznik GD, Zollikofer CPE, Ponce de León MS, Allentoft ME, Moltke I, et al. 2015. The ancestry and affiliations of Kennewick Man. Nature :1–10.

Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PLF, et al. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468:1053–1060.

Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, Parra M V, Rojas W, Duque C, Mesa N, et al. 2012. Reconstructing Native American population history. Nature 488:370–374.

Renfrew C. 1992. Archaeology, genetics and linguistic diversity. Man:445–478.

Renfrew C, Boyle K V. 2000. Archaeogenetics: DNA and the population prehistory of Europe. McDonald Inst of Archeological

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928–933.

Sawyer S, Renaud G, Viola B, Hublin J-J, Gansauge M-T, Shunkov M V., Derevianko AP, Prüfer K, Kelso J, Pääbo S. 2015. Nuclear and mitochondrial DNA sequences from two Denisovan individuals. Proc. Natl. Acad. Sci. 112:201519905.

Schliesinger J. 2015a. Ethnic groups of Thailand: non-Tai-speaking peoples. Booksmango

Schliesinger J. 2015b. Tai Groups of Thailand Vol 1: Introduction and Overview. Booksmango

Schwarz S, Morelli G, Kusecek B, Manica A, Balloux F, Owen RJ, Graham DY, van der Merwe S, Achtman M, Suerbaum S. 2008. Horizontal versus Familial Transmission of Helicobacter pylori.Maiden MCJ, editor. PLoS Pathog. 4:e1000180.

Searle JB, Jones CS, Gunduz I, Scascitelli M, Jones EP, Herman JS, Rambau RV, Noble LR, Berry R., Gimenez MD, et al. 2009. Of mice and (Viking?) men: phylogeography of British and Irish house mice. Proc. R. Soc. B Biol. Sci. 276:201–207.

Seielstad MT, Minch E, Cavalli-Sforza LL. 1998. Genetic evidence for a higher female migration rate in humans. Nat. Genet. 20:278–280.

Sengupta S, Zhivotovsky LA, King R, Mehdi SQ, Edmonds CA, Chow C-ET, Lin AA, Mitra M, Sil SK, Ramesh A, et al. 2006. Polarity and Temporality of High-Resolution Y-Chromosome Distributions in India Identify Both Indigenous and Exogenous Expansions and Reveal Minor Genetic Influence of Central Asian Pastoralists. Am. J. Hum. Genet. 78:202–221.

Shennan S. 2001. Demography and Cultural Innovation: a Model and its Implications for the Emergence of Modern Human Culture. Cambridge Archaeol. J. 11:5–16.

Shi H, Qi X, Zhong H, Peng Y, Zhang X, Ma RZ, Su B. 2013. Genetic Evidence of an East Asian Origin and Paleolithic Northward Migration of Y-chromosome Haplogroup N. PLoS One 8:1–9.

Skoglund P, Northoff BH, Shunkov M V., Derevianko AP, Pääbo S, Krause J, Jakobsson M. 2014. Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. Proc. Natl. Acad. Sci. 111:2229–2234.

Skoglund P, Posth C, Sirak K, Spriggs M, Valentin F, Bedford S, Clark GR, Reepmeyer C, Petchey F, Fernandes D, et al. 2016. Genomic insights into the peopling of the Southwest Pacific. Nature 538:510–513.

