From networks to population-specific adaptations: studying positive selection in humans

Begoña Dobón Berenguer

TESI DOCTORAL UPF / 2018

DIRECTORS DE LA TESI

Dr. Jaume Bertranpetit i Busquets

Dr. Hafid Laayouni El Alaoui

DEPARTAMENT DE CIÈNCIES EXPERIMENTALS I DE LA SALUT

Universitat Pompeu Fabra Barcelona

Dedicatòria

A mon pare. Ha sigut dur acabar-ho sense tu.

Acknowledgments

A la meua família, que sempre m'ha animat a estudiar el que més m'interessés, encara que això suposés viure lluny d'ells i perdrem la mitat de les reunions familiars. Gràcies per el vostre suport, sé que no sempre teniu molt clar el que faig, espere que esta tesi ho aclareixi un poc. Silvia, lo mejor del máster de Bioinformática fue conocerte a ti. Con gente como tú a mi lado, sé que no importa que pase en esta vida, todo se puede superar. Gracias por enseñarme que somos mucho más fuertes de lo que creemos.

All the members of the Evolutionary Systems Biology lab, past and present, thanks for being there to answer my (stupid) questions. I hope I was able to answer some of yours in return. To me, the people you work with is, if not more, as important as the work you do. To Pierre, Marc, Ludovica and Mayukh, I learnt a lot from all of you, I admit that I was very impressed by how much you knew about everything (still am) and I hoped to be as knowledgeable as you by the end of this thesis (I am not). To Sandra, Jessica, Pablo, Gabriela and Guille, I'm so glad you came to the group and we spend these years together (and go to the beer sessions, dinners, BBK, … it's been a lot of fun!).

To my supervisor, Jaume, who had to put his foot down and tell me (more than once) to stop doing analysis, to write the paper and send it once and for all. If not for him I would have never finished. Every time I got lost on the research, you would help me see the big picture again. To my other supervisor, Hafid, who always encouraged me to "complain" and speak up: about them not having time, not giving feedback, … I have been lucky to have supervisors who really care about their students and see that they growth as researchers, and as persons.

This research has been supported by F.P.U. grant FPU13/06813 from the Ministerio de Educación, Cultura y Deporte (Spain).

Abstract

The evolution of the genome is driven, among other factors, by the composition of the genome, the functional output of the gene products, and by environmental pressures. Among the environmental factors, pathogens are one of the strongest selective pressures. In this thesis we describe two examples of this: i) the convergent evolution in immune-related genes in East Africa populations despite having different ethnical and genetic backgrounds, ii) the rapid adaptation in variants associated with differential cytokine production in the Roma people since their migration from the Indian subcontinent. We also propose that positive selection acted in the cytochrome P450 system after the out of Africa, whereas genetic drift is the main force behind the genetic variability present in taste receptors genes. Gene evolution is also affected by the location and connectivity of their products within the metabolic network. Positive selection detected at interspecific and intraspecific levels show opposite but complementary patterns: the first is detected in peripheric genes, whereas the second is detected mainly in central genes.

Abstract

Resum

L'evolució del genoma depèn, entre altres factors, de la composició del genoma, del paper funcional que realitzen els productes del gen, i de les pressions mediambientals. Dins dels factors mediambientals, els patògens són una de les pressions selectives més fortes. En aquesta tesi descrivim dos exemples d'aquest fet: i) l'evolució convergent en gens amb funcions immunològiques a poblacions de l'est d'Àfrica, malgrat pertànyer a diferent grups ètnics i fons genètics, ii) la ràpida adaptació als Roma en variants associades amb la producció diferencial de citocines des de la seva migració des del subcontinent Indi. També proposem que la selecció positiva va actuar a la súper família dels citocroms després de la expansió des del continent Africà, mentre que la deriva gènica és la principal força darrere de la variabilitat genètica observada als receptors del gust. L'evolució dels gens també és veu influïda per la ubicació i connectivitat dels seus productes dins de la xarxa metabòlica. La detecció de la selecció positiva a nivell interespecífic i intraespecífic mostra un patró oposat però complementari: la primera és detectada a la perifèria de la xarxa, mentre que la segona és detectada principalment en gens centrals.

Preface

Not all human populations have been equally represented in genetic studies. One of the efforts of this thesis has been in including traditionally neglected populations to provide a wider representation of the existent human genetic variability. This generated new insights into the evolutionary history of two groups: African populations, specifically from East Africa (Sudan, South Sudan and Ethiopia), and an ethnic minority, the Roma people currently living in Romania. Besides, both studies supported the notion that pathogens and infectious diseases have been one of the strongest selective pressures during human evolution.

Then, we followed a network approach to address one of the main problems in biology: the gap between genotype and phenotype, and how to interpret the findings from genome-wide scans of selection. This thesis presents the first comparison of the relationship between both, intra and interspecific variation, and the topological structure of the human metabolic network. The adoption of a network approach to detect and interpret signals of positive selection corroborated the idea that adaptive selection acts in distinct parts of the network depending on the evolutionary time-scale considered.

Contents

List of figures

List of tables

INTRODUCTION

1. Origin of humans

Modern humans originated in the African continent around 200 thousand years ago (KYA). However, whether the *Homo sapiens* evolved from a single population or if there was ancient substructure early on our history is still debated (Scerri et al., 2018). Around 60 KYA a group of modern humans migrated out of Africa (OOA) and spread across the world (Henn, Cavalli-Sforza, & Feldman, 2012; Tishkoff et al., 2009) (Figure 1).

During our evolution, humans have faced extreme environments, new pathogens, novel food sources, and encounters with other hominins. To survive and thrive, our species had to adapt. OOA populations interbred with archaic populations, resulting in between 1% to 6% of modern genomes with Neandertal or Denisovan origin, that, in some conditions, has provided a selective advantage in humans (Abi-Rached et al., 2011; Green et al., 2010; Huerta-Sánchez et al., 2013; Racimo, Sankararaman, Nielsen, & Huerta-Sánchez, 2015; Reich et al., 2010). Around 10 KYA, the appearance of agriculture in several regions of world started the Mesolithic-Neolithic transition. Human populations, up until then, nomadic hunter-gatherers, changed into sedentary and semi-nomadic communities of agriculturists and pastoralists. This was the seed for a dramatic population expansion that within the past 4000 years, caused over 100 migration and admixture events between human populations (Hellenthal et al., 2014).

This complex pattern of migrations and admixture seems a commonplace occurrence during the evolution of the Homo genus (Ackermann, Mackay, & Arnold, 2016). Only this year, the offspring of a Neanderthal mother and a Denisovan father was reported (Slon et al., 2018). If such an individual existed, how many are still to be discovered? How many others have existed and disappeared without leaving fossil remains? Without doubt the discoveries made by the latest genetic studies are challenging our definition of species and our ideas of how humans evolved.

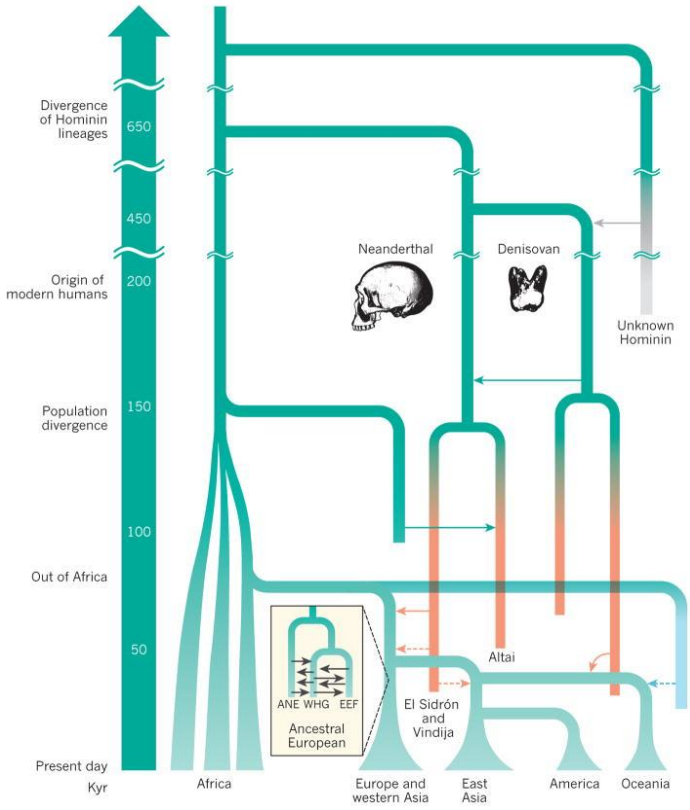


Figure 1. Simplified model of human evolutionary history.

Origin of present-day population. Arrows indicate known and theorized admixture events between modern humans and between archaic and modern humans. (Nielsen et al., 2017)

1.2. Underrepresented human populations in genetic studies

In the last 10 years, the cost of SNP genotyping arrays and genome sequencing has dropped considerably, allowing the genotyping and sequencing of hundreds of thousands of individuals from populations across the globe. This has been possible by projects led by international consortiums that, besides generating a wealth of genetic data, developed most the tools needed for its analysis

(Abecasis et al., 2012; Auton et al., 2015; †The International HapMap Consortium et al., 2003; T. I. H. Consortium et al., 2007).

Commercial SNP arrays contain between 200,000 and 2,000,000 SNP, usually with uniform coverage of the human genome - but see the Immunochip (Cortes & Brown, 2011; Trynka et al., 2011) for a special case. The main issue with genotyping arrays is SNP ascertainment bias, the SNPs to be genotyped must be known beforehand. Typically, genotyping arrays have been biased towards alleles discovered in European populations with a minor allele frequency (MAF) in the general population higher than 1% (Abecasis et al., 2012; Lachance & Tishkoff, 2013b).

This in turn, has impacted the last decade of genome-wide association studies (GWAS) performed using these arrays and the imputation panels generated with representative populations sequenced at low coverage (Abecasis et al., 2012; Delaneau & Marchini, 2014; Visscher et al., 2017). GWAS studies have proved helpful in studying the etiology of common adult diseases (Visscher et al., 2017). However, due to the OOA non-African populations carry a subset of human genetic variation and it is not always possible to extrapolate the insights obtained from genetic studies outside the population that generated them. The replicability of GWAS results is high in populations of European and Asian ancestry (85.6% and 45.8%), but much lower for populations of African ancestry (9.6%) (Marigorta & Navarro, 2013).

Thus, we need to increase the representation of neglected human populations to discover the true levels of genetic variability and how the genotype influences the phenotype. The scientific community is well aware of this, and it is trying to remedy it by generating next generation sequencing (NGS) data from the main geographic areas (Europe, Middle-East, North Africa, Sub-Saharan Africa, Central-South Asia, East Asia, Oceania and America). The latest phase of the 1000 Genomes Project includes a wide representation of African, and East and South East Asian populations (Auton et al., 2015). Specially interesting initiatives are the African Genome Variation Project (AGVP) (Gurdasani et al., 2014), and open access platforms such as AfricArxiv (www.africaxiv.org), which are helping the inclusion of African

populations and, as importantly, African researchers and universities, in genetic studies.

In this thesis, the two first chapters represent my small contribution to include underrepresented populations in genetic studies: East African populations (Results: 1. The genetics of East African populations: A Nilo-Saharan component in the African genetic landscape) and a European minority, the Roma (Results: 2. The shaping of immunological response through natural selection after migration: the case of the Roma).

1.2.1. East African populations

East Africa is a strategic region to study human genetic diversity due to the presence of ethnically, linguistically, and geographically diverse populations (Tishkoff et al., 2009). One of the main factors contributing to this diversity is the Nile River, which has acted as a genetic corridor allowing gene flow between North and Sub-Saharan Africa.

The first result of the thesis (Results: 1. The genetics of East African populations: A Nilo-Saharan component in the African genetic landscape) consists of a genetic study of East African populations from the Sudanese region, that comprises Sudan, South Sudan and Ethiopia. One issue to have in mind when working in genetic studies, is that national frontiers do not always correspond to population boundaries. In this work, we considered nine populations or ethnic-linguistic groups using the following definition: a group of people that share common culture, language, origin, practices, and live in the same geographic region.

We described a genetic component that identifies Sudanese Nilo-Saharan speaking groups (Darfurians and part of Nuba populations) and South Sudan Nilotes. The genetic homogeneity of these populations contrasted with the populations from the north and eastern parts of the region (Nubians, Arabs, Beja, and Ethiopians), which showed genetic admixture of the Nilo-Saharan component with a European genetic component. A broader study confirmed

these conclusions and estimated the admixture event to ~ 700 years ago (Hollfelder et al., 2017).

Taking advantage of the genotyping array used, the Immunochip (Illumina Infinium single-nucleotide polymorphism microarray) (Cortes & Brown, 2011; Trynka et al., 2011), we also analyzed how infectious pressures affected the genetic variation of East African populations. We found that selective pressures on host defense genes generated lower genetic distances between populations in those genes, hinting at convergent evolution of the immune system.

1.2.2. The Roma people

The Roma people, with a population of 10-12 million, are the largest ethnic minority in Europe. Previous genetic studies indicate that the initial proto-roma population departed from the northwestern region of the India subcontinent around 1000-1500 years ago (Mendizabal, Lao, & UM, 2012; Mendizabal, Valente, & Gusmao, 2011). After crossing Persia and Armenia, they settled for two hundred years in the Balkan peninsula between the 11^{th} and 12^{th} centuries (Achim, 1998; A. Fraser, 1992). Some groups kept migrating west, and by the 15^{th} century their presence is mentioned in the Iberian peninsula (A. Fraser, 1992). Remarkably, the Roma diaspora can be considered the last human migration of Asian origin into Europe (Achim, 1998).

The history of the Roma migration indicates that any selection study in this group will have to account for the confounding effects of small population size, strong genetic drift (multiple bottlenecks), isolation, and uneven migration patterns with the surrounding host populations.

The second result of this thesis (Results: 2. The shaping of immunological response through natural selection after migration: the case of the Roma), consists of the analyses of whole-genome sequences of Roma individuals along with individuals from their host population, Romania. The main goal of the study was to identify the selective pressures the Roma faced before and after their migration. To understand how the immune system of the

Roma adapted to different pathogens, we performed a genome-wide scan of selection and combined the results with functional studies of an immunological quantitative trait loci (QTL): cytokine production (Figure 2). This integrative approach can help distinguishing between the real targets of positive selection and false positives (Barreiro & Quintana-Murci, 2010).

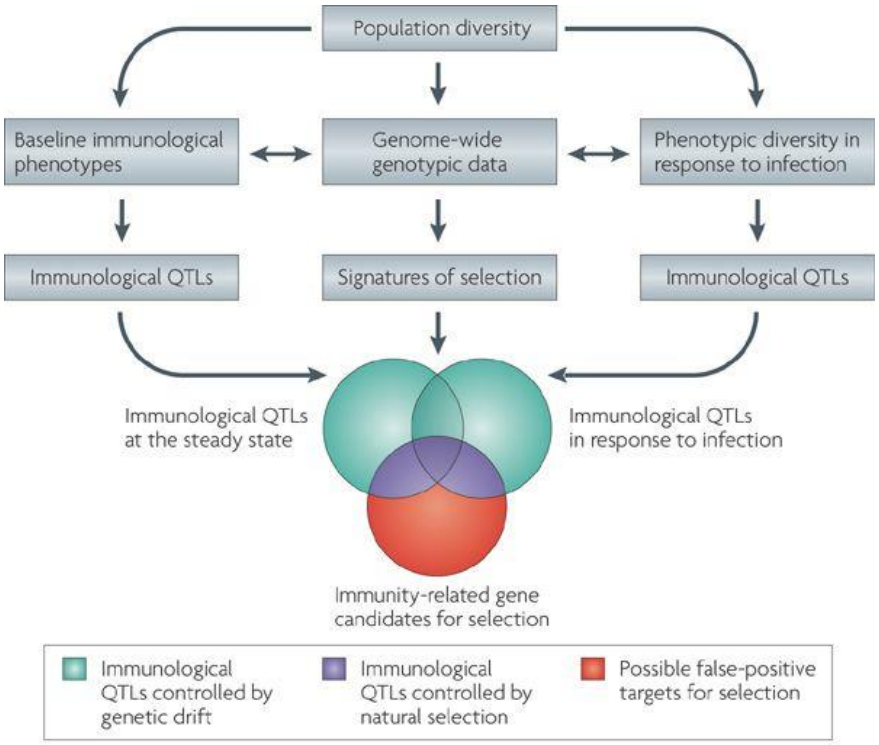


Figure 2. Identification of adaptive immunological phenotypes.

The combination of scans of selection with functional studies to detect immunological quantitative traits loci (QTLs) can identify the real targets of positive selection (Barreiro & Quintana-Murci, 2010).

2. Detecting signals of positive selection

In 1858, Darwin and Wallace, introduced the idea of natural selection to explain how species evolve. The main concept of natural selection is that a heritable trait that increases the individual's fitness and its ability to survive and reproduce in a given environment, will eventually increase in frequency in the population.

The study of natural selection can help us clarify, among many other questions: how our species evolved, what is the origin of the phenotypic variation that we observe, and why some diseases are more prevalent in specific populations.

All the phenotypic and genetic variability present in humans is the substrate in which natural selection acts on. Natural selection encompasses different modes of selection: positive, purifying, balancing, and sexual selection.

