Doctoral (Ph.D.) Thesis

UNFOLDING THE EFFECTS OF MIGRATION TO THE ANCESTRY OF ROMANI PEOPLE BASED ON GENOME-WIDE MARKER DATA

Zsolt Bánfai

Head of the Doctoral School:

Dr. Ferenc Gallyas

professor

Supervisor:

Dr. Béla Melegh

professor

Pécs, 2021

Department of Medical Genetics, Clinical Centre, University of Pécs

Introduction

Population genetics based on genome-wide data

Technological advancements in information technology, such as high-performance computing, the development of computer storage, next generation sequencing, microarray-based genotyping allowed scientists the genome-wide comparison of sequences and genotype data of populations. This was a major step towards population genomics. Population genomics allows the study of population genetic phenomena on a large scale, on a much more complex level, investigating simultaneously hundreds of thousands of loci or markers. It also enables the investigation of population structure, population ancestry, processes like population admixture and migration, and therefore allows the revelation and characterization of human historic events. Population genomics has also clinical usages, it allows the investigation how genetic variation relates different diseases and syndromes and how are they responsible generally to the human health.

The development of sequencing and genotyping technologies also facilitated the formation of research groups in academic institutions, of which goal was to utilize these massive data in population genomic studies involving also human historical and clinical purposes. At the same time, these research groups also set their goal to develop new mathematical and statistical methods which are able to handle and exploit the potential of large genetic data, and to overcome the computational and analytical challenges posing these massive data to researchers. These groups often make their software available to anyone for academic research. Such groups are, including but not limited to, the Pritchard Lab of Stanford University led by Jonathan Karl Pritchard, the Tang Lab led by Hua Tang also from Stanford University, the Reich Lab led by David Reich and associated with the Harvard Medical University, the Broad Institute and the Massachusetts Institute of Technology, Daniel Lawson's lab from the University of Bristol, Brian Browning's laboratory from the University of Washington or Manfred Kayser's group from the Erasmus University, Rotterdam.

The continuous work of these research groups enabled the comprehensive study of whole-genome sequencing data and genome-wide marker data by developing novel population stratification, ancestry analysis, haplotype phasing and inferring, identical by state (IBS) and identical by descent (IBD) segment detection as well as admixture analysis methods. Besides of population history purposes, analytical methods for clinical purposes are also developing. One of the first comprehensive software packages for GWA studies was the PLINK software package. PLINK featured all of the basic population genetics statistics, such as fixation index (F_{st}) calculations, testing for Hardy-Weinberg equilibrium, inbreeding coefficient, missingness tests, Mendel errors and many more basic statistics on genotyped biallelic marker data. Furthermore, it is capable of linkage disequilibrium (LD) calculations, IBD and

IBS calculations, testing for epistatic interactions, population stratification and of course one can conduct GWA studies both for quantitative traits and basic case/control investigations. Of course, many developers aim both clinical and population historical purposes and creates robust algorithms for both purposes. In case of the EIGENSOFT software package, developed by the colleagues of the David Reich Lab, there are two population stratification methods available, which meet the needs of both of these areas.

Romani people

The Romani people (also known as Roma or Gypsies) are a transnational, diasporic ethnic group from South Asia that includes several socially and genetically diverse groups. The census size of the Roma population is currently estimated at 10-15 million, and most of them is located in Europe especially in the central region of Europe. Another significant European population can be found in the Iberian Peninsula. Outside Europe, they can be found in the Caucasus region, the Middle East and in the Americas. Their genetic make-up was shaped by at least one founder event, genetic drift, and other events such as differential admixture with other people mainly from Western Eurasia. Roma subgroups were created by the limited gene flow caused by social habits similar to those of the Indians, like the caste system and genetically closed Sub-ethnic groups, also called as clans, thus creating a characteristic substructure in the Roma population. Like Indians, there is also a high incidence of marriage of close relatives within these Roma sub-ethnic groups. Human genetic research targeting the genetics and the genome of Roma has identified a number of rare disease-causing mutations in the Roma population following Mendelian inheritance and are exclusive to Roma. These may have been due to the effect of the special population genetic events and characteristics already described, through which the genetic and genomic study of the Roma population has become an important goal of medical genetics. Earlier, these studies were based on comparative statistical analysis of certain loci and markers, advances in the 2000s also allowed genome-wide medical genetic case-control studies.