Skoglund P, Reich D. 2016. A genomic view of the peopling of the Americas. Curr. Opin. Genet. Dev. 41:27–35.
Skoglund P, Thompson JC, Prendergast ME, Mittnik A, Sirak K, Hajdinjak M, Salie T, Rohland N, Mallick S, Peltzer A, et al. 2017. Reconstructing Prehistoric African Population Structure. Cell 171:59–71.e21.
Slatkin M, Racimo F. 2016. Ancient DNA and human history. Proc. Natl. Acad. Sci. 113:6380–6387.
Sokal RR. 1988. Genetic, geographic, and linguistic distances in Europe. Proc. Natl. Acad. Sci. 85:1722–1726.
Stearns SC, Koella JC. 2008. Evolution in health and disease. Oxford University Press
Stoneking M, Delfin F. 2010. The Human Genetic History of East Asia: Weaving a Complex Tapestry. Curr. Biol. 20:R188–R193.
Stoneking M, Krause J. 2011. Learning about human population history from ancient and modern genomes. Nat. Rev. Genet. 12:603–614.
Suerbaum S, Michetti P. 2002. Helicobacter pylori Infection. N. Engl. J. Med. 347:1175–1186.
Tackney JC, Potter BA, Raff J, Powers M, Watkins WS, Warner D, Reuther JD, Irish JD, O'Rourke DH. 2015. Two contemporaneous mitogenomes from terminal Pleistocene burials in eastern Beringia. Proc. Natl. Acad. Sci. 112:13833–13838.
Tassi F, Ghirotto S, Mezzavilla M, Vilaça ST, De Santi L, Barbujani G. 2015. Early modern human dispersal from Africa: genomic evidence for multiple waves of migration. Investig. Genet. 6:13.
Thorell K, Yahara K, Berthenet E, Lawson DJ, Mikhail J, Kato I, Mendez A, Rizzato C, Bravo MM, Suzuki R, et al. 2017. Rapid evolution of distinct Helicobacter pylori subpopulations in the Americas. PLoS Genet. 13:e1006546.
Tishkoff SA, Williams SM. 2002. Genetic analysis of African populations: human evolution and complex disease. Nat. Rev. Genet. 3:611–621.
Tomb J-F, White O, Kerlavage AR, Clayton RA, others. 1997. The complete genome sequence of the gastric pathogen Helicobacter pylori. Nature 388:539.
Vai S, Ghirotto S, Pilli E, Tassi F, Lari M, Rizzi E, Matas-Lalueza L, Ramirez O, Lalueza-Fox C, Achilli A, et al. 2015. Genealogical Relationships between Early Medieval and Modern Inhabitants of Piedmont.Calafell F, editor. PLoS One 10:e0116801.
Vajda E. 2009. Loanwords in ket. Loanwords World's Lang. A Comp. Handb.:471–495.
Vandkilde H. 2007. Culture and Change in Central European Prehistory 6th to 1st millennium BC. Aarhus Universitetsforlag
Wang CC, Wang LX, Shrestha R, Zhang M, Huang XY, Hu K, Jin L, Li H. 2014. Genetic structure of Qiangic populations residing in the Western Sichuan corridor. PLoS One 9.
Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L. 2010. ABCtoolbox: a versatile toolkit for approximate Bayesian computations. BMC Bioinformatics 11:116.
Weir BS, Cockerham CC. 1984. Estimating *F*-statistics for the analysis of population structure. Evolution (N. Y). 38:1358–1370.
Wen B, Li H, Lu D, Song X, others. 2004. Genetic evidence supports demic diffusion of Han culture. Nature 431:302.
Wirth T, Meyer A, Achtman M. 2005. Deciphering host migrations and origins by means of their microbes. Mol. Ecol. 14:3289–3306.
Yunusbayev B, Metspalu M, Metspalu E, Valeev A, Litvinov S, Valiev R, Akhmetova V, Balanovska E, Balanovsky O, Turdikulova S, et al. 2015. The Genetic Legacy of the Expansion of Turkic-Speaking Nomads across Eurasia. PLoS Genet. 11:5850.
Zhong H, Shi H, Qi X-B, Xiao C-J, Jin L, Ma RZ, Su B. 2010. Global distribution of Y-chromosome haplogroup C reveals the prehistoric migration routes of African exodus and early settlement in East Asia. J. Hum. Genet. 5540:428–435.

RINGRAZIAMENTI

Chiunque sia arrivato a leggere fino a questo punto è in qualche modo responsabile per questo dottorato, il che non è necessariamente una fatto positivo. A causa di ciò, sebbene i ringraziamenti seguenti coinvolgano una manciata di persone in particolare, invito il lettore casuale a non sentirsi esente da un minimo di senso di colpa.

Grazie a Guido Barbujani per aver deciso, più di tre anni fa, che l'aver studiato l'ecologia dei pipistrelli poteva costituire una buona base per occuparmi di genetica umana. Dopo avermi seguito durante un dottorato e un libro spero non si sia ricreduto troppo.

Grazie a Silvia Ghirotto per avermi insegnato l'ABC della genetica di popolazioni e ad avermi dato fiducia, trasformando qualcuno che non aveva idea di cosa fosse uno SNP in una persona che può spiegarlo mentendo in maniera convincente.

Grazie a Francesca Tassi per avermi badato, spronato ed essere venuta in mio aiuto tutte le volte in cui ne avevo bisogno durante questi tre anni.

Grazie ad Andrea Benazzo per la pazienza con cui ha sempre corretto le bestialità nei miei modelli.

Grazie a Giorgio Bertorelle e a Silvia Fuselli per i consigli e i regali di Natale più azzeccati che io abbia mai ricevuto.

Grazie a Roberto per i topi, a Emiliano per l'utile corso di RAD, a Roberta per i dolci, a Gloria per il JC e a Patricia per le lezioni di portoghese.

Grazie a Francesco, Matteo, Valentina, Andrea, Alessandro, Giacomo, Leanne, Luca e Maria per ricordarmi ogni giorno che Bologna è vicina, fiammante e cromata.

Grazie a Mattia, Christian, Carlo, Gianmarco, Sofia, Andrea e Pietro per assecondare le mie passioni cinematografiche, anche se solo una domenica ogni tanto.

Grazie a Luca, Lisa, Riccardo, Lea, Enrico e Stefano per sopportare i miei discorsi tutte le volte che usciamo.

Grazie a Teresa e Flora per molte più cose di quelle che potrei scrivere qui ma, se devo sceglierne due, soprattutto per la parmigiana di melanzane e l'avermi portato al parco 16,000 volte.

Grazie a Marco per essere il fratello Brunelli minore e migliore, senza farmelo pesare poi troppo.

Grazie a Raffaella e Giampaolo, per non aver bruciato la videocassetta di "Nata libera", avermi comprato i fascicoli sui dinosauri e supportato in ogni cosa per quasi trent'anni.

Grazie a Giulia per esserci stata e continuare ad esserci ogni giorno.