Positive selection: also called adaptive selection, consists on the increase in frequency in the population of a beneficial mutation, usually until it reaches fixation. It is the process that allows the creation of new phenotypes and drives the adaptation to new environments. Its detection and how it is affected by other factors, such as network topology, is the main topic of this thesis.

Purifying selection: also called negative selection, consists on the removal of deleterious mutations from the population. It is considered to be the main force acting on the functional elements of the genome, as a mutation with a functional consequence is more likely to have a damaging than a beneficial effect.

Balancing selection: this process maintains different alleles in the population at medium-high frequency. In this case, neither of the alleles will reach fixation, as it is the presence of genetic diversity the advantageous factor.

Sexual selection: it is a special case of natural selection. It acts when there no random mating in the population, but mating choice favors a given phenotype.

There are several methods to detect positive selection by comparing either the genetic variation between species (interspecific variation or divergence data) or the genetic variation between individuals of the same species (intraspecific variation or polymorphism data) based on the footprints left in the genome (Figure 3).

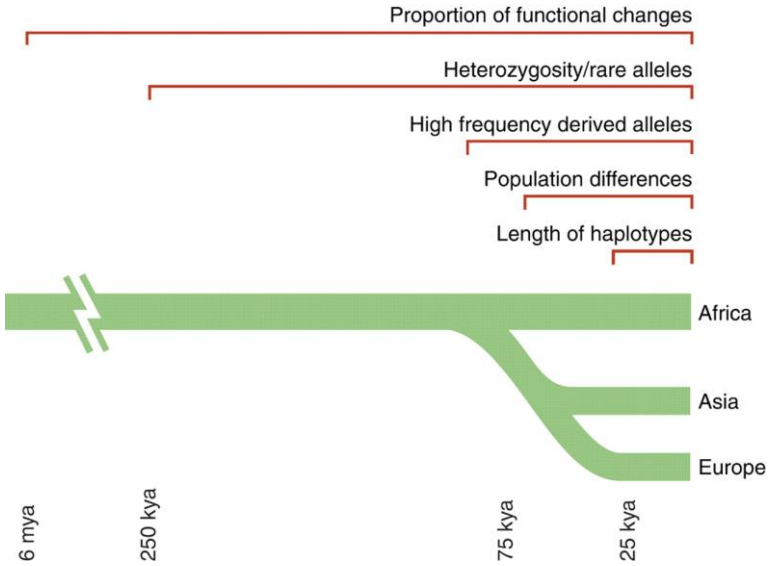


Figure 3. Time scales for the signatures of selection.

Positive selection creates different signatures in the genome that are erased at different time-frames. From (Sabeti et al., 2006).

2.1. Using divergence data (interspecific data)

One way to detect positive selection on protein-coding genes is to compare orthologous sequences and identify substitutions in the protein sequence (Figure 4). A synonymous mutation does not change the amino acid sequence and its effect can be assumed to be neutral. Thus, the number of synonymous substitutions per synonymous site (dS), can be considered the background level of

mutations in a protein before selection takes place. A nonsynonymous mutation will change the amino acid sequence and it is expected to affect the protein function. The number of nonsynonymous substitutions per nonsynonymous site (dN) will reflect the selective pressure on the protein. The ratio between these two rates will estimate the selective pressure acting on the protein ($\omega = dS/dN$). If nonsynonymous mutations are deleterious, they will be removed by purifying selection ($\omega < 1$); whereas is they are advantageous they will be fixed by positive selection ($\omega > 1$). If $\omega = 1$ the protein evolves neutrally.

It is expected than most amino acid sites will be under strong purifying selection, and only a few sites will be under positive selection. Thus, obtaining a global value of ω higher than one for a gene is very unlikely. To formally test for positive selection, we use a likelihood ratio test (LRT) to compare nested models and assess whether adding sites under positive selection explains better the data. The data consists of a multiple sequence alignment of a protein-coding gene, where the sequences belong to species neither too similar nor too divergent. The models are: a null model, where all sites in the multiple sequence alignment evolve neutrally ($\omega = 1$) or under purifying selection ($\omega < 1$); and the alternative model, where an extra category is added to the previous model and some sites are under positive selection ($\omega > 1$).

Through this thesis, six models, all implemented in the codeml package of PAML 4 (Yang, 2007), have been applied to study the evolution of protein coding genes.

M0 (one ratio): It is the simplest model, as it assumes the same ω for all branches or sequences in the multiple sequence alignment. M0 averages the ω along all sites to estimate a global value for the multiple sequence alignment.

M1a (NearlyNeutral) vs. M2a (Positive selection): The M1a model fits the data to two classes of sites: sites evolving neutrally ($\omega = 1$) or under purifying selection ($\omega = 0$). The exact value of $\omega < 1$ is estimated from the data. The alternative model (M2a) adds a third site class ($\omega > 1$), where some sites can be targeted by positive selection.

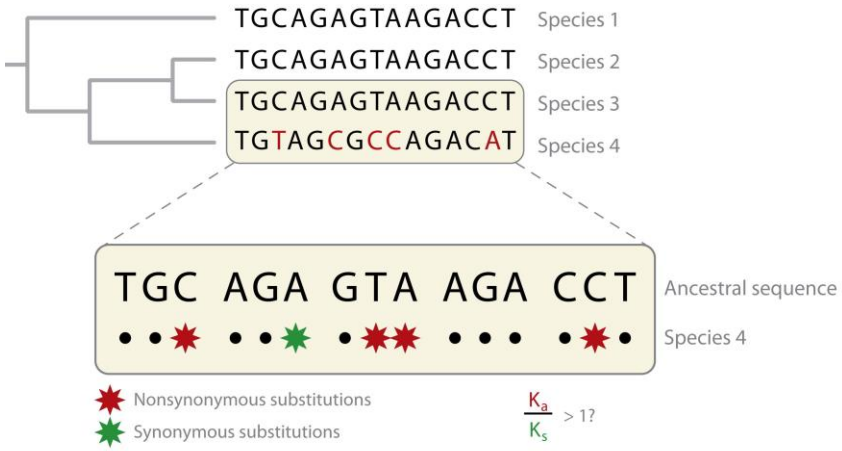


Figure 4. Detecting positive selection using divergence data.

The comparison of the rate of nonsynonymous substitutions (dN or K_a) to the rate of synonymous substitution (dS or K_s) is an indication of the selective pressure acting on a protein coding sequence. Modified from (Vitti, Grossman, & Sabeti, 2013).

M7 (beta) vs. M8 (beta & ω): In the model M7 (beta), codons in the sequence alignment are fit to seven classes: one with $\omega = 1$ and the rest with $\omega < 1$ drawn from a β distribution. The model M8 (β&ω) adds another site class allowed to reach ω values higher than 1.

Branch-site test of positive selection (Test 2): While the other models allow ω to vary between sites, this model also allows ω to vary between branches. It detects positive selection on a given branch. In the null model, sites are fit to two ω values on both the branch of interest (foreground) and the others (background): $\omega < 1$ and $\omega = 1$. In the alternative model, a third ω value is added in the foreground branch: $\omega > 1$. Both, M2a and the branch site test 2 are more conservative than the M8 model.

2.2. Using polymorphism data (intraspecific data)

A selective event leaves a genomic signature that can be identified by comparing sequence or genotype data from different individuals of the same species. In this thesis, the focus was on changes caused

by single nucleotide polymorphisms (SNPs), leaving out the effect of other factors that contribute to the phenotype: copy number variants (CNV), epigenetic changes, alternative splicing, secondary modification of proteins, and so on.

When a new mutation appears that is highly beneficial, it can increase so rapidly in frequency in the population that recombination has not enough time to act and all the linked variants also increase in frequency, creating a valley of decreased genetic diversity around the selected allele (Figure 5a). This process of beneficial mutation allele "sweeping" its surrounding genomic area with it, is called genetic hitchhiking o selective sweep (Smith & Haigh, 1974).

Site frequency spectrum-based tests: The increase of the haplotype carrying the beneficial allele will create a region with low diversity, an increase of rare and derived alleles, and a decrease of alleles with intermediate frequencies (Figure 5b). For example, Tajima's D (Tajima, 1989) detects regions with an excess of rare alleles, but a population expansion after a recent bottleneck will cause the same signature.

Haplotype-based tests: The rapid increase of frequency of the beneficial allele creates a long region of linkage disequilibrium (LD) around it (Figure 5c). For example, the Cross-Population Extended Haplotype Homozygosity (XP-EHH) test compares the length of haplotypes between two populations around a putatively selected variant (Sabeti et al., 2007).

Population differentiation-based tests: Populations under different selective pressures will have different alleles selected, thus differences in allele frequencies can pinpoint local adaptation (Figure 5d). For example, F_{ST} (Weir & Hill, 2002) measures the proportion of genetic diversity due to allele frequency differences among populations. While F_{ST} is non-directional, deltaDAF measures the differences of derived allele frequency between one population and a reference, and by using the derived allele information it can point which population is under selection.

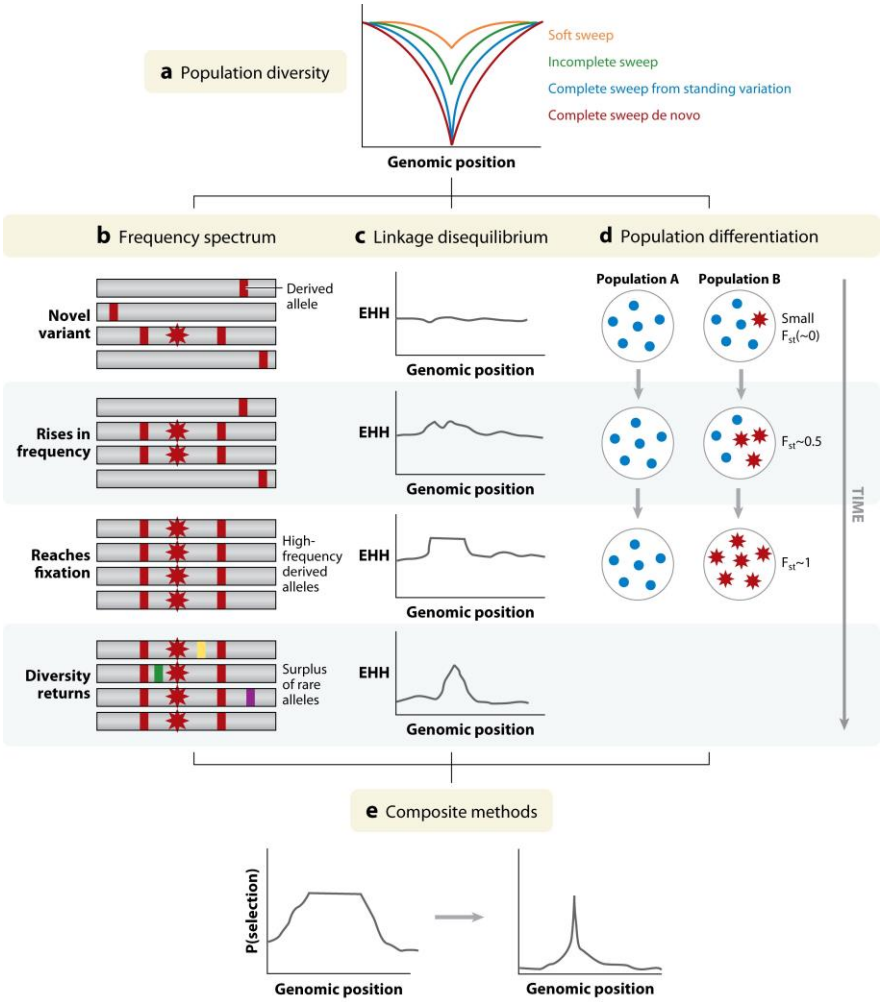


Figure 5. Detecting selective sweeps using polymorphism data.

a) Valley of low genetic diversity around selected allele. Effects of a selective sweep in b) the site frequency spectrum, c) linkage disequilibrium, and d) allele frequencies in the selected region. e) Composite methods can combine information from multiple signatures and enhance the detection of a selective sweep. From (Vitti et al., 2013).

Composite methods: However, tests applied individually are not enough support for proving positive selection, as only the combination of different tests may permit to distinguish the effects of demography (population expansion, population structure, isolation, admixture, bottlenecks and founder effect) from the

selection events (Zeng, Fu, Shi, & Wu, 2006) (Figure 5e). For examples, the Cross-population Composite Likelihood Ratio (XP-CLR) test combines information of allele frequency differentiation (F_{ST}) in an extended genomic region with LD information (Chen, Patterson, & Reich, 2010). More complex methods, such as the Hierarchical Boosting (HB), use a machine-learning approach to classify different selection scenarios, while considering population-specific demography (Pybus et al., 2015).

2.2.1. Confounding factors

There are several factors that affect the detection of positive selection using polymorphism data, some of them are caused by the genotyping and sequencing technics, while others are inherent to how populations originate and change with time.

Ascertainment bias: This is a major consequence of using genotype data, in which the SNPs to genotype have to be selected *a priori*. Effectively, this caused genetic studies to favor the analysis of common variants of European origin (Lachance & Tishkoff, 2013b). This factor was particularly relevant for the first study presented (Results: 1. The genetics of East African populations: A Nilo-Saharan component in the African genetic landscape) as the array used, the Immunochip (Trynka et al., 2011), presents an uneven coverage of the genome and thus, restricted the type of selection tests that could be applied. This bias is partially solved with the use of NGS, but NGS technologies carry their own set of biases, for example, genotype calling algorithms tend to underestimate the true number of heterozygous sites and coverage depth affects the discovery of rare variants (Abecasis et al., 2012).

Background selection: in a comparable way than variants linked to a beneficial selected allele increase in frequency with it, neutral variants linked to a deleterious variant will we removed from the population along with the damaging allele, creating a similar region of low variability.

Population structure: usually we assume that all individuals within a population have the same probability to reproduce.

However, there are geographic, linguistic, social, and religious barriers that separate groups of the population and prevent the random mating of individuals. If there is hidden substructure in the population it will generate an excess of variants at intermediate frequency and higher genetic variability than expected.

Migration: when individuals move from one population to other it will increase the genetic variability within the population and, at the same time, decrease the genetic differentiation between the populations.

Isolation: when a population remains separated from the others, without external genetic flow, it will increase the genetic differentiation with other populations.

Population expansion: sudden increase of the population size. It will generate an excess of rare variants and a decrease of the overall variability of the population.

Population bottleneck: sudden decrease of the population size, usually followed by the recovery or surpass of their original numbers. As with a sudden population expansion, it will generate an excess of rare variants and a decrease of the overall variability of the population, in this case, due to the removal of part of the variation.

Founder effect: a special case of bottleneck, where a subset of the population splits and migrates to a different location.

For these reasons, it is extremely important to have a clear sampling strategy to maximize the discovery of genetic diversity, to study of the population demographic history and ancestry and to keep in mind all the possible biases caused by merging datasets originated in different studies with different technologies. This is true for any genetic study, but especially if we intent to perform a selection analysis.

2.2.1. Examples of positive selection in humans

There are few examples of positive selection in humans where there is strong and irrefutable evidence of both, the genotype and the phenotype under selection (Figure 6). Examples of recent human adaptation include immune-related genes, lactose tolerance, response to hypoxia, and polygenic traits such as height (Fan, Hansen, Lo, & Tishkoff, 2016; Lachance & Tishkoff, 2013a). However, in most cases we are left with either, a phenotype, clearly under selection, but with no knowledge of the molecular mechanism, or with a genomic signature in a gene with unknown function.

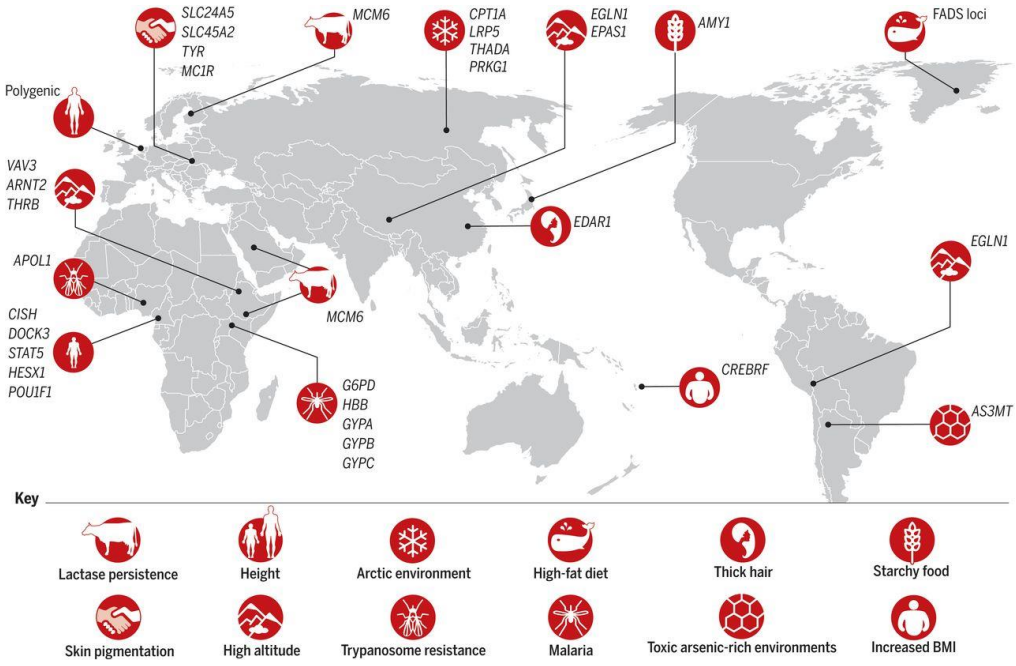


Figure 6. Examples of population-specific adaptations in humans.