Before the emergence of genetic investigations, there have been several studies on their origins based on historical, linguistic, anthropological evidence. Linguistics and anthropology were thought to discover similarities between certain ethnic groups in India and the Roma. According to comparative linguistics, Hindi is most similar to Roma among the languages of India. Anthropological research has pointed to a phenomenon previously mentioned in the health characteristics of the Roma people, that a caste system similar to that of the Indians is present in Roma society, and that endogamy and inbreeding can also be found in several Roma groups.

Objectives

- Describing the contribution of distinct regional populations to Roma ancestry from India to Europe
- Determining the significance of the Caucasus region in the ancestry of Roma
- Characterizing admixture events occurred during the migration of the proto-Roma
- Investigating the effects of the Ottoman invasion of East-Central Europe on Romani people and on European populations

Materials and Methods

Materials

Our study is based on DNA isolated from the EDTA-anticoagulated whole blood samples of the participants. These participants were individuals from Central Europe (from Hungary or from the neighborhood of Hungary) who declared themselves consciously as Roma for at least three generations back. All participants gave their written informed consent to participate in the study. They all got personal verbal information prior their signed consent, which was approved for this study by the Regional Research Ethics Committee of Pécs. The samples were anonymized. This study follows the principles expressed in the Declaration of Helsinki.

Datasets

The Central European Roma samples, included in our study, were partly collected, and genotyped in an international collaboration with the Harvard Medical School. The genotyping was carried out on the Affymetrix Genome-Wide Human SNP 6.0 array chip (n=27; 726,016 SNPs). The other part of the sample set was obtained from the database of the University of Rotterdam (n=152, 868,174 SNPs), which was made available upon request. This dataset was genotyped on the same microarray platform and represented a total of 179 individuals. The genotype calling and quality control were carried out using the recommended settings of the Birdseedv2 algorithm, which can be found in the Affymetrix Power Tools command line software bundle. Raw genotype calls were converted to binary PLINK format using a domestic script and also PLINK1.9, and were filtered further according to the genotype data missingness in all samples. SNPs with missing genotypes were removed from the data using PLINK1.9 missingness per marker feature as inclusion criteria. This means that all markers were removed which failed the genotyping process in at least one sample, meaning that the genotype for the given marker was not found uniformly in all 179 samples.

Genetic distances were also added to the marker data with PLINK using the HapMap Phase 2 GRCh37 genetic map, which allowed to carry out certain linkage-based tests. The resulting dataset contained data from 599,472 SNPs after genotyping and after the follow-up additional quality control.

Based on preliminary population structure studies, Roma individuals strongly admixed with non-Roma Europeans was removed from the data, resulting that 21 of the Roma samples were removed from the data. 158 Central European Roma samples were found suitable for our study.

We also applied several datasets that was either in our repository already, was available freely from public online repositories or made available to us upon request. We also used a dataset of which usage is subject to permission.

We used samples from the publicly available HGDP dataset (n=1044 from 57 populations, 660,918 SNPs genotyped on Illumina 650 Y array). This dataset contains data from various worldwide populations. We used two additional free datasets available from public online repositories, which were datasets of the Estonian Biocentre described in two distinct papers. We will refer to these data as the "Caucasus dataset" (n= 204 from 13 populations, 555,767 SNPs) and the "Jew dataset" (n=466 from 39 populations, 555,736 SNPs) from now on reflecting to the focus of these articles. We also applied the upon request available dataset of Indian ethnic groups (n=121 from 23 ethnic groups, 524,053 SNPs, genotyped on Affymetrix 1 M and Illumina 650K arrays) which was described previously in a study conducted by the Harvard Medical School dealing with the population history of India. The authorized access requiring dataset was the POPRES (n=4077 from 57 populations, 453,617 SNPs genotyped on Affymetrix 500 K platform) from the repository of the NCBI dbGaP. The dataset similarly to the HGDP data contains worldwide population data, but in our study, we applied only European and Indian samples. Hungarian data in our repository, comprising of n=238 individuals and 898,723 SNPs after preliminary filtering analyses, were also used in some of the investigations.