For each selected trait or selective pressure is indicated the loci under selection. (Fan et al., 2016)

3. Network framework applied to evolutionary studies

One of the challenges of genome-wide scans of selection is the interpretation of the signals. Most of the signals of selection fall outside coding regions, and of those that are in genes, they are likely to participate in several basic cellular processes or their function have not been described. That, and the evidence from GWAS that for any complex trait many genes contribute to the genetic variation in the population (Visscher et al., 2017) has prompted the formulation of the omnigenic model (Boyle, Li, & Pritchard, 2017). This model suggests that the strong interconnection between gene regulatory networks will cause that any gene co-expressed in a disease-relevant tissue will be related to the disease, even if it does not participate in the key pathway associated to the disease. Evolutionary studies about gene co-expression and protein evolution support this idea: genes encoding interacting or co-expressed proteins co-evolve together (Clark, Alani, & Aquadro, 2012; Lovell & Robertson, 2010).

For this reason, one of the most common practices is to perform some sort of enrichment analysis on the signals. The goal is to test whether we found in our signals more genes associated to a given biological process, pathway, phenotype, or disease than we expect by chance. Following the idea of studying patterns of positive selection in genes with a common function, a study about the evolution of genes related to taste and phase I biotransformation in humans is presented in the third chapter (See Results: 3. Is there adaptation in the human genome for taste perception and phase I biotransformation). We propose that genetic drift and not adaptive selection caused the genetic variability observed in taste receptors genes. Conversely, we report important genetic adaptations in the cytochrome P450 system.

3.1. Metabolic pathways as networks

The human genome can be regarded as a group of interconnected elements, or in other words, as a network. In biological networks,

the elements or nodes, can be genes, or the proteins encoded by them. The connections, edges or links joining the nodes, can be protein-protein interactions, shared metabolites, and so on. Network theory can help us understand the evolutionary constraints imposed by the intrinsic structure of the system.

The last chapter of this thesis presents the first evolutionary analysis of the human metabolic network at both intraspecific and interspecific level (Results: 4. Influence of network topology on the evolution of metabolic enzymes in humans and mammals). In this work we compared how the structure of the network and the connection between the nodes, affects their evolution. To do that we analyzed two different databases: i) the latest consensus metabolic network reconstruction, Recon3D (Brunk et al., 2018), and ii) the metabolic pathways from HumanCyc, a Pathway/Genome database (Romero et al., 2005).

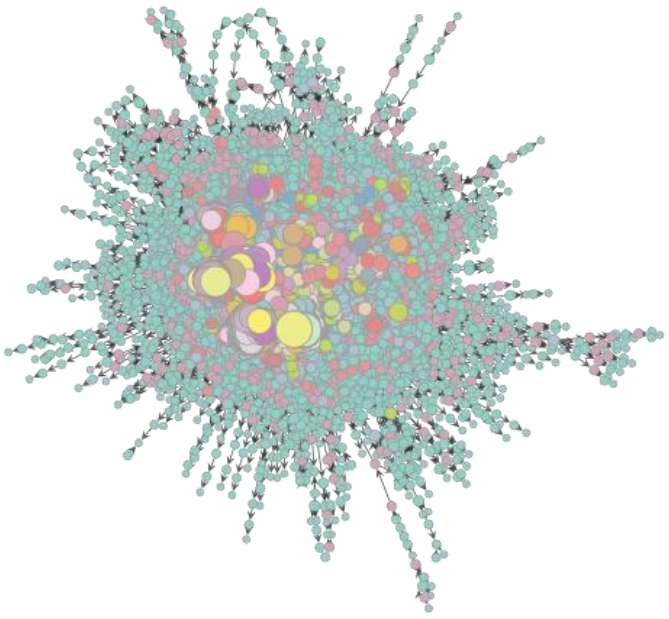


Figure 7. Giant connected component of the human metabolic network.

Nodes size and color correlates with node out-degree.

The choice of these two sources was motivated by the problem of how to define a metabolic pathway and its boundaries. Where does it start and where does it end? It is naïve to treat pathways as separate entities when we are precisely interested in the connections between elements. But at the same time, it is impossible (yet) to include all interactions and obtain a meaningful and interpretable picture from a "hairball" (Figure 7). This poses the tricky question of, what is it better: an overly simplistic model or an overly complex model?

In the end, both are approximations of the reality that we want to study and emphasize distinct aspects of it. In a large-scale network we will be able to infer global patterns and account for cross-talk effects between biological processes, with the drawback that the interactions are less reliable and might be incomplete. However, comparing hundreds of small-scale networks might allow us to uncover local shared patterns with an easier biological interpretation. Montanucci et al. (2018) followed this last approach. The analysis of hundreds of human metabolic pathways showed that purifying selection is stronger in enzymes performing the first reaction of a pathway and that have many connections. It also allowed the comparison of the strength of purifying selection across different the layers of metabolic functions: genes participating in inner core pathways are more constrained.

3.3. How to characterize a network?

In a reaction graph, nodes are enzymatic reactions, and by extension the genes that encode them. Edges are shared metabolites, and directed links indicate which enzymatic reaction produces the substrates of a given reaction. When we represent a metabolic pathway as a reaction graph, the transformation can create several connected components that are isolated (Figure 8).

Number of connected components: A connected component of a network is a subset of the graph where each par the nodes is connected by a continuous path. Figure 8 shows a network with three connected components. When we transformed the metabolic

network to a directed reaction graph (Results: 4. Influence of network topology on the evolution of metabolic enzymes in humans and mammals), we obtained: a giant connected component (88.72% of the nodes), 145 small connected components (1.37% of the nodes), and 821 isolated nodes (7.75% of the nodes).

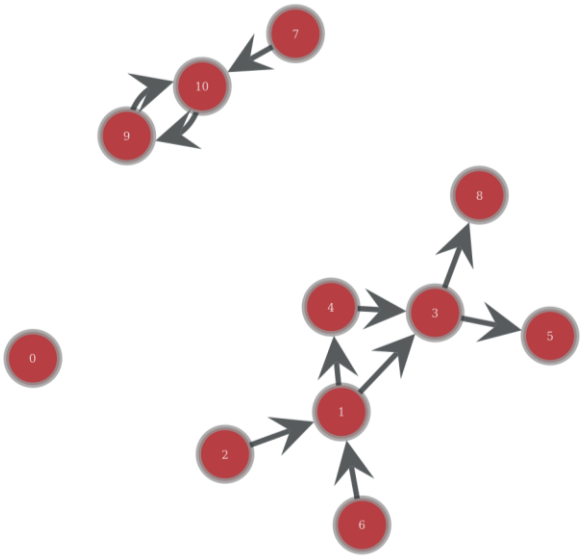


Figure 8. From a metabolic network to a reaction graph.

Transformation of a metabolic network into a reaction graph, where nodes are enzymatic reactions linked by shared metabolites. The transformation creates several connected components.

3.3.1. Node centralities

Each centrality measure rates the importance of a node differently, thus the need to apply different measures depending on how we define important nodes. Are important nodes the ones with more connections? Or the ones with fewer connections but that link distinct parts of the network?

In the work presented in this thesis, the centrality measures used are: degree (in and out-degree), spin, closeness centrality, betweenness centrality, and eigenvector centrality.

Degree (in/out-degree): Number of neighbors of a node. The neighbors of a given node are the nodes connected by a link to that node. In Figure 9, the degree of node 0 is 0.5. In a directed network, we can separate degree in: in-degree (number of incoming links), and out-degree (number of out-going links). In this case, node 0 has in-degree 0.3333 and out-degree 0.1667. If nodes represent enzymatic reactions, a gene with high degree is a highly connected gene in which a mutation would affect many other genes.

Spin: Difference between incoming and outgoing links of a given link, normalized by the node degree. A node with a negative spin has more outgoing than incoming links (node 2 in Figure 9), whereas a positive spin indicates more incoming than outgoing links (node 0). The extreme cases are: nodes with no outgoing links (spin = 1, nodes 4 and 6 in Figure 9), no incoming links (spin = -1, nodes 1 and 5) or with the same number of incoming and outgoing links (spin = 0, node 3).

Betweenness centrality: Number of times a node acts as a bridge between two other nodes. This is calculated using the shortest path between each pair of nodes. Nodes with high betweenness connect different clusters of the network. A node with high betweenness could be describe as a node where "all paths go through it" (Node 3 in Figure 9).

Closeness centrality: It is a measure of how "close" a node is from all the others, the higher the value, the closer or more central a node is. It is calculated as the inverse of the sum of all distances of the node to all other nodes. For directed networks, closeness has not a straightforward interpretation. A node with high closeness is a node where "all paths lead to it", that would be node 2 in Figure 9.

Eigenvector centrality: It is calculated based on the number of neighbors of a given node and on the centrality of those neighbors. A node with high eigenvector centrality is connected to one or more nodes with a high number of connections.

Position within the pathway: In a directed graph, nodes have a position along the pathway: top (in-degree = 0), bottom (out-degree = 0) or intermediate. In Figure 9, nodes 1 and 5 are in top positions,

whereas nodes 4 and 6 are in bottom positions. The rest have intermediate positions.

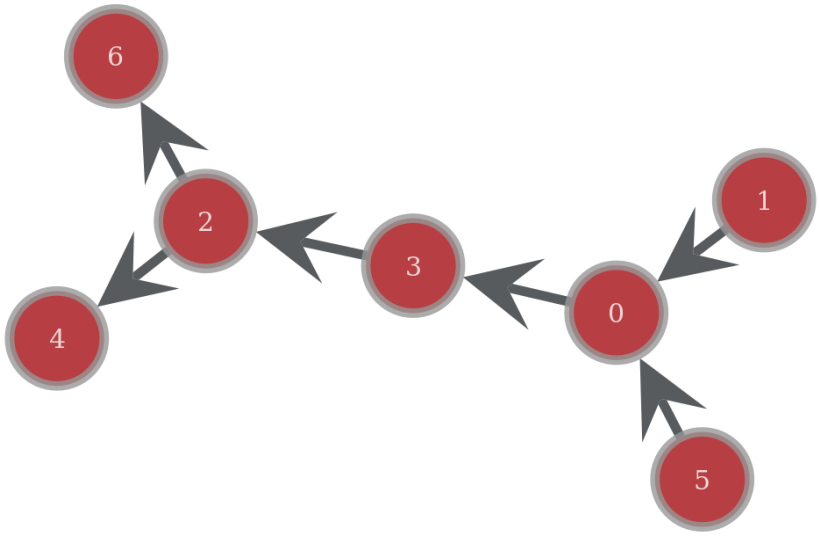


Figure 9. Small directed network.

Depending on the centrality measured chosen, the more important nodes change, see Table 1 *for the centrality values of each node.*

Node	*Degree*	*Indegree*	*Outdegree*	*Spin*	*Closeness*	*Betweenness*
0	0.500	0.333	0.167	0.333	0.296	0.267
1	0.167	0	0.167	-1	0.298	0
2	0.500	0.167	0.333	-0.333	0.333	0.267
3	0.333	0.167	0.167	0	0.300	0.300
4	0.167	0.167	0	1	0	0
5	0.167	0	0.167	-1	0.298	0
6	0.167	0.167	0	1	0	0

Table 1. Centrality measures of the directed network.

Values of centrality measures (degree, in-degree, out-degree, closeness and betweenness) of the nodes in the directed network illustrated in Figure 9.

3.3.1. Network structure and gene evolution

Previous studies have studied the relation between pathway structure and gene evolution in different systems (Table 2). Most of the studies focused on the effect of purifying selection on node evolution, measured as the dN/dS ratio (ω). Both in whole metabolic networks, protein-protein interaction networks (PIN), and in individual pathways from different organisms, the results show that purifying selection is stronger in highly connected and more central genes.

It is important to mention that not all interactions between nodes have the same confidence or meaning. In PIN, links between nodes can represent at the same time several types of interactions (physical interactions, tissue co-expression). And more importantly, these have no directionality. In the case of metabolic networks directionality can be derived from the physiological route of metabolite production, but not all studies took advantage of that feature (Greenberg, Stockwell, & Clark, 2008).

Accounting for the direction of the interaction allows to measure the action of natural selection depending on the position of the node. However, there are no clear results of the how the strength of purifying selection varies along pathways. The results obtained vary depending on the organism and type of network: in biosynthetic pathways in plants upstream genes are more conserved (Livingstone & Anderson, 2009; Rausher, Miller, & Tiffin, 1999), while downstream genes are more conserved in the Insulin/TOR signal transduction pathway in mammals and flies (Alvarez-Ponce, Aguade, & Rozas, 2008; Alvarez-Ponce, Aguadé, & Rozas, 2011).

Few studies that have analyzed where positive selection acts within a network. Positive selection has acted in peripheric genes in the human (Luisi et al., 2015) and yeast PIN (Chakraborty & Alvarez-Ponce, 2016). Remarkably, the same study found the opposite trend in the fly PIN: positive selection was detected mostly in central genes (Chakraborty & Alvarez-Ponce, 2016).

Only two studies have analyzed the effect of positive selection based on polymorphism data (intraspecific) in the human PIN (Luisi

et al., 2015; Qian, Zhou, & Tang, 2015). Both studies reached the same conclusion: recent positive selection has acted in more central genes. A similar result was observed in the Insulin/TOR signal transduction pathway (Luisi et al., 2012).

Type of network	Organism	Result	Study
Metabolic pathways	Fly	Purifying selection stronger in highly connected genes (interspecific)	(Greenberg et al., 2008)
Metabolic network	Yeast	Purifying selection stronger in highly connected genes (interspecific)	(Vitkup, Kharchenko, & Wagner, 2006)
Metabolic network	Mammals	Purifying selection stronger in central genes (interspecific)	(Hudson & Conant, 2011)
Protein-protein interaction network	Yeast	Purifying selection stronger in highly connected genes (interspecific)	(H. B. Fraser, Hirsh, Steinmetz, Scharfe, & Feldman, 2002)
Protein-protein interaction network	Mammals	Purifying selection stronger in central genes (interspecific)	(Luisi et al., 2015)
Protein-protein interaction network	Human	Purifying selection stronger in central genes (intraspecific)	(Luisi et al., 2015)
Phototransduction pathway	Mammals	Purifying selection stronger in central genes (interspecific)	(Invergo, Montanucci, Laayouni, & Bertranpetit, 2013)
Asparagine N-glycosylation pathway	Primates	Purifying selection stronger in downstream genes (interspecific)	(Montanucci, Laayouni, Dall'Olio, & Bertranpetit, 2011)
Insulin/TOR signal transduction pathway	Fly	Purifying selection stronger in downstream genes (interspecific)	(Alvarez-Ponce et al., 2008)
Insulin/TOR signal transduction pathway	Vertebrates	Purifying selection stronger in downstream genes (interspecific)	(Alvarez-Ponce et al., 2011)
Anthocyanin biosynthetic pathway	Plants	Purifying selection stronger in upstream genes (interspecific)	(Rausher et al., 1999)
Carotenoid biosynthetic pathway	Plants	Purifying selection stronger in upstream genes (interspecific)	(Livingstone & Anderson, 2009)
Metabolic pathways	Mammals / Human	Purifying selection stronger in upstream genes and genes with high in-degree	(Montanucci et al., 2018)

		Purifying selection stronger in genes of the inner core layer (interspecific)	
Metabolic pathways	Fly	Positive selection acted in genes at branch points (intraspecific)	(Flowers et al., 2007)
Asparagine N-glycosylation pathway	Human	Positive selection acted in genes at branch points (intraspecific)	(Dall'Olio et al., 2012)
Metabolic pathways	Mammals / Human	Positive selection acted in genes with high out-degree (interspecific)	(Montanucci et al., 2018)
Protein-protein interaction network	Mammals / Human	Positive selection acted in peripheric genes (interspecific)	(Kim, Korbel, & Gerstein, 2007; Luisi et al., 2015)
Protein-protein interaction network	Yeast	Positive selection acted in peripheric genes (interspecific)	(Chakraborty & Alvarez-Ponce, 2016)
Protein-protein interaction network	Fly	Positive selection acted in central genes (interspecific)	(Chakraborty & Alvarez-Ponce, 2016)
Protein-protein interaction network	Human	Positive selection acted in central genes (intraspecific)	(Luisi et al., 2015; Qian et al., 2015)
Insulin/TOR signal transduction pathway	Human	Positive selection acted in central genes (intraspecific)	(Luisi et al., 2012)

Table 2. Summary of previous studies relating network topology and evolutionary rates.

Studies are classified based on the type of network, organism and whether they estimated purifying selection or positive selection based on divergence (interspecific) or polymorphism (intraspecific) data.

Thus, while differences in the methodology to estimate evolutionary measures and the choice of database, could be behind some of the contradictory results, it is also possible that constraints imposed by network structure are lineage-specific (Chakraborty & Alvarez-Ponce, 2016).

More interestingly, the difference between long-term positive selection (as measured by interspecific data) and recent or short-term positive selection (as measured by intraspecific data), suggests that positive selection has targeted distinct parts of the network at

different evolutionary time-scales (Luisi et al., 2015). The authors argued that this phenomenon is explained by the Geometric Model of Adaptation (FGM) (Fisher, 1930) and by the features of the tests to detect positive selection at intra and interspecific levels.

The FGM describes the phenotype of an organism as point in a high-dimensional space, where the dimensions are independent phenotypic traits. As the phenotypic complexity increases (more dimensions), it will be less likely that a mutation will have beneficial effects, as it will be impossible to be advantageous in all dimensions or traits at the same time: it will be more likely to be beneficial if the effect of the mutations is small. However, this depends on how far from the optimum fitness is the organism. When an organism is far from the optimum, mutations with large effects are more advantageous, whereas if it is close to the optimum a mutation with small effect will be more advantageous (Fisher, 1930; Tenaillon, 2014). One can draw a comparison with golf: a golfer's first shot intends to get the ball as close to the hole as possible (without going too far), subsequent shots should be smaller and precise.

This framework should be interpreted in the light of how we detect positive selection. Long-term positive selection is detected by calculating the rate of synonymous and nonsynonymous substitutions (ω). Therefore, it can be estimated only in protein-coding regions. Conversely, short-term positive selection is detected on both coding and non-coding regions of the genome, as it is based on polymorphisms.

Thus, if we join these two points, the result that positive selection acted in peripheric genes (interspecific) is explained by the study of adaptation over a long time-scale on protein-coding genes: it is more likely that mutations with smaller effects will be more advantageous. Thus, more events of positive selection are detected in genes with small effects on the phenotype (less central or less connected).

The result that positive selection acted in central genes (intraspecific) is explained by the study of recent adaptation in the whole genome of a species far from the optimum (OOA, Mesolithic-Neolithic transition, extreme environments…):

mutations with larger effects will be more advantageous. We will detect more positive selection in genes with larger effects (more central or more connected) and in regulatory regions that affect many phenotypic traits (Luisi et al., 2015).

Further studies, where positive selection is detected in the same system at both intra and interspecific level are needed to confirm this idea. Also, to know whether there is a general rule for the constraints imposed by network structure, or if the patterns are organism-specific, we need a systematic analysis of diverse types of pathways in more organisms.

OBJECTIVES

The main objective of this thesis is to study how selection acted at a molecular level on human populations. To understand the forces driving adaptive evolution, two approaches were selected: one, the detection of population-specific adaptations and, second, the identification of the constraints given by the position and connections of the enzymes in a metabolic pathway.

First, to increase the knowledge on human genetic variability and how infectious diseases affected the evolution of our genomes, two studies were performed. A genetic study on East African populations, highlighting the need of including diverse groups to truly capture the diversity of a region (See Results: 1. The genetics of East African populations: A Nilo-Saharan component in the African genetic landscape). Second, a study on the Roma people was carried out to analyze their recent evolutionary history and see how their migration affected the selective pressures shaping their immune response (See Results: 2. The shaping of immunological response through natural selection after migration: the case of the Roma).

Then, based on the idea of studying patterns of positive selection in genes involved in the same process, a study about the evolution of genes related to taste and phase I biotransformation in humans was performed (See Results: 3. Is there adaptation in the human genome for taste perception and phase I biotransformation).

Last, the structure of the human metabolic network was analyzed together with signals of positive selection estimated from intra and interspecific variation. The goal was to study how the events of positive selection are distributed across a large-scale network, and to compare the emergent patterns to the results obtained when studying individual metabolic pathways (See Results: 4. Influence of network topology on the evolution of metabolic enzymes in humans and mammals).

RESULTS

1. The genetics of East African populations: A Nilo-Saharan component in the African genetic landscape

Dobon B, Hassan HY, Laayouni H, Luisi P, Ricaño-Ponce I, Zhernakova A, et al. The genetics of East African populations: a Nilo-Saharan component in the African genetic landscape. Sci Rep. 2015 Sep 28;5(1):9996. DOI: 10.1038/srep09996

2. The shaping of immunological response through natural selection after migration: the case of the Roma

Begoña Dobón[1], Rob ter Horst[2], Hafid Laayouni[1,3], Mayukh Mondal[4], Elena Bosch[1], Jaume Bertranpetit[1*], Mihai G. Netea[2,5*].

[1] Institut de Biologia Evolutiva (UPF-CSIC), Universitat Pompeu Fabra, Doctor Aiguader 88 (PRBB), 08003 Barcelona, Spain.
[2]Department of Internal Medicine, Radboud University Medical Center, 6525 GA Nijmegen, the Netherlands.
[3]Bioinformatics Studies, ESCI-UPF, Pg. Pujades 1, 08003 Barcelona, Spain
[4]Institute of Genomics, University of Tartu, Tartu, Estonia
[5]Department of Internal Medicine, Radboud University Medical Center, 6525 GA Nijmegen, the Netherlands; Department for Genomics & Immunoregulation, Life and Medical Sciences Institute (LIMES), University of Bonn, 53115 Bonn, Germany.

*Corresponding authors: jaume.bertranpetit@upf.edu or mihai.netea@radboudumc.nl

Keywords: Roma, India, positive selection, cytokine production, immune system

Abstract
The Roma people departed the Indian subcontinent around 1000-1500 years ago, and part of them settled in the Romanian territory around the 14th century. The analysis of whole-genome sequences of Roma individuals along with individuals from their host population, Romania, suggests that pathogens where an important selective pressure, before and after the Roma diaspora, and that Romanian Roma have suffered rapid adaptation in variants associated with differential cytokine production.

Background
The Roma people, also called Romani/Rroma or with the derogatory term of Gypsies, represent the largest ethnic minority in Europe. Due to the nomadic lifestyle of some of the groups and the

social exclusion that the Roma have suffered, their real number in the continent is unknown, but estimates vary between 10 and 12 million. For still unknown reasons, the Roma departed the Indian subcontinent around 1000-1500 years ago. They traveled through Persia and Armenia, reaching the Balkan peninsula between the 11th and 12th centuries (Achim, 1998; Fraser, 1992). The first record of the presence of Roma in Romanian territory dates to the 14th century (Achim, 1998). Nowadays, they represent between 3-5% of the Romanian population, being the second minority in the country after Hungarians (6.5%).

The study of the history of the Roma, which lacks written records, has relied on linguistic, sociological, and later, genetic studies. The main topics investigated in genetic studies about the Roma are: i) place of origin in the Indian subcontinent; ii) migrations from and to population from the host countries, and iii) similarities between Romani groups from different countries. The analyses of uniparental markers (mitochondrial DNA and Y-chromosome haplogroups) gave support to the linguistic studies suggesting the Indian origin of the Roma (Martínez-Cruz et al., 2015; Mendizabal, Valente, & Gusmao, 2011; Rai et al., 2012). Genome-wide data further narrowed the putative population of origin to those currently inhabiting the northwestern region of the Indian subcontinent (Mendizabal, Lao, & UM, 2012; Mendizabal et al., 2011; Moorjani, Patterson, et al., 2013).

In this study we present whole-genome sequences of Roma individuals along with individuals from their host population, Romania, to investigate what selective pressures the Roma faced before and after leaving India. This approach has been proven successful to identify signatures of convergent evolution in immunological genes caused by the plague in the Roma (Laayouni et al., 2014).

Methods

Samples

We generated whole genome sequences of 50 Roma (Romani) and 50 Romanian individuals from Romania with an average of 15X coverage (see Supplementary Note 1 for technical details). Informed consent was obtained from all individuals. After strict quality control (Supplementary Note 2 and 3, Supplementary

Figures 1-6) we were left with 40 unrelated Romanians and 40 unrelated Roma for the main analyses (see Supplementary Table 1 for reasons of exclusion).

Population demographic analyses

We combined the new data generated in this study with populations covering the genetic diversity present on continental India: 10 Rajput (RAJ), 10 Uttar Pradesh Upper Caste Brahmins (UBR), nine Vellalar (VLR), 10 Irula (ILA), 10 Riang (RIA), and nine Birhor (BIR) (Mondal et al., 2016). For further information about the Indian populations see Mondal et al. 2016. We also added 40 unrelated individuals from each of the following populations from 1000 Genomes Project Phase 3 (Auton et al., 2015): CEU (Utah Residents (CEPH) with Northern and Western European Ancestry), TSI (Tuscans in Italia), FIN (Finnish in Finland), GBR (British in England and Scotland), IBS (Iberian population in Spain), CHB (Han Chinese in Beijing, China), JPT (Japanese in Tokyo, Japan), CHS (Southern Han Chinese), CDX (Chinese Dai in Xishuangbanna, China), KHV (Kinh in Ho Chi Minh City, Vietnam), GIH (Gujarati Indian from Houston, Texas), PJL (Punjabi from Lahore, Pakistan), BEB (Bengali from Bangladesh), STU (Sri Lankan Tamil from the UK), ITU (Indian Telugu from the UK), YRI (Yoruba in Ibadan, Nigeria), LWK (Luhya in Webuye, Kenya), GWD (Gambian in Western Divisions in the Gambia), MSL (Mende in Sierra Leone), and ESN (Esan in Nigeria).

We filtered the data with PLINK 2.0 (Chang et al., 2015) to keep only bi-allelic autosomal SNPs with MAF > 0.05, under Hardy-Weinberg Equilibrium and without missing data, obtaining a dataset with 938 samples and 5,216,078 SNPs. This dataset was pruned for linkage disequilibrium (LD) in 515,723 SNPs. We performed a principal component analysis (PCA) with EIGENSOFT v6.1 (Patterson, Price, & Reich, 2006) on the pruned dataset without the African populations (YRI, LWK, GWD, MSL, and ESN) and runs of homozygosity (ROH) were estimated by PLINK. The admixture analysis with ADMIXTURE v1.23 (Alexander et al., 2009) was run with values of K ranging from 2 to 9, each 25 times with 5-fold cross-validation and different seed to estimate the best supported model.

We applied the 3-Population Test implemented in qp3Pop, Admixtools (Patterson et al., 2012), to test whether Roma are an admixed population in the form of f3(Roma; European, Indian), where European and Indian are populations from the 1000 Genomes Project (Auton et al. 2015) and from Mondal et al. (2016). We also calculated the populations that Roma share more genetic drift with the outgroup f3(Romani; X, YRI), where X is a European, an Indian population, or the newly sequenced Romanians.

Scan of selection

The objective of this analysis was to identify the selective pressures that Roma people faced prior to leaving their place of origin (North India) and after their settlement in Romania. We used the Cross-population Extended Haplotype Homozygosity (XP-EHH) test (Sabeti et al., 2007) to detect recent signals of positive selection that are shared by two populations but not the third: signals shared by Roma and North Indians (and not Romanians) will be older while those shared by Roma and Romanians (and not North Indians) will be more recent. XP-EHH was run using selscan (Szpiech & Hernandez, 2014) after phasing the data with SHAPEIT2 (Delaneau & Marchini, 2014) with the 1.000 Genomes phase 3 reference panel of haplotypes. As needed by selscan, we phased each population separately without allowing for missing data. We obtained the genetic position and ancestral allele information from the 1000 Genomes Project (Auton et al., 2015).

We analyzed 40 Roma and 40 Romanians and 10 Rajput individuals from Mondal et al. (2016). To minimize any biases due to sequencing technologies or variant calling algorithms, we selected as the best proxy for a North Indian population from the putative area of origin of the Roma people (Mendizabal et al., 2012), the Rajput population from our dataset Rajput. XP-EHH was run using default parameters, only reducing the maximum allowed gap between two SNPs from 200.000 to 20.000 bp to avoid spurious peaks. We performed three comparisons: Roma vs. Romanian, Roma vs. North India, and Romanian vs. North India. We calculated the average value of XP-EHH in 30kb windows with an overlap of 5kb.

We selected the windows shared between the 5% upper tail genome-wide distribution of the Romanian vs. North India and

Roma vs. North India comparisons. From those we removed the windows belonging to the 5% upper and lower tail of the Roma vs. Romanian comparison. We removed SNPs with |XP-EHH| < 2. These signals would indicate the selective pressures that Roma people faced when they established themselves in Romania, from now on called "Recent Shared Signals". Following the same rationale, we selected the windows shared between the 5% upper tail genome-wide distribution of the Roma vs. Romanian and North India vs. Romanian comparisons. From those we removed the windows belonging to the 5% upper and lower tail of the Roma vs. North India comparison. We removed SNPs with |XP-EHH| < 2. These signals would indicate the selective pressures that Roma people faced prior to leaving India, from now on called "Old Shared Signals". We performed a two-sided Gene Ontology (GO) enrichment analyses (Enrichment/Depletion) and pathway annotation network tests with Cytoscape (Shannon et al., 2003) plug-in ClueGo (Bindea et al., 2009) in both Recent and Old Shared Signals. P-values were corrected for multiple testing by the Benjamini-Hochberg procedure.

cQTL enrichment analysis

Windows belonging to the Shared Signals, both Recent and Old, were pruned for LD (SNPs within 1Mb, $R^2 > 0.8$) and intersected with the cQTL (cytokine QTL) dataset from (Li et al., 2016). We assessed whether there was an enrichment of cQTLs in the selection signals by a randomization test. We selected SNPs identified as cQTLs with a p-value threshold <= 1e-5.

Functional validation of the pathways

We tested how the inhibition of three pathways obtained in the selection analyses affected cytokine production capacity of the cell during an infection: i) mTOR mediated cellular metabolism pathway was inhibited with Rapamycine and Ascorbate; ii) Adenylate cyclase pathway was inhibited with KH7; iii) Histone deacetylation pathway mediated by HDAC9 was inhibited with TMP269. DMSO was used as control when testing the Adenylate Cyclase pathway, whereas RPMI was used for the other two pathways. A total of 6 stimulations were tested: RPMI as negative control; *Y.pestis antigua* (10^6/ml); *Y.pestis antigua* (10^5/ml); Influenza (x10); MTB (5 µg/ml); and *C. albicans* (10^6/ml). For every stimulation we measured the expression levels of 7 cytokines:

IL-1β, IL-6, TNF, IL-10, IFNγ, IL-17 and IL-22. These were all measured for a total of 8 subjects (3 batches of sizes 3-3-2). Differences were assessed by permutations within batches (paired t-test, two-sided p-value).

Results and Discussion

The Roma people are genetically more similar to Europeans than to Indian populations

To explore the genetic relationship between the Roma and other worldwide populations, after quality control, we performed a principal component analysis (PCA) (Figure 1a). PC1 separates between European and Indian populations from Asian, whereas PC2 differentiates European populations (including Romanians) from Indian and Roma populations. Roma fall in a cline between European/Romanians and Indian populations, with the closest Indian populations being those geographically located in North India: Rajput (RAJ), Uttar Pradesh Upper Caste Brahmins (UBR) (UBR) and Punjabi (PJL). When only Roma and Romanians populations are analyzed PC1 separates Romanies from Romanians, with the later forming a tight cluster (Supplementary Figure 6). We observed several individuals, both Roma and Romanian, spread between the two clusters, indicating genetic flow between both populations and lack of genetic and social correspondence.

In an admixture analysis (Figure 1b, Supplementary Figure 8a), the Roma appear as an admixed population with a 30% Indian and a 70% European component (K = 3). It is at K = 4 when Roma show their own genetic component (best supported model, Supplementary Figure 8b). This component can also be seen in small proportions in the European Romanian (RMN), TSI and IBS populations (populations with a known presence of Roma in their countries). Then, we estimated how genetically similar are the Roma to other worldwide populations by the outgroup f3-statistics. Roma share more genetic drift with Central or Eastern European populations than with Indians (Figure 2). It has been suggested that the European ancestry present within India, increases the genetic similarity of the Roma with other European populations (Moorjani et al. 2013). Further analyses with more robust statistics, such as Dstat or qpAdmix (Patterson et al., 2012), and the combination of methods that complement the admixture analysis (Lawson, van Dorp, & Falush, 2018), are needed to disentangle the ancestry of the

Roma, as their putative population of origin has themselves a complex recent demographic history (Moorjani, Thangaraj, et al., 2013).