We examined Roma admixture using two approaches, therefore we created two different main datasets for our investigations. The first approach was to create a dataset containing the regional population groups falling on the migration route of the Romani people. These groups were Europeans, populations of the Caucasus region, the Middle East, and the populations of South Asia. Latter comprises of Pakistani and Indian populations. In order to create this dataset, we used the HGDP data, the Caucasus, the Jew, and the Indian datasets. In the second approach, we examined the impact of the Ottoman Empire on the peoples of the Ottoman-occupied countries and on the Roma. In this approach, we examined European, Caucasus and Middle Eastern samples, for which, in addition to the datasets listed above, we also used the POPRES data and our dataset containing Hungarian samples.

Methods

To study the structure and relationship of the populations involved in the investigation, we used the SMARTPCA program from the EIGENSOFT 6.01 software package, which still gets updated algorithms and is developed in the cooperation of the Harvard Medical School and the Broad Institute. Unlike most software we applied in this study, which employ statistical methods, this software is based on a mathematical algorithmic principle. SMARTPCA performs principal component analysis (PCA) on an allele frequency matrix, narrowing multidimensional data to quantifiable and interpretable dimensions (with minimal information loss), which are also relevant to a given research. It also performs formal significance tests based on the computed eigenvalues. With its help, one can observe patterns on genome-wide marker data that reflect the actual relationships of the investigated ethnic groups, and it also provides the F_{st} matrix of investigated populations, which is the pairwise average allele frequency differentiation matrix of ethnic groups, and it reflects the degree of relatedness of the populations to each other.

Ancestry estimation and clustering analysis were performed with ADMIXTURE, which is a clustering software based on a statistical method. The ADMIXTURE software uses an expectation maximization algorithm implementing maximum likelihood estimation in order to estimate the extent of the relationship between the studied populations and given hypothetical ancestors. One can determine the maximum number of possible common ancestors (value of K) with a cross-validation method.

The TreeMix 1.13 algorithm, like ADMIXTURE and several other ancestry estimation algorithms, is based on the working principle of the STRUCTURE program. The software generates a so-called "maximum-likelihood" graph using the autosomal allele frequency data. With its help, we can also infer the relationship of distinct ethnic groups, since it shows population splits and admixture events, but it can also be used to estimate migration events.

For the formal testing of admixture events, we applied algorithms included in the ADMIXTOOLS 4.1 Software Package. These statistics are based on the measurement of allele frequency correlations between investigated populations and referred as f-statistics by its developers at Broad Institute, however the 4-population test is implemented in the package as D-statistics.

The Refined IBD algorithm of Beagle 4.1 was also applied in certain tests. The algorithm seeks for identical by descent (IBD) segments among Roma and the populations of investigated regional groups. We applied the algorithm mainly in order to assess the extent of Roma ancestry from regional populations. An average pairwise IBD sharing between Roma (population I) and regional populations (population J) from the output of Beagle was calculated using the following formula:

$$Average\ pairwise\ IBD\ sharing = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} IBD_{ij}}{n \cdot m}$$

where IBD_{ij} is the length of IBD segment shared between individuals i and j, and n, m are the number of individuals in population I and J.

Since higher number of shared long IBD segments implies more recent admixture between two given populations, we examined the issue, whether there is a detectable difference in the distribution of shared IBD segment length among Roma and distinct regional populations falling on the migration route of Roma. We calculated the distribution of the average number of IBD lengths between pairs of individuals from Romani people and regional populations using the output of Beagle. IBD segments were classified by length, the number of segments were counted in each length classes and were divided with the number of all possible pairs of individuals.

As we attempted with the IBD length analysis, we tried to infer the chronology and the date of the gene flow between Roma and regional populations implementing the ALDER 1.03 algorithm. ALDER is capable to estimate the date of population admixture. Similar to its predecessor, the ROLLOFF algorithm, ALDER is also based on the decay of admixture LD. The algorithm calculates the correlations between SNPs in an admixed target population weighted accordingly to the allele frequency difference in ancestral populations. Ancestral populations (or more precisely contemporary surrogates of ancestral populations) serve as reference populations to the algorithm. The results are highly affected by background LD; therefore, the algorithm uses allele frequencies of the reference populations to amplify the signal of admixture LD, which helps filtering out the background LD. The enhanced algorithm of ALDER provides sophisticated weighted LD statistics, and also has the ability to fully avoid biased estimates caused by the background LD. Another major advantage is that the algorithm is capable to apply the target population itself as a reference resulting in virtually unbiased statistics.