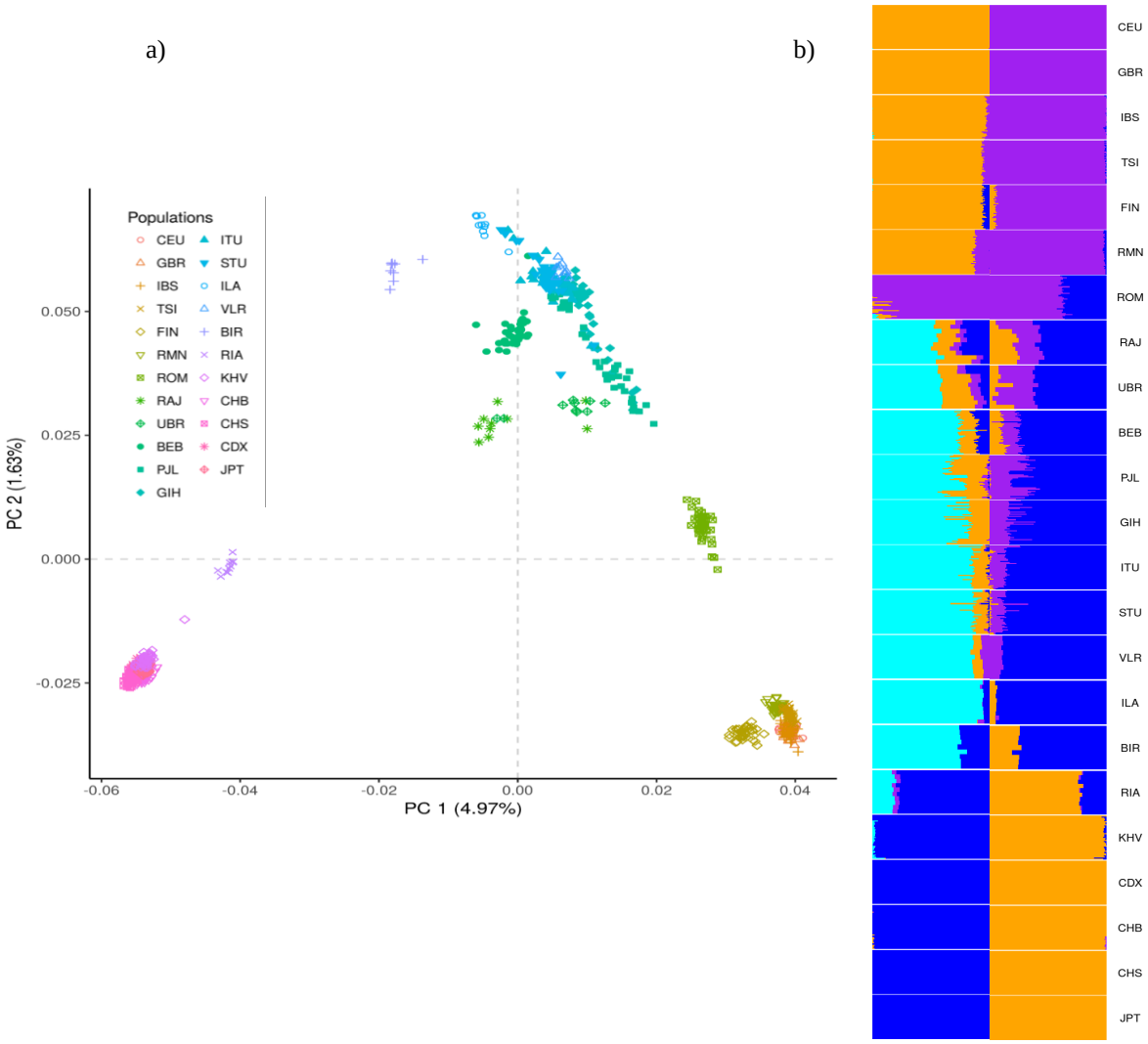


Figure 1. a) Principal component analysis of the Romani (ROM) from Romania in the context of other worldwide populations. Showing the first two principal components and the variance explained by them; b) Clustering analysis showing K = 3 and K = 4. Romanies show their own component in K = 4 (best supported model). See Samples section for a description on the populations.

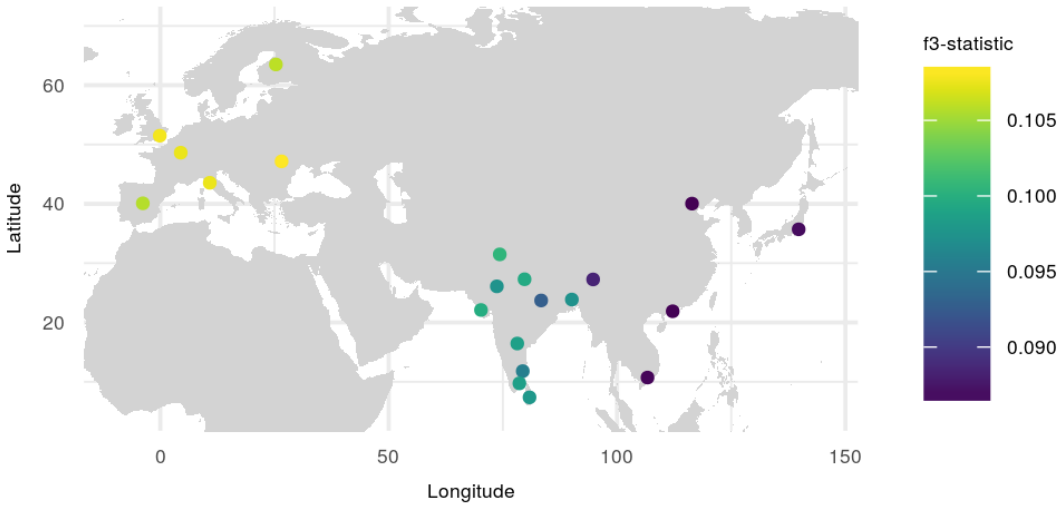
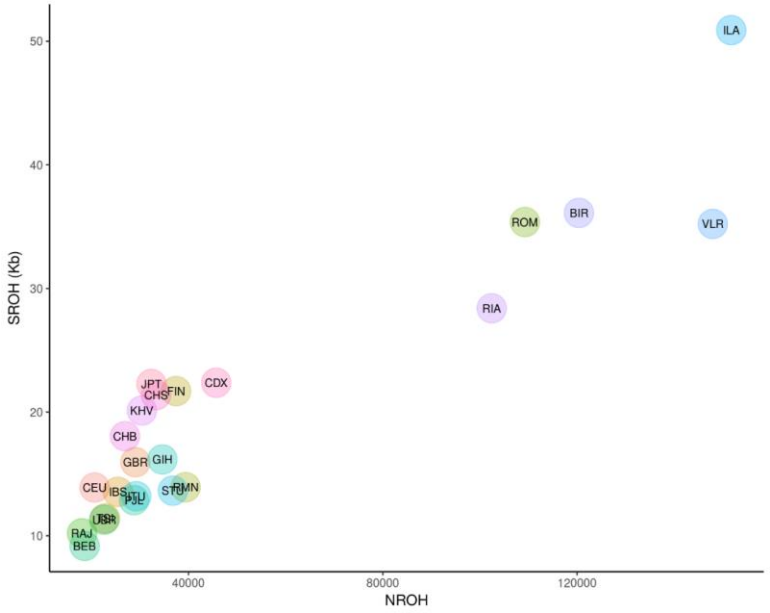


Figure 2. Proportion of shared genetic drift between the Roma and extant worldwide populations measured using f3 (Roma; X, YRI); where X is either Romanians, another European population, or an Indian population. Roma share more drift with Europeans than with Indians populations.

Signals of bottlenecks and endogamy in Roma people

By comparing the number and length of runs of homozygosity (ROH) we can infer the demographic history of a population (Ceballos, Joshi, Clark, Ramsay, & Wilson, 2018). Romanies show a unique profile with respect to other European populations (Figure 3a). As a population that suffered a strong bottleneck after the departure from India and kept a small effective population size, it has more of both longer and shorter ROHs that other populations with higher effective sizes (Figure 3b and Supplementary Figure 10). The practice of consanguineous marriage in the Roma (Kalaydjieva, Morar, Chaix, & Tang, 2005) increases the variance in the length of ROHs as the offspring of those unions will have a small number of very long ROHs. Tribal Indian populations (RIA, BIR, VLR), present a similar profile, with ILA showing the strongest bottleneck signal.

a)


SROH (Kb)
NROH
50
40
30
20
10
40000
80000
120000
ILA
BIR
ROM
VLR
RIA
CDX
JPT
CHS
FIN
KHV
CHB
GBR
GIH
CEU
IBS
STU
RMN
PJL
ITU
RAJ
BEB

b)

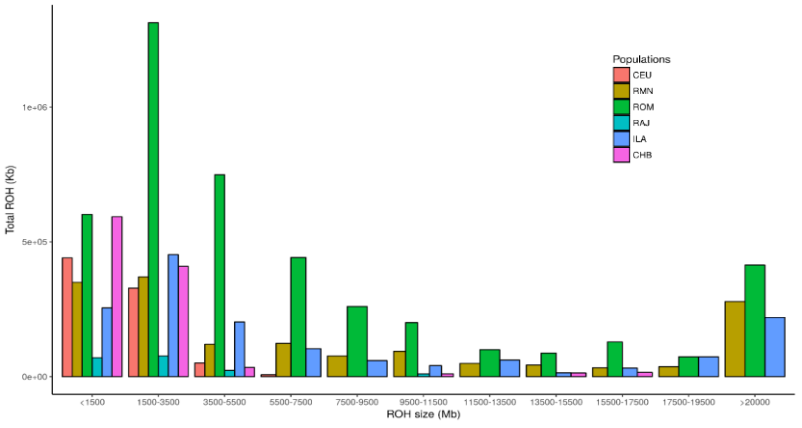
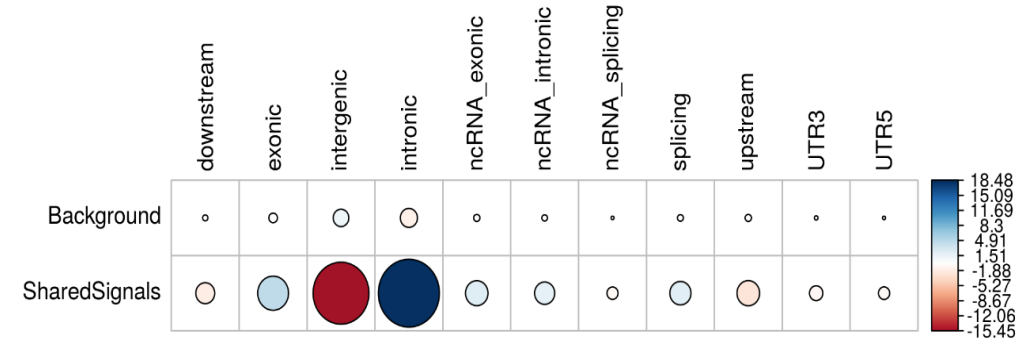
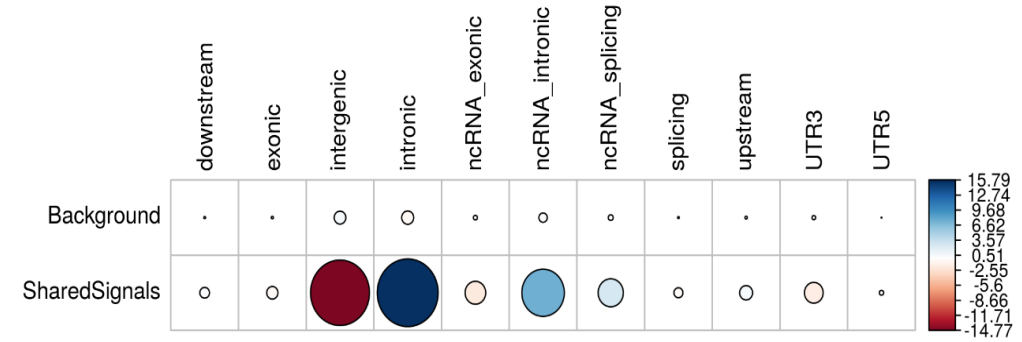

Total ROH (Kb)
ROH size (Mb)
Populations
CEU
RMN
ROM
RAJ
ILA
CHB
1e+06
5e+05
0e+00
<1500
1500-3500
3500-5500
5500-7500
7500-9500
9500-11500
11500-13500
13500-15500
15500-17500
17500-19500
>20000

Figure 3. a) Total number of ROH (NROH) versus the sum of the total length of ROH in Kb (SROH) in worldwide populations. Each dot represents population means. b) Distribution of the total length of runs of homozygosity (ROHs) classified by length categories in Romanies (ROM), Romanians (RMN), CEU, CHB, Rajput (RAJ) and Irula (ILA). See Supplementary Figure 10 for a comparison with all worldwide populations.

a)



b)



c)

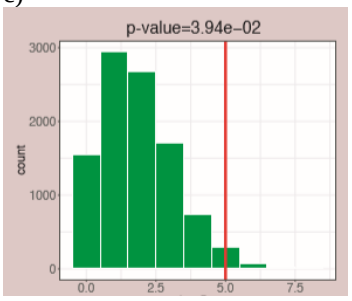


Figure 4. Enrichment of SNPs functional categories in the signals of selection. a) Recent Shared Signals (Pearson's Chi-squared test, χ^2 = 632.89, p-value < 2.2e-16); b) Old shared Signals (Pearson's Chi-squared test, χ^2 = 541.57, p-value < 2.2e-16); c) cQTL enrichment. Number of cQTLs found in the Recent Shared

Signals (red line) compared to a sampling distribution. Circle size is proportional to the contribution of each SNPs functional category to the total χ^2 score indicated by the Pearson residuals. Positive values (in blue) indicate that the proportion of that category is higher than expected in the signals whereas negative values (in red) indicate that that signals are depleted in that functional category.

Genome distribution of the selection signals

In the analysis of the footprints of positive selection left in the genome, two levels of analysis are compared: recent and old signals of adaptive selection. The first will be composed of the common signals between Romanies and Romanians, and not in Indians (RAJ) (Supplementary Figure 11a) and the second of the common between ROM and RAJ and not the RMN (Supplementary Figure 11b).

In the Recent Shared Signals (28,640 SNPs) we found more SNPs located in the genic region (intronic and exonic) than expected by comparing with a genome-wide distribution; and less belonging to intergenic regions (Pearson's Chi-squared test, $\chi^2 = 632.89$, p-value < 2.2e-16) (Figure 4a). In the Old Shared Signals (7,205 SNPs) we found more SNPs located in the intronic region of genes and ncRNAs than expected (Pearson's Chi-squared test, $\chi^2 = 541.57$, p-value < 2.2e-16) (Figure 4b), but not in exons, highly enriched in recent signals. Accordingly, we identified 28 non-synonymous changes potentially linked to Recent Shared Signals but only 4 in Old Shared Signals (Supplementary Note 4). Notably, for both Recent and Old Shared Signals, we found a striking enrichment in intronic regions (and a dearth in intergenic), indicating that most of the signals are expected to be related to gene regulation and not to amino acid changes in the coded proteins. Indeed, recent selection in Romani and Romanians targeted variants that affect the expression of cytokines (Figure 4c). No such enrichment was found in Old shared signals.

What selective pressures faced the Romani people after settling in Romania?

As a first approximation to identify the selective pressures faced by the Roma people in their new environment, we performed an enrichment pathway analysis on genes in the Shared selection signals (Figure 5a). Several of the pathways enriched in the Recent Shared Signals are related to housekeeping processes, involved in

functions that cannot be linked to specific phenotypes that could be at the base of the action of selection. This result seems to follow the omnigenic model (Boyle, Li, & Pritchard, 2017), in which the strong interconnexion among the gene regulation networks may cause to find signals (of susceptibility in GWAS studies, of positive selection in genome scans) that are not of direct relevance for the selected phenotype.

For that reason, we have chosen specific pathways related to the immunological function. These pathways have been deeply characterized, and the knowledge of the regulatory networks and their physiological function may allow a direct link between a complex genotype and the phenotype. As a preliminary analysis detected an enrichment of cQTLs (cytokine QTL) in the signals, and among the enriched pathways we found the GO term regulation of cytokine-mediated signaling pathway, we selected two pathways likely to influence cytokine production capacity for further validation: mammalian target of rapamycin (mTOR) mediated cellular metabolism and signaling pathways (carbohydrate biosynthetic process and Ras GTPase binding) (Huang & Fingar, 2014; Weichhart & Säemann, 2008) and adenylate cyclase pathway.

Inhibition of mTOR mediated cellular metabolism pathway generates a pro-inflammatory profile characterized by a decrease of IL-10 and IL-17, and an increase of IL-1β (Figure 5c), supporting the idea that mTOR participates in the establishment of a proinflammatory or anti-inflammatory profile in immune cells (Weichhart & Säemann, 2008). The inhibition of this pathway affects the expression of cytokines in the presence of all pathogens tested except the influenza virus. We found a similar pro-inflammatory profile when we inhibited the adenylate cyclase pathway. However, in this case the strongest stimulus is given the infection by the influenza virus, as it strongly increases the production of several cytokines (IL-1β, IL-6, TNF and IFNγ). This suggests that, even though these pathways respond preferentially to a specific type of pathogen, its response is executed through similar mechanisms.

Moreover, a third pathway was selected due to the presence of MEF2B (Myocyte Enhancer Factor 2B) among the top 100 signals: the histone deacetylation pathway mediated by HDAC9. HDAC9

represses MEF2-related transcriptional activity (Zhou, Marks, Rifkind, & Richon, 2001) and MEF2 is required for B cell proliferation and survival after antigen receptor stimulation (Jain et al., 2015). We did not observe a clear pattern on how the inhibition of HDAC9 impacts cytokine production (Figure 5c). This could be because the histone deacetylation pathway is not immunological mediated, or it responds specifically to a given pathogen that we did not test for.

What selective pressures faced the Romani people in India?