Results

Population structure and ancestry analyses helped to place the Roma on a Eurasian context and showed that the investigated regional populations establish three major groups with variously definite clusters. These were the slightly loose groups of Europeans and South Asians and the tightly clustered group of populations from the Middle East and from the Caucasus. Central Asians and a few groups of the South Asians are showed to be outliers due to their relatively high East Asian ancestry proportion. The analysis results reflected the actual geographical positions of these regional groups. Plotting Roma samples to these populations, we can see that Roma are scattered severely between Europe and South Asia, due to their migration throughout Eurasia and basically because their nomadic nature. However, most Roma individuals plotted more tightly to each other, and can be found between South Asia and the populations of the Caucasus, Middle East, and Central Asia. TreeMix strengthened these results, placing Roma between the Middle Eastern populations, the Caucasus region and between South Asians.

F_{st} calculations showed that Roma have the highest differentiation with South Asian populations and the closer the investigated populations to Europe are, the lower the value of F_{st} gets. These are expected results and reflects the chronological order in which the Roma contacted with the populations of each region during their migration from Northwest India. However, we found a minimum F_{st} at the populations of the Caucasus region and Turkey. It suggests that this region could play an important role in the ancestry of Romani people. Our tests also show that Roma could have a remarkable Turkish ancestry even compared to the neighboring Caucasus region.

Based on the results of the population structure and ancestry estimation tests, we investigated formally the proposed admixture events in order to provide evidence and also to assess their magnitude. Thus, we applied the 4-population and F_4-ratio estimation tests which can prove that the investigated groups are admixed. The formal tests confirmed that the Caucasus, Middle East, and Central Asia are admixed with the ancestors of Roma. F_4-ratio estimation gave the result which suggests that Roma have a high extent of ancestry proportion from these populations. According to Admixture graph fitting results, our model of Roma ancestry fits well into the data, therefore confirming also the proposed significant admixture of Roma and the populations from the Caucasus region. The average pairwise IBD sharing estimation results show that populations of the Caucasus region have slightly higher IBD share with the Roma than Middle Eastern and Central Asian populations, which is similar to the share of South Asian populations. Our IBD length distribution analyses are concordant with data provided by average IBD share, since its results showed that the Caucasus region has a greater number of shared long IBD segments than Middle East and Central Asia. This could also suggest that Caucasus populations have admixed with Roma people more recently and have a higher proportion in Roma ancestry.

Our investigations regarding the impact of the Ottomans on East-Central Europe show that Roma and Turks are admixed. Assessing the impact of the Ottoman occupation, we also investigated the admixture of East-Central European populations with Turks using the 4-population test. This test might reveal Turkish ancestry from the Ottoman occupation of East-Central European territories. We also confirmed with average IBD sharing estimation that this admixture could originate from the former Ottoman rule of East-Central Europe, since Turk-related ancestry of Roma shows a higher degree than the ancestry of populations living adjacent regions of Anatolia. Estimating the average IBD sharing of East-Central Europeans with Turks also showed that East-Central European populations have higher average IBD share with Turkic people than other formerly Ottoman-occupied regions (Caucasus, Middle East). In order to assess the significance of average IBD sharing difference values between investigated formerly Ottoman-occupied populations, we also calculated the average IBD sharing of Turks with Sardinians, which group lives separately from the European continent on the Sardinia Island, therefore isolated

largely from European demographic events. This investigation strengthened that the Ottoman occupation of East-Central Europe has a detectable impact on its population.

In order to estimate the dates for the admixture of the ancestors of Roma with the populations of the Caucasus region, Middle East, Central and South Asia, we conducted analyses based on admixture LD, implementing ALDER. Results show that Roma admixture with South Asians is the oldest admixture event occurred during the migration of Roma, as expected. ALDER placed the admixture event of the ancestry of Roma with Caucasus and the Middle East to a similar date, but admixture with the groups of the Caucasus is more recent.

We also estimated the date of the admixtures in order to further investigate and provide evidence for the proposed admixture of Roma and East-Central Europeans with Turks during the Ottoman occupation. In case of the Romani people, the resulting admixture date was biased due to the supposed multiple admixture events with Middle Eastern populations, and it placed the date of the admixture somewhat earlier. The admixture date interval obtained from East-Central Europeans corresponds to the time interval when Ottomans were present in East-Central Europe. ALDER strengthened our proposal that a part of the Turkish ancestry in East-Central Europeans can originate from the 150 years Ottoman presence in Europe, and ancestry from Turkic people in Roma could be derived also from the times of the Ottoman occupied Europe.