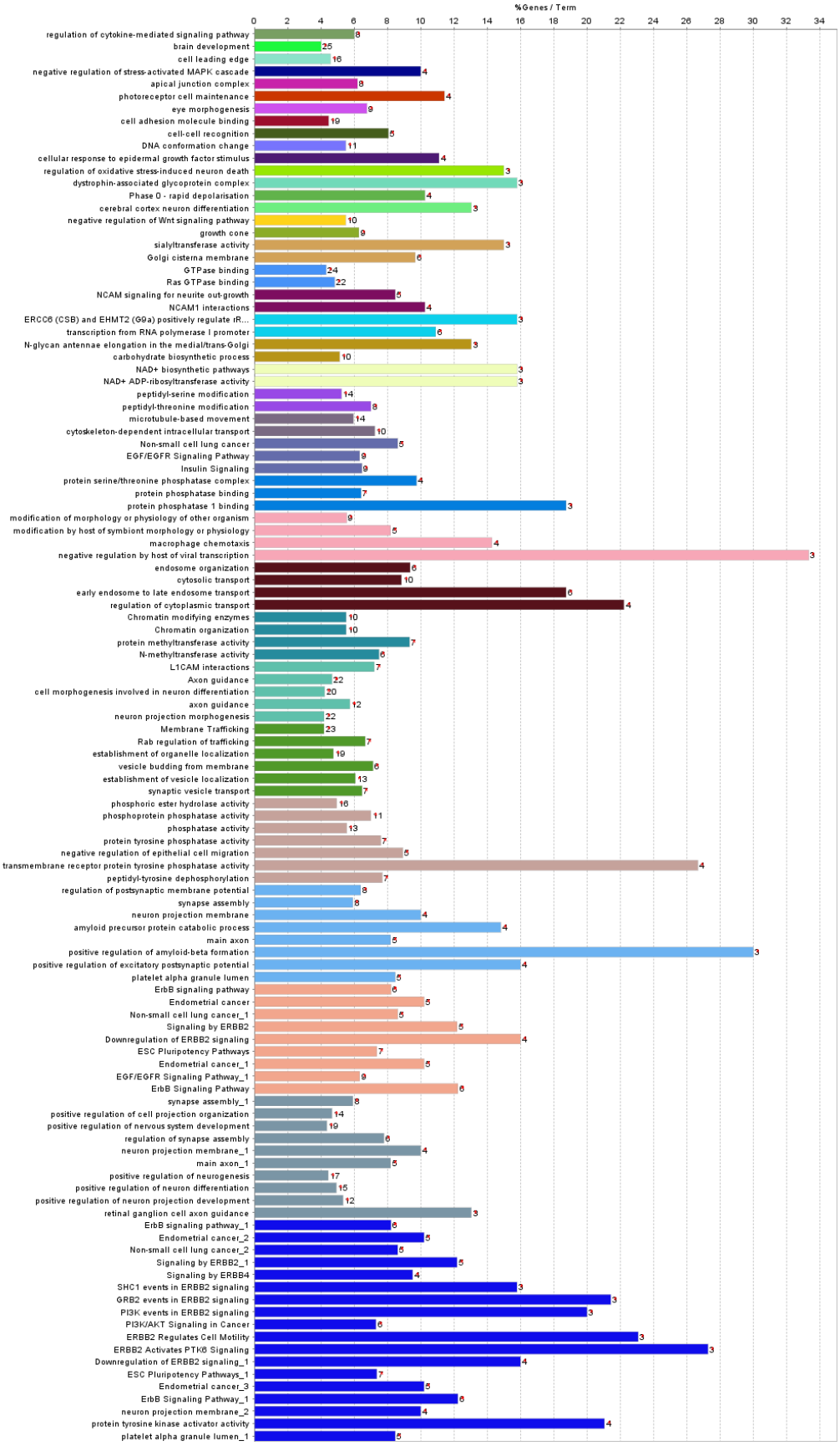
Following the same rationale as when estimating the Recent Shared Signals, we identified the selection signals shared between Romanies and a North Indian population (Old Shared Signals). A general view of the enrichment pathway analysis in the Old Shared signals shows only 20 significant enriched pathways (Figure 5b), distributed in two main groups: terms related to the metabolism of nucleotides, and terms related to immunological processes.

Among the pathways found, there was an enrichment in genes involved in tuberculosis. Tuberculosis is one of the main causes of death worldwide, with India and China being the countries with the highest incidence of the disease (Sulis, Roggi, Matteelli, & Raviglione, 2014). Further functional studies are needed to determine the importance and prevalence of this disease in the history of the Roma.

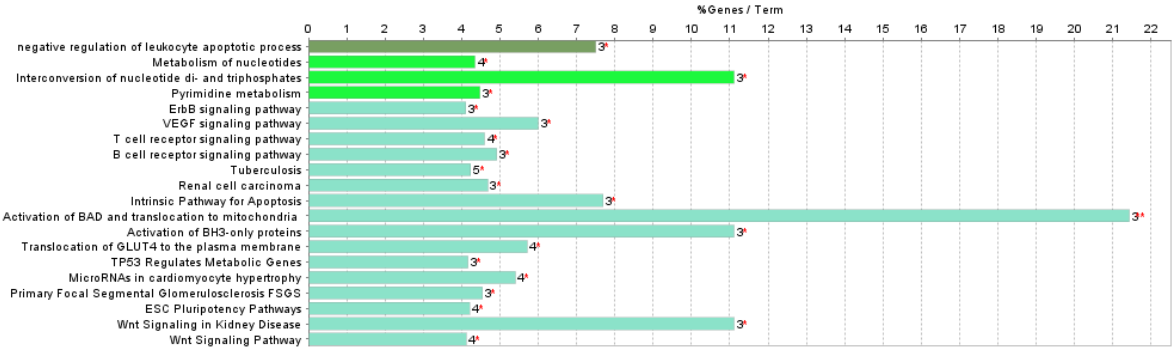
Conclusion

We have analyzed whole-genome sequences of Roma individuals together with individuals from their host population, Romania, and their source population, India. We show that immunological processes are one of the strongest selective pressures during human evolution (Barreiro & Quintana-Murci, 2010; Daub et al., 2013) and they have left their mark in the genome of the Roma. Specifically, positive selection targeted regulatory variants that affect cytokine production in the Roma.

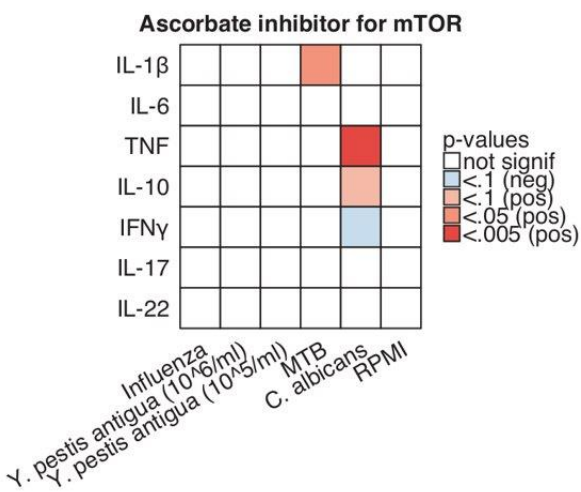
a)

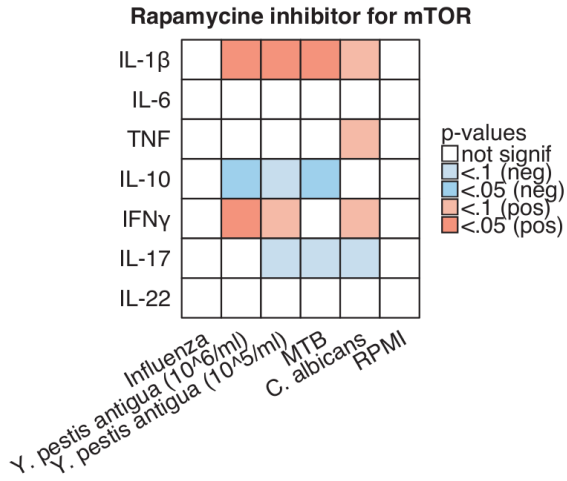

%Genes / Term
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34
regulation of cytokine-mediated signaling pathway 8
brain development 25
cell leading edge 16
negative regulation of stress-activated MAPK cascade 4
apical junction complex 8
photoreceptor cell maintenance 4
eye morphogenesis 9
cell adhesion molecule binding 19
cell-cell recognition 5
DNA conformation change 11
cellular response to epidermal growth factor stimulus 4
regulation of oxidative stress-induced neuron death 3
dystrophin-associated glycoprotein complex 3
Phase 0 - rapid depolarisation 4
cerebral cortex neuron differentiation 3
negative regulation of Wnt signaling pathway 10
growth cone 9
sialyltransferase activity 3
Golgi cisterna membrane 6
GTPase binding 24
Ras GTPase binding 22
NCAM signaling for neurite out-growth 5
NCAM1 interactions 4
ERCC6 (CSB) and EHMT2 (G9a) positively regulate rR... 3
transcription from RNA polymerase I promoter 6
N-glycan antennae elongation in the medial/trans-Golgi 3
carbohydrate biosynthetic process 10
NAD+ biosynthetic pathways 3
NAD+ ADP-ribosyltransferase activity 3
peptidyl-serine modification 14
peptidyl-threonine modification 8
microtubule-based movement 14
cytoskeleton-dependent intracellular transport 10
Non-small cell lung cancer 5
EGF/EGFR Signaling Pathway 9
Insulin Signaling 9
protein serine/threonine phosphatase complex 4
protein phosphatase binding 7
protein phosphatase 1 binding 3
modification of morphology or physiology of other organism 9
modification by host of symbiont morphology or physiology 5
macrophage chemotaxis 4
negative regulation by host of viral transcription 3
endosome organization 6
cytosolic transport 10
early endosome to late endosome transport 6
regulation of cytoplasmic transport 4
Chromatin modifying enzymes 10
Chromatin organization 10
protein methyltransferase activity 7
N-methyltransferase activity 6
L1CAM interactions 7
Axon guidance 22
cell morphogenesis involved in neuron differentiation 20
axon guidance 12
neuron projection morphogenesis 22
Membrane Trafficking 23
Rab regulation of trafficking 7
establishment of organelle localization 19
vesicle budding from membrane 6
establishment of vesicle localization 13
synaptic vesicle transport 7
phosphoric ester hydrolase activity 16
phosphoprotein phosphatase activity 11
phosphatase activity 13
protein tyrosine phosphatase activity 7
negative regulation of epithelial cell migration 5
transmembrane receptor protein tyrosine phosphatase activity 4
peptidyl-tyrosine dephosphorylation 7
regulation of postsynaptic membrane potential 8
synapse assembly 8
neuron projection membrane 4
amyloid precursor protein catabolic process 4
main axon 5
positive regulation of amyloid-beta formation 3
positive regulation of excitatory postsynaptic potential 4
platelet alpha granule lumen 5
ErbB signaling pathway 6
Endometrial cancer 5
Non-small cell lung cancer_1 5
Signaling by ERBB2 5
Downregulation of ERBB2 signaling 4
ESC Pluripotency Pathways 7
Endometrial cancer_1 5
EGF/EGFR Signaling Pathway_1 9
ErbB Signaling Pathway 6
synapse assembly_1 8
positive regulation of cell projection organization 14
positive regulation of nervous system development 19
regulation of synapse assembly 6
neuron projection membrane_1 4
main axon_1 5
positive regulation of neurogenesis 17
positive regulation of neuron differentiation 15
positive regulation of neuron projection development 12
retinal ganglion cell axon guidance 3
ErbB signaling pathway_1 6
Endometrial cancer_2 5
Non-small cell lung cancer_2 5
Signaling by ERBB2_1 5
Signaling by ERBB4 4
SHC1 events in ERBB2 signaling 3
GRB2 events in ERBB2 signaling 3
PI3K events in ERBB2 signaling 3
PI3K/AKT Signaling in Cancer 6
ERBB2 Regulates Cell Motility 3
ERBB2 Activates PTK6 Signaling 3
Downregulation of ERBB2 signaling_1 4
ESC Pluripotency Pathways_1 7
Endometrial cancer_3 5
ErbB Signaling Pathway_1 6
neuron projection membrane_2 4
protein tyrosine kinase activator activity 4
platelet alpha granule lumen_1 5

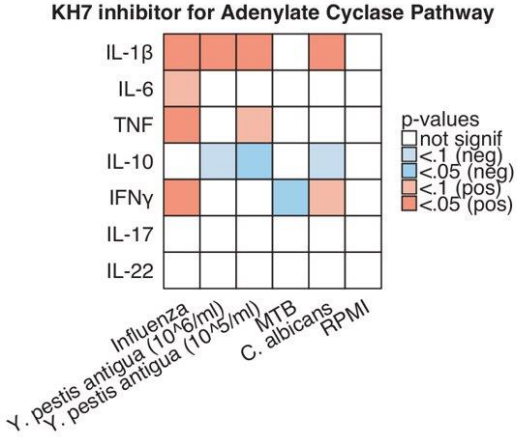
b)


%Genes / Term
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
negative regulation of leukocyte apoptotic process
Metabolism of nucleotides
Interconversion of nucleotide di- and triphosphates
Pyrimidine metabolism
ErbB signaling pathway
VEGF signaling pathway
T cell receptor signaling pathway
B cell receptor signaling pathway
Tuberculosis
Renal cell carcinoma
Intrinsic Pathway for Apoptosis
Activation of BAD and translocation to mitochondria
Activation of BH3-only proteins
Translocation of GLUT4 to the plasma membrane
TP53 Regulates Metabolic Genes
MicroRNAs in cardiomyocyte hypertrophy
Primary Focal Segmental Glomerulosclerosis FSGS
ESC Pluripotency Pathways
Wnt Signaling in Kidney Disease
Wnt Signaling Pathway

c)


Ascorbate inhibitor for mTOR
IL-1β
IL-6
TNF
IL-10
IFNγ
IL-17
IL-22
p-values
not signif
<.1 (neg)
<.1 (pos)
<.05 (pos)
<.005 (pos)
Influenza
Y. pestis antigua (10^6/ml)
Y. pestis antigua (10^5/ml)
MTB
C. albicans
RPMI


Rapamycine inhibitor for mTOR
IL-1β
IL-6
TNF
IL-10
IFNγ
IL-17
IL-22
p-values
not signif
<.1 (neg)
<.05 (neg)
<.1 (pos)
<.05 (pos)
Influenza
Y. pestis antigua (10^6/ml)
Y. pestis antigua (10^5/ml)
MTB
C. albicans
RPMI


KH7 inhibitor for Adenylate Cyclase Pathway
IL-1β
IL-6
TNF
IL-10
IFNγ
IL-17
IL-22
p-values
not signif
<.1 (neg)
<.05 (neg)
<.1 (pos)
<.05 (pos)
Influenza
Y. pestis antigua (10^6/ml)
Y. pestis antigua (10^5/ml)
MTB
C. albicans
RPMI

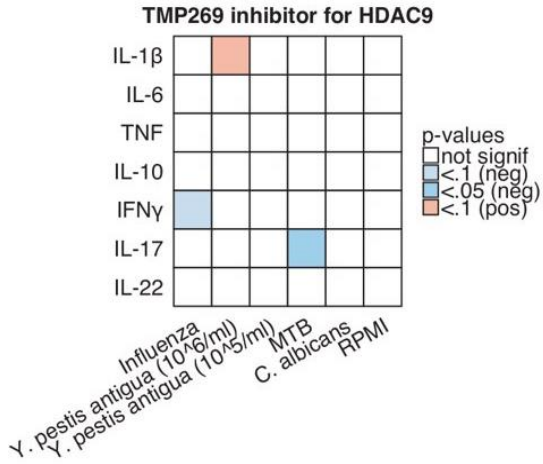

TMP269 inhibitor for HDAC9
IL-1β
IL-6
TNF
IL-10
IFNγ
IL-17
IL-22
p-values
not signif
<.1 (neg)
<.05 (neg)
<.1 (pos)
Influenza
Y. pestis antigua (10^6/ml)
Y. pestis antigua (10^5/ml)
MTB
C. albicans
RPMI

Figure 5. Pathway enrichment analysis based on the genes belonging to a) Recent Shared Signals; b) Old Shared Signals. For each term it is indicated the number (at the end of the bar) and the percentage (length of the bar) of genes belonging to that term that are found under positive selection. Consecutive terms that share at least 50% of the genes are depicted with the same color. Only statistically significant terms are shown (p-value < 0.05, BH-FDR). c) Change in cytokine production after infection and inhibition of mTOR cellular mediated pathway, adenylate cyclase pathway and HDAC9 pathway. Statistically significant changes in expression are indicated with colors: red indicates an increase in cytokine production after inhibition, whereas blue indicates a decrease. P-values were corrected for multiple testing by FDR.

Data availability

Data (BAM, FASTQ and VCF files) is available at European Nucleotide Archive (Accession number: PRJEB28641) and will be released after publication.

Supplementary Information:

- Supplementary Notes 1-4.
- Supplementary Figures 1-11.
- Supplementary Tables 1-4.