Discussion

Applying genome-wide autosomal marker data, we were able to assess the contribution of the Caucasus to the ancestry of Roma, which seems to be significant also compared to the two main sources of their ancestry (South Asia and Europe). Our analyses show that the Caucasus region can be the third most important source of Roma ancestry taking into consideration their migration route, which included also the Central Asian and the Middle Eastern regions. Our results suggest that the area of the Caspian and Black Seas is a significant source for the genetic legacy of Romani people, this region plays and important role in their ancestry and migration history.

We also confirmed that the expansion of the Ottoman Empire into East-Central Europe left its mark on local populations, contributing significantly to their Turkic ancestry. According to investigations based on Y chromosome markers, the R1b haplogroup is a primary characteristic of West Europeans and significance of this haplogroup strongly decreases towards East Europe, where R1a is predominant. Most of the East-Central European ethnic groups share R1a and R1b to approximately the same extent. Our genome-wide autosomal marker data provided a similar geographical pattern in the genetic makeup of European populations, as these studies based on Y haplogroups found. Ancestry analysis, testing for

admixture and IBD segment analyses further separated the formerly Ottoman-occupied East-Central European groups from the rest of Europe, showing that ancestry derived from the Middle Eastern area is also observable in the autosomal data. According to the literature, the Middle Eastern region derived Y haplogroups E3b and J were observed at high extent in East-Central Europeans. Our results correspond to this observation, since we found that Middle Eastern ancestry also shows the highest extent in OEC groups in case of genome-wide autosomal marker data. We revealed that Romani people in the East-Central European area could have acquired Turkish ancestry not only during their migration towards Europe, but also at the time of Ottoman presence in Europe; at a time when Roma were already the largest ethnic minority of the region.

Summary

- Applying IBD analysis, we successfully estimated the share of regional populations like people of South Asia, the Middle East (and Central Asia), the Caucasus region and Europe in the ancestry of Roma, therefore we estimated the significance of these region in the Roma migration history.
- With the help of f-statistics, we confirmed that Caucasus related ancestry in Roma are detectable, therefore significant, and also characterized it with various statistical approaches.
- Using f-statistics and LD-based methods, we further confirmed and characterized the admixture events involving regional population on the migration route of Roma, and therefore also strengthened the concept of their main migration route itself with population genetic approaches.
- Also applying population stratification, ancestry estimation, formal admixture analyses, IBD and LD-based analyses, we found that the effect of the Ottoman occupation on East-Central Europe and on Roma left a detectable genetic impact, therefore showing that the role of the Middle East in the Roma migration has an even smaller role than our previous estimations pointed it out, highlighting the significance of the neighboring Caucasus region in Roma ancestry and Roma migration.

Papers on which the thesis is based

1. **Revealing the impact of the Caucasus region on the genetic legacy of Romani people from genome-wide data**
   Z Bánfai, V Ádám, E Pöstyéni, G Büki, M Czakó, A Miseta, B Melegh.
   PLoS One. 2018 Sep 10;13(9):e0202890. doi: 10.1371/journal.pone.0202890

   Impact factor: 2.776

2. **Revealing the genetic impact of the Ottoman occupation on ethnic groups of East-Central Europe and on the Roma population of the area**
   Z Bánfai, BI Melegh, K Sumegi, K Hadzsiev, A Miseta, M Kásler, B Melegh
   Front Genet. 2019 Jun 13;10:558. doi: 10.3389/fgene.2019.00558

   Impact factor: 3.517

Other publications (10)

Impact factor: ∑ 55.57

Cumulative impact factor: 61.863

Acknowledgements

I am primarily grateful to my supervisor Professor Béla Melegh, who aroused my interest in the field of human genetics and human population genetics based on genome-wide data and state of the art algorithms and enabled me to join the Doctoral School of the Medical School of the University of Pécs within the framework of the Interdisciplinary Medical Science. He monitored my professional activity throughout, guided and assisted my research work.

Special thanks to the colleagues and PhD students of the Department of Medical Genetics, especially to Katalin Sümegi, András Szabó, Gergely Büki, Anita Maász.

I am grateful to all the assistants of the Department of Medical Genetics who helped me with their competent, conscientious work and professional experience.

Finally, I owe a debt of gratitude to my family who made this work possible with their endless and understanding patience and support.