References

Achim, V. (1998). *The Roma in Romanian History*. Central European University Press. Retrieved from http://books.openedition.org/ceup/1532

Alexander, D. H., Novembre, J., Lange, K., Alexander D.H.. Novembre J., & K., L. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., … Schloss, J. A. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Barreiro, L. B., & Quintana-Murci, L. (2010). From evolutionary genetics to human immunology: How selection shapes host defence genes. *Nature Reviews Genetics*, *11*(1), 17–30. https://doi.org/10.1038/nrg2698

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M.,

Kirilovsky, A., … Galon, J. (2009). ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, *25*(8), 1091–1093. https://doi.org/10.1093/bioinformatics/btp101

Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*, *169*(7), 1177–1186. https://doi.org/10.1016/j.cell.2017.05.038

Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M., & Wilson, J. F. (2018). Runs of homozygosity: windows into population history and trait architecture. *Nature Reviews Genetics*, *19*(4), 220–234. https://doi.org/10.1038/nrg.2017.109

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1), 7. https://doi.org/10.1186/s13742-015-0047-8

Daub, J. T., Hofer, T., Cutivet, E., Dupanloup, I., Quintana-Murci, L., Robinson-Rechavi, M., & Excoffier, L. (2013). Evidence for Polygenic Adaptation to Pathogens in the Human Genome. *Molecular Biology and Evolution*, *30*(7), 1544–1558. https://doi.org/10.1093/molbev/mst080

Delaneau, O., & Marchini, J. (2014). Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nature Communications*, *5*, 3934. https://doi.org/10.1038/ncomms4934

Fraser, A. (1992). *The Gypsies*. Blackwell Publishers.

Huang, K., & Fingar, D. C. (2014). Growing knowledge of the mTOR signaling network. *Seminars in Cell & Developmental Biology*, *36*, 79–90. https://doi.org/10.1016/j.semcdb.2014.09.011

Jain, P., Lavorgna, A., Sehgal, M., Gao, L., Ginwala, R., Sagar, D., … Khan, Z. K. (2015). Myocyte enhancer factor (MEF)-2 plays essential roles in T-cell transformation associated with HTLV-1 infection by stabilizing complex between Tax and CREB. *Retrovirology*, *12*(1), 23. https://doi.org/10.1186/s12977-015-0140-1

Kalaydjieva, L., Morar, B., Chaix, R., & Tang, H. (2005). A newly discovered founder population: the Roma/Gypsies. *BioEssays*, *27*(10), 1084–1094. https://doi.org/10.1002/bies.20287

Laayouni, H., Oosting, M., Luisi, P., Ioana, M., Alonso, S., Ricano-Ponce, I., … Netea, M. G. (2014). Convergent evolution in European and Rroma populations reveals pressure exerted by

plague on Toll-like receptors. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1317723111

Lawson, D. J., van Dorp, L., & Falush, D. (2018). A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature Communications*, *9*(1), 3258. https://doi.org/10.1038/s41467-018-05257-7

Li, Y., Oosting, M., Smeekens, S. P., Jaeger, M., Aguirre-Gamboa, R., Le, K. T. T., … Netea, M. G. (2016). A Functional Genomics Approach to Understand Variation in Cytokine Production in Humans. *Cell*, *167*(4), 1099–1110.e14. https://doi.org/10.1016/j.cell.2016.10.017

Martínez-Cruz, B., Mendizabal, I., Harmant, C., de Pablo, R., Ioana, M., Angelicheva, D., … Comas, D. (2015). Origins, admixture and founder lineages in European Roma. *European Journal of Human Genetics*, (August), 1–7. https://doi.org/10.1038/ejhg.2015.201

Mendizabal, I., Lao, O., & UM, M. (2012). Reconstructing the population history of European Romani from genome-wide data. *Curr Biol*, *22*, 2342–2349. Retrieved from http://dx.doi.org/10.1016/j.cub.2012.10.039

Mendizabal, I., Valente, C., & Gusmao, A. (2011). Reconstructing the Indian origin and dispersal of the European Roma: a maternal genetic perspective. *PLoS One*, *6*, e15988. Retrieved from http://dx.doi.org/10.1371/journal.pone.0015988

Mondal, M., Casals, F., Xu, T., Dall'Olio, G. M., Pybus, M., Netea, M. G., … Bertranpetit, J. (2016). Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation. *Nature Genetics*, *48*(9), 1066–1070. https://doi.org/10.1038/ng.3621

Moorjani, P., Patterson, N., Loh, P.-R., Lipson, M., Kisfali, P., Melegh, B. I., … Melegh, B. (2013). Reconstructing Roma history from genome-wide data. *PLoS One*, *8*, e58633. https://doi.org/10.1371/journal.pone.0058633.g001

Moorjani, P., Thangaraj, K., Patterson, N., Lipson, M., Loh, P.-R., Govindaraj, P., … Singh, L. (2013). Genetic Evidence for Recent Population Mixture in India. *The American Journal of Human Genetics*, *93*(3), 422–438. https://doi.org/10.1016/j.ajhg.2013.07.006

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., … Reich, D. (2012). Ancient admixture in human history.

Genetics, *192*(3), 1065–1093. https://doi.org/10.1534/genetics.112.145037

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, *2*(12), e190. https://doi.org/10.1371/journal.pgen.0020190

Rai, N., Chaubey, G., Tamang, R., Pathak, A. K., Singh, V. K., Karmin, M., … Thangaraj, K. (2012). The Phylogeography of Y-Chromosome Haplogroup H1a1a-M82 Reveals the Likely Indian Origin of the European Romani Populations. *PLoS ONE*, *7*(11), 1–7. https://doi.org/10.1371/journal.pone.0048477

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., … Hapmap, T. I. (2007). Genome-wide detection and characterization of positive selection in human populations. *October*, *449*(7164), 913–918. https://doi.org/10.1038/nature06250.Genome-wide

Shannon, P., Markiel, A., Owen Ozier, 2, Baliga, N. S., Wang, J. T., Ramage, D., … Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, (13), 2498–2504. https://doi.org/10.1101/gr.1239303.metabolite

Sulis, G., Roggi, A., Matteelli, A., & Raviglione, M. C. (2014). Tuberculosis: Epidemiology and Control. *Mediterranean Journal of Hematology and Infectious Diseases*, *6*(1), e2014070. https://doi.org/10.4084/MJHID.2014.070

Szpiech, Z. A., & Hernandez, R. D. (2014). Selscan: An efficient multithreaded program to perform EHH-based scans for positive selection. *Molecular Biology and Evolution*, *31*(10), 2824–2827. https://doi.org/10.1093/molbev/msu211

Weichhart, T., & Säemann, M. D. (2008). The PI3K/Akt/mTOR pathway in innate immune cells: Emerging therapeutic applications. *Annals of the Rheumatic Diseases*, *67*(SUPPL. 3), 70–75. https://doi.org/10.1136/ard.2008.098459

Zhou, X., Marks, P. A., Rifkind, R. A., & Richon, V. M. (2001). Cloning and characterization of a histone deacetylase, HDAC9. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(19), 10572–10577. https://doi.org/10.1073/pnas.191375098

3. Is there adaptation in the human genome for taste perception and phase I biotransformation?

Dobon B, Rossell C, Walsh S, Bertranpetit J. Is there adaptation in the human genome for taste perception and phase I biotransformation? BMC Evol Biol. 2019 Dec 31;19(1):39. DOI: 10.1186/s12862-019-1366-7

4. Influence of network topology on the evolution of metabolic enzymes in humans and mammals

Montanucci L, Laayouni H, Dobon B, Keys KL, Bertranpetit J, Peretó J. Influence of pathway topology and functional class on the molecular evolution of human metabolic genes. PLoS One. 2018 Dec 1;13(12). DOI: 10.1371/journal.pone.0208782

DISCUSSION

With the development of sequencing technologies, population and evolutionary geneticists have been able to interrogate the genome to understand the molecular basis of natural selection. By applying diverse statistical methods, we can analyze the changes in the genome and describe the evolutionary history undergone by different species and/or populations. The recent advances in genomics data, such as the improvement of sequencing technologies to the point of being able to sequence ancient samples, has supposed a revolution in the field of evolutionary biology and in the study of the molecular mechanisms behind natural selection. In this thesis, the principal goal was to detect, quantify and understand positive selection in several human populations.

The study of positive selection is heavily affected by the approximations, methodologies and sampling strategy used in the study design. Genome coverage (specific loci or genome-wide data), genotyping or sequencing data will determine which part of the genome we are able to interrogate, while the detection of putative events of selection is a function of the sample size and the quality of the data. The four papers presented in this thesis aim to advance our understanding of natural selection using a diverse set of tools and approximations available at the time. In each work, careful considerations were taken to acknowledge, remedy, and take advantage of the limitations of these methods.

The first study presented in this thesis (Results: 1. The genetics of East African populations: A Nilo-Saharan component in the African genetic landscape) is a good example of how to take advantage of the constraints imposed by the characteristics of the genotyping data used. We analyzed genetic data from 9 ethno-linguistic groups from the Sudanese region in East Africa (Sudan, South Sudan and Ethiopia). The inclusion of several ethnic groups from one of the most diverse regions in the world allowed us to characterize a genetic component that identifies a non-admixed set of East African populations: Sudanese Nilo-Saharan speaking groups (Darfurians and part of Nuba populations) and South Sudan Nilotes. The samples were genotyped in the Immunochip (Illumina Infinium

single-nucleotide polymorphism microarray), that contrary to other commercial arrays does not have a uniform coverage of the human genome. The Immunochip was designed for immunogenetics studies and contains 196,524 polymorphisms, most of them associated with major autoimmune and inflammatory diseases (Cortes & Brown, 2011; Trynka et al., 2011). The array also includes ancestry informative markers, but as it was primarily designed for use in white European populations (Cortes & Brown, 2011), we were concerned about how this would affect the discrimination of population structure in African populations. To assess this, we repeated our analysis on population structure using different subsets of markers and we determined that our inferences on population stratification were robust to sample size, batch effects, and genome coverage. Then, as the Immunochip presents a dense SNP coverage of immune-related genes, we were able to analyze how different infectious pressures affected the genome of these populations. We found that selective pressures on anti-malarial and anti-bacterial host defense genes generated lower genetic distances between populations of different genetic backgrounds.

In the second work presented in this paper we investigated the evolutionary history of the Roma (Results: 2. The shaping of immunological response through natural selection after migration: the case of the Roma). Our hypothesis was that by analyzing whole-genome sequences of Roma individuals along with individuals from their host population, Romania, and from their source population, North India, we would be able to detect selective events in the Roma from before and after they left India. The main benefit of whole-genome sequencing data is that our scans of selection are not restricted to *a priori* selected loci. This permitted us to detect adaptive selection in regulatory regions associated with cytokine production. We validated these results by detecting pathways enriched in signals of selection and with an immunological function and then, performing functional experiments trying to identify which pathogens drove the adaptation affecting cytokine expression.

In the third work presented, we selected a list of candidate genes related to a common biological function: substance identification and detoxification (Results: 3. Is there adaptation in the human

genome for taste perception and phase I biotransformation). To analyze the evolution of these genes, we used the results of three publicly available scans of selection: the 1000 Genomes Selection Browser 1.0 (Pybus et al., 2014), the Hierarchical Boosting data (Pybus et al., 2015), and the Human population genomics browser PopHuman (Mulet et al., 2018). The main advantage of using several neutrality statistics and tests of selection calculated on genome-wide data is the ability to detect the signature of positive selection underlying the signal observed. Besides, the Hierarchical Boosting is a composite method that combines several selection tests and that is robust to the confounding effects of population-specific demography. This allowed us to discover that genes related to taste and phase I biotransformation in humans followed different evolutionary trajectories. While it is clear that positive selection acted in the cytochrome P450 system after the out of Africa, it appears that genetic drift has been the main force causing the genetic variability that we observe nowadays in taste receptors genes.

In the last work presented, we took a step forward and added the information provided by the interconnections among genes to the study of natural selection (Results: 4. Influence of network topology on the evolution of metabolic enzymes in humans and mammals). We transformed the human metabolic network into a reaction graph, then, by mapping the events of positive selection onto the structure of the network, we could analyze how the connections between the enzymes affect the distribution of the selective events. Results of this study show the great explanatory power of the topology of the network to explain part of the variation in the distribution of signals of selection through the metabolic network. For instance, purifying selection is stronger in genes catalyzing the last steps in metabolic pathways and, when looking at the strength of recent positive selection in human populations genes at the bottom of metabolic pathways have higher positive selection values than those nodes participating in top steps. One of the drawbacks of these analyses is that the sample size (number of genes analyzed) gives the impression that the correlation between two measures is very strong (highly significant). However, except in few cases, the effect size of the relationships is small (very low coefficient of determination: r^2). Still, despite the difficult interpretation of some of the results, this study supports the idea that the structure of the network influences

and constrains the distribution of the selective events at different evolutionary times, as found in the human protein-protein interaction network (Luisi et al., 2015).

This are exciting times to study human evolution, as the cost of sequencing technologies has lowered enough to allow the generation of data from different populations and big sample sizes to allow robust statistical analyses. Also, the sequencing of previous and newly discovered fossils of archaic hominins is painting a complex picture of the evolution of our species that will need the development of new and more sophisticated demographic models and selection tests to account for this complexity.

Bibliography

Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. a, Durbin, R. M., Handsaker, R. E., … McVean, G. a. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*(7422), 56–65. https://doi.org/10.1038/nature11632

Abi-Rached, L., Jobin, M. J., Kulkarni, S., McWhinnie, A., Dalva, K., Gragert, L., … Parham, P. (2011). The Shaping of Modern Human Immune Systems by Multiregional Admixture with Archaic Humans. *Science (New York, N.Y.)*, *334*(6052), 89–94. https://doi.org/10.1126/science.1209202

Achim, V. (1998). *The Roma in Romanian History*. Central European University Press. Retrieved from http://books.openedition.org/ceup/1532

Ackermann, R. R., Mackay, A., & Arnold, M. L. (2016). The Hybrid Origin of "Modern" Humans. *Evolutionary Biology*, *43*(1), 1–11. https://doi.org/10.1007/s11692-015-9348-1

Alvarez-Ponce, D., Aguade, M., & Rozas, J. (2008). Network-level molecular evolutionary analysis of the insulin/TOR signal transduction pathway across 12 Drosophila genomes. *Genome Research*, *19*(2), 234–242. https://doi.org/10.1101/gr.084038.108

Alvarez-Ponce, D., Aguadé, M., & Rozas, J. (2011). Comparative Genomics of the Vertebrate Insulin/TOR Signal Transduction Pathway: A Network-Level Analysis of Selective Pressures. *Genome Biology and Evolution*, *3*(1), 87–101. https://doi.org/10.1093/gbe/evq084

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., … Schloss, J. A. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Barreiro, L. B., & Quintana-Murci, L. (2010). From evolutionary genetics to human immunology: How selection shapes host defence genes. *Nature Reviews Genetics*, *11*(1), 17–30. https://doi.org/10.1038/nrg2698

Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*, *169*(7), 1177–1186. https://doi.org/10.1016/j.cell.2017.05.038

Brunk, E., Sahoo, S., Zielinski, D. C., Altunkaya, A., Dräger, A., Mih, N., … Palsson, B. O. (2018). Recon3D enables a three-dimensional view of gene variation in human metabolism. *Nature Biotechnology*, *36*(3), 272. https://doi.org/10.1038/nbt.4072

Chakraborty, S., & Alvarez-Ponce, D. (2016). Positive Selection and Centrality in the Yeast and Fly Protein-Protein Interaction Networks. *BioMed Research International*, *2016*, 1–12. https://doi.org/10.1155/2016/4658506

Chen, H., Patterson, N., & Reich, D. (2010). Population differentiation as a test for selective sweeps. *Genome Research*, *20*(3), 1–10.

https://doi.org/10.1101/gr.100545.109.3
Clark, N. L., Alani, E., & Aquadro, C. F. (2012). Evolutionary rate covariation reveals shared functionality and coexpression of genes. *Genome Research*, *22*(4), 714–720. https://doi.org/10.1101/gr.132647.111
Consortium, †The International HapMap, Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., … Tanaka, T. (2003). The International HapMap Project. *Nature*, *426*, 789. Retrieved from http://dx.doi.org/10.1038/nature02168
Consortium, T. I. H., Frazer (Principal Investigator), K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., … Stewart, J. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, *449*, 851. Retrieved from http://dx.doi.org/10.1038/nature06258
Cortes, A., & Brown, M. A. (2011). Promise and pitfalls of the Immunochip. *Arthritis Research & Therapy*, *13*(1), 101. https://doi.org/10.1186/ar3204
Dall'Olio, G. M., Laayouni, H., Luisi, P., Sikora, M., Montanucci, L., & Bertranpetit, J. (2012). Distribution of events of positive selection and population differentiation in a metabolic pathway: the case of asparagine N-glycosylation. *BMC Evolutionary Biology*, *12*, 98. https://doi.org/10.1186/1471-2148-12-98
Delaneau, O., & Marchini, J. (2014). Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nature Communications*, *5*, 3934. https://doi.org/10.1038/ncomms4934
Fan, S., Hansen, M. E. B., Lo, Y., & Tishkoff, S. A. (2016). Going global by adapting local: A review of recent human adaptation. *Science (New York, N.Y.)*. https://doi.org/10.1126/science.aaf5098
Fisher, R. A. (1930). *The Genetical Theory Of Natural Selection*. At The Clarendon Press.
Flowers, J. M., Sezgin, E., Kumagai, S., Duvernell, D. D., Matzkin, L. M., Schmidt, P. S., & Eanes, W. F. (2007). Adaptive Evolution of Metabolic Pathways in Drosophila. *Molecular Biology and Evolution*, *24*(6), 1347–1354. Retrieved from http://dx.doi.org/10.1093/molbev/msm057
Fraser, A. (1992). *The Gypsies*. Blackwell Publishers.
Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C., & Feldman, M. W. (2002). Evolutionary rate in the protein interaction network. *Science (New York, N.Y.)*, *296*(5568), 750–752. https://doi.org/10.1126/science.1068696
Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., … Pääbo, S. (2010). A Draft Sequence of the Neandertal Genome. *Science*, *328*(5979), 710 LP-722. Retrieved from http://science.sciencemag.org/content/328/5979/710.abstract
Greenberg, A. J., Stockwell, S. R., & Clark, A. G. (2008). Evolutionary constraint and adaptation in the metabolic network of Drosophila. *Molecular Biology and Evolution*, *25*(12), 2537–2546. https://doi.org/10.1093/molbev/msn205

Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., … Sandhu, M. S. (2014). The African Genome Variation Project shapes medical genetics in Africa. *Nature*. https://doi.org/10.1038/nature13997

Hellenthal, G., Busby, G. B. J., Band, G., Wilson, J. F., Capelli, C., Falush, D., & Myers, S. (2014). A Genetic Atlas of Human Admixture History. *Science*, *343*(6172), 747 LP-751. Retrieved from http://science.sciencemag.org/content/343/6172/747.abstract

Henn, B. M., Cavalli-Sforza, L. L., & Feldman, M. W. (2012). The great human expansion. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(44), 17758–17764. https://doi.org/10.1073/pnas.1212380109

Hollfelder, N., Schlebusch, C. M., Günther, T., Babiker, H., Hassan, H. Y., & Jakobsson, M. (2017). Northeast African genomic variation shaped by the continuity of indigenous groups and Eurasian migrations. *PLOS Genetics*, *13*(8), e1006976. https://doi.org/10.1371/journal.pgen.1006976

Hudson, C. M., & Conant, G. C. (2011). Expression level, cellular compartment and metabolic network position all influence the average selective constraint on mammalian enzymes. *BMC Evolutionary Biology*, *11*(1), 89. https://doi.org/10.1186/1471-2148-11-89

Huerta-Sánchez, E., DeGiorgio, M., Pagani, L., Tarekegn, A., Ekong, R., Antao, T., … Nielsen, R. (2013). Genetic Signatures Reveal High-Altitude Adaptation in a Set of Ethiopian Populations. *Molecular Biology and Evolution*, *30*(8), 1877–1888. https://doi.org/10.1093/molbev/mst089

Invergo, B. M., Montanucci, L., Laayouni, H., & Bertranpetit, J. (2013). A system-level, molecular evolutionary analysis of mammalian phototransduction. *BMC Evolutionary Biology*, *13*(1), 52. https://doi.org/10.1186/1471-2148-13-52

Kim, P. M., Korbel, J. O., & Gerstein, M. B. (2007). Positive selection at the protein network periphery: Evaluation in terms of structural constraints and cellular context. *Proceedings of the National Academy of Sciences*.

Lachance, J., & Tishkoff, S. a. (2013a). Population Genomics of Human Adaptation. *Annual Review of Ecology, Evolution, and Systematics*, *44*(1), 123–143. https://doi.org/10.1146/annurev-ecolsys-110512-135833

Lachance, J., & Tishkoff, S. a. (2013b). SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. *BioEssays*, *35*(9), 780–786. https://doi.org/10.1002/bies.201300014

Livingstone, K., & Anderson, S. (2009). Patterns of Variation in the Evolution of Carotenoid Biosynthetic Pathway Enzymes of Higher Plants. *Journal of Heredity*, *100*(6), 754–761. Retrieved from http://dx.doi.org/10.1093/jhered/esp026

Lovell, S. C., & Robertson, D. L. (2010). An Integrated View of Molecular Coevolution in Protein–Protein Interactions. *Molecular Biology and*

Evolution, *27*(11), 2567–2575. Retrieved from http://dx.doi.org/10.1093/molbev/msq144

Luisi, P., Alvarez-Ponce, D., Dall'Olio, G. M., Sikora, M., Bertranpetit, J., & Laayouni, H. (2012). Network-level and population genetics analysis of the insulin/TOR signal transduction pathway across human populations. *Molecular Biology and Evolution*, *29*(5), 1379–1392. https://doi.org/10.1093/molbev/msr298

Luisi, P., Alvarez-Ponce, D., Pybus, M., Fares, M. A., Bertranpetit, J., & Laayouni, H. (2015). Recent positive selection has acted on genes encoding proteins with more interactions within the whole human interactome. *Genome Biology and Evolution*, *7*(4), 1141–1154. https://doi.org/10.1093/gbe/evv055

Marigorta, U. M., & Navarro, A. (2013). High Trans-ethnic Replicability of GWAS Results Implies Common Causal Variants. *PLOS Genetics*, *9*(6), e1003566. https://doi.org/10.1371/journal.pgen.1003566

Mendizabal, I., Lao, O., & UM, M. (2012). Reconstructing the population history of European Romani from genome-wide data. *Curr Biol*, *22*, 2342–2349. Retrieved from http://dx.doi.org/10.1016/j.cub.2012.10.039

Mendizabal, I., Valente, C., & Gusmao, A. (2011). Reconstructing the Indian origin and dispersal of the European Roma: a maternal genetic perspective. *PLoS One*, *6*, e15988. Retrieved from http://dx.doi.org/10.1371/journal.pone.0015988

Montanucci, L., Laayouni, H., Dall'Olio, G. M., & Bertranpetit, J. (2011). Molecular Evolution and Network-Level Analysis of the N-Glycosylation Metabolic Pathway Across Primates. *Molecular Biology and Evolution*, *28*(1), 813–823. https://doi.org/10.1093/molbev/msq259

Montanucci, L., Laayouni, H., Dobón, B., Keys, K. L., Bertranpetit, J., & Peretó, J. (2018). Influence of pathway topology and functional class on the molecular evolution of human metabolic genes. *BioRxiv*. https://doi.org/doi.org/10.1101/292714

Mulet, R., Villegas-mir, P., Velasco, D., Bertranpetit, J., Laayouni, H., Barbadilla, A., … Barbadilla, A. (2018). PopHuman: The human population genomics browser. *Nucleic Acids Research*, *46*(D1), D1003–D1010. https://doi.org/10.1093/nar/gkx943

Nielsen, R., Akey, J. M., Jakobsson, M., Pritchard, J. K., Tishkoff, S., & Willerslev, E. (2017). Tracing the peopling of the world through genomics. *Nature*. https://doi.org/10.1038/nature21347

Pybus, M., Dall'Olio, G. M., Luisi, P., Uzkudun, M., Carreño-Torres, A., Pavlidis, P., … Engelken, J. (2014). 1000 Genomes Selection Browser 1.0: A genome browser dedicated to signatures of natural selection in modern humans. *Nucleic Acids Research*, *42*(D1), 903–909. https://doi.org/10.1093/nar/gkt1188

Pybus, M., Luisi, P., Dall'Olio, G. M., Uzkudun, M., Laayouni, H., Bertranpetit, J., & Engelken, J. (2015). Hierarchical boosting: A machine-learning framework to detect and classify hard selective sweeps in human populations. *Bioinformatics*, *31*(24), 3946–3952. https://doi.org/10.1093/bioinformatics/btv493

Qian, W., Zhou, H., & Tang, K. (2015). Recent Coselection in Human Populations Revealed by Protein–Protein Interaction Network. *Genome Biology and Evolution*, *7*(1), 136–153. https://doi.org/10.1093/gbe/evu270

Racimo, F., Sankararaman, S., Nielsen, R., & Huerta-Sánchez, E. (2015). Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics*, *16*, 359. Retrieved from http://dx.doi.org/10.1038/nrg3936

Rausher, M. D., Miller, R. E., & Tiffin, P. (1999). Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. *Molecular Biology and Evolution*, *16*(2), 266–274. https://doi.org/10.1093/oxfordjournals.molbev.a026108

Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., … Pääbo, S. (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, *468*, 1053. Retrieved from http://dx.doi.org/10.1038/nature09710

Romero, P., Wagg, J., Green, M. L., Kaiser, D., Krummenacker, M., & Karp, P. D. (2005). Computational prediction of human metabolic pathways from the complete human genome. *Genome Biology*, *6*(1), R2. https://doi.org/10.1186/gb-2004-6-1-r2

Sabeti, P. C., Schaffner, S. F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., … Lander, E. S. (2006). Positive natural selection in the human lineage. *Science (New York, N.Y.)*, *312*(5780), 1614–1620. https://doi.org/10.1126/science.1124309

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., … Hapmap, T. I. (2007). Genome-wide detection and characterization of positive selection in human populations. *October*, *449*(7164), 913–918. https://doi.org/10.1038/nature06250.Genome-wide

Scerri, E. M. L., Thomas, M. G., Manica, A., Gunz, P., Stock, J. T., Stringer, C., … Chikhi, L. (2018). Did Our Species Evolve in Subdivided Populations across Africa, and Why Does It Matter? *Trends in Ecology & Evolution*, *33*(8), 582–594. https://doi.org/10.1016/j.tree.2018.05.005

Slon, V., Mafessoni, F., Vernot, B., de Filippo, C., Grote, S., Viola, B., … Pääbo, S. (2018). The genome of the offspring of a Neanderthal mother and a Denisovan father. *Nature*, *561*(7721), 113–116. https://doi.org/10.1038/s41586-018-0455-x

Smith, J. M., & Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical Research*, *23*(1), 23–35. https://doi.org/DOI: 10.1017/S0016672300014634

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*(3), 585 LP-595. Retrieved from http://www.genetics.org/content/123/3/585.abstract

Tenaillon, O. (2014). The Utility of Fisher's Geometric Model in Evolutionary Genetics. *Annual Review of Ecology, Evolution, and Systematics*, *45*(1), 179–201. https://doi.org/10.1146/annurev-ecolsys-120213-091846

Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A.,

Froment, A., … Williams, S. M. (2009). The genetic structure and history of Africans and African Americans. *Science (New York, N.Y.)*, *324*(5930), 1035–1044. https://doi.org/10.1126/science.1172257

Trynka, G., Hunt, K. a, Bockett, N. a N., Romanos, J., Mistry, V., Szperl, A., … van Heel, D. a. (2011). Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nature Genetics*, *43*(12), 1193–1201. https://doi.org/10.1038/ng.998

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics*, *101*(1), 5–22. https://doi.org/https://doi.org/10.1016/j.ajhg.2017.06.005

Vitkup, D., Kharchenko, P., & Wagner, A. (2006). Influence of metabolic network structure and function on enzyme evolution. *Genome Biology*, *7*(5). https://doi.org/10.1186/gb-2006-7-5-r39

Vitti, J. J., Grossman, S. R., & Sabeti, P. C. (2013). Detecting Natural Selection in Genomic Data. *Annual Review of Genetics*, *47*(1), 97–120. https://doi.org/10.1146/annurev-genet-111212-133526

Weir, B. S., & Hill, W. G. (2002). Estimating F-statistics. *Annual Review of Genetics*, *36*, 721–750. https://doi.org/doi: 10.1146/annurev.genet.36 050802.093940

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591. https://doi.org/10.1093/molbev/msm088

Zeng, K., Fu, Y. X., Shi, S., & Wu, C. I. (2006). Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics*, *174*(November), 1431–1439. https://doi.org/10.1534/genetics.106.061432

Appendix

1. List of publications

Montanucci, L., Laayouni, H., **Dobon, B**., Keys, K. L., Bertranpetit, J., & Peretó, J. (2018). Influence of pathway topology and functional class on the molecular evolution of human metabolic genes. BioRxiv. https://doi.org/doi.org/10.1101/292714

Torruella-Loran, I., Laayouni, H., **Dobon, B**., Gallego, A., Balcells, I., Garcia-Ramallo, E., & Espinosa-Parrilla, Y. (2016). MicroRNA Genetic Variation: From Population Analysis to Functional Implications of Three Allele Variants Associated with Cancer. Human Mutation, 37(10), 1060–1073. https://doi.org/10.1002/humu.23045

Dobon, B., Hassan, H. Y., Laayouni, H., Luisi, P., Ricaño-Ponce, I., Zhernakova, A., … Bertranpetit, J. (2015). The genetics of East African populations: A Nilo-Saharan component in the African genetic landscape. Scientific Reports, 5(November 2014), 9996. https://doi.org/10.1038/srep09996

2. List of manuscripts in preparation

Dobon B, ter Horst R., Laayouni H., Mondal M., Bosch E., Bertranpetit J. & Netea MG. The shaping of immunological response through natural selection after migration: the case of the Roma.

Dobon B., Montanucci L., Peretó J., Bertranpetit J. & Laayouni H. Influence of network topology on the evolution of metabolic enzymes in humans and mammals.

Dobon B., Rossell C., Walsh S. & Bertranpetit J. Is there adaptation in the human genome for taste perception and phase I biotransformation? (*Submitted*)

3. Supplementary materials

3.1. The shaping of immunological response through natural selection after migration: the case of the Roma

Supplementary Note 1. Data preprocessing

Samples

We generated whole genome sequences of 50 Romani and 50 Romanian individuals from Romania. To study the ancestry of Romani people and avoid any ascertainment bias in the variant calling (process described in Raw Sequence Processing and Mapping) we included whole genome sequences of a set of Indian populations that reflect the maximum genetic diversity on continental India described in (Mondal et al., 2016): 10 Uttar Pradesh Upper Caste Brahmins (UBR), 10 Rajput (RAJ), 10 Irula (ILA), 10 Birhor (BIR), nine Riang (RIA), nine Vellalar (VLR), one Punjabi (PUN), one Bengali (BEN), six Onge (ONG) and four Jarawa (JAR). We also included one individual from each of the following populations to have a worldwide representation of human genetic diversity: French (FRN), Sardinian (SAR), Dai (DAI), Han Chinese (HAN), Mandenka (MAD), Mbuti (MBT), Papuan (PAP), San (SAN), and Yoruba (YRI) (Meyer *et al*., 2012).

Library preparations, sequencing, and base calling

DNA was extracted from blood samples and were sequenced at the Beijing Genomics Institute (BGI; Beijing, China). For every sample, 1 μg of genomic DNA was sheared into short fragments on Covaris E210 system (CovarisInc). The overhang at the ends of DNA fragments were converted into blunt ends by T4 DNA polymerase and Klenow enzyme. After ligation with adapters on both ends, DNA fragments of ~500 bp were selected by agarose gel electrophoresis and purified. Polymerase Chain Reaction (PCR) was performed to obtain sufficient DNA for a sequencing library. The quality of the library was checked by agarose gel electrophoresis. Sequencing was performed on Illumina HiSeq 2000 to produce paired-end reads of 90 bp. Base calling was completed following the manufacturer's base-calling pipeline.

Raw Sequence Processing and Mapping

Fastq conversion, mapping and BAM processing was performed following the procedure described in the Supplementary material of Mondal et al., (2016).

- **Fastq Conversion**: Sequences from BGI were in Illumina 1.5+ FASTQ format. All the BGI FASTQ files were converted to Illumina 1.8+ using seqtk () with the -VQ64 flag (./seqtkseq -VQ64). FASTQ files from human populations from Meyer *et. al* (2012) were downloaded and converted to Illumina 1.8+ format using seqtk. The following steps (Mapping, BAM processing and Variant Calling) were applied to all sequences of Romani and Romanian populations, along with sequences from Indian and worldwide datasets.

- **Mapping:** All sequences, in Illumina 1.8+ FASTQ format, were mapped using BWA (Li, 2013). Hg19 was used as a reference and mapped using the BWA mem algorithm. Only paired-end reads were kept. The BWA -w 50 flag was used to give the size of the band width. BWA output was then converted to binary format (bam) using SAMtools (version 0.1.18) (Li et al., 2009) and sorted using SortSam from Picard tools (version 1.100).

- **BAM Processing:** Bam processing was completed by following the "Best Practices" recommendations in GATK (version 3.5) (McKenna et al., 2010). After converting the mapped files to the binary format, CleanSam from Picard tools (http://picard.sourceforge.net) was used to remove unmapped sequences, and MarkDuplicates to mark duplicates. The bam files were then indexed using SAMtools. Since indels can cause inaccurate mapping in the genome, IndelRealigner from GATK was used to realign them, with 1000 Genomes Project phase 1 Indel as a reference file (interval file) (Abecasis et al., 2012). BaseRecalibrator and PrintReads were used from GATK to calibrate bases for various statistics (i.e. reported quality score, machine cycle, positions of the SNP in the read, etc.) for SNPs not present in dbSNP version 137 (Sherry, 2001).

MergeSamFiles from Picard tools was used to merge lanes for the same individuals before variant calling by GATK.

Variant Calling

Variant calling for Romani and Romanian sequences, along with Indian and worldwide sequences, were done by GATK. Per-sample calling was done with default options and using -- max-alternate_alleles 20 (to capture all genetic diversity present in the populations) by running HaplotypeCaller in GVCF mode on each sample's BAM file. Then, the joint genotyping of the gVCFs produced was done by running GenotypeGVCFs on all of them together to generate a raw SNP and indel Variant Calling File (VCF).

VCF Recalibration

The raw VCF was filtered using post variant calling recalibration steps as listed in GATK "Best Practices". VariantRecalibration and ApplyRecalibration from GATK were used to calculate various statistics for novel variants (both for SNPs and indels) and then recalibrated according to their needs. We applied the following steps:

SNPs with the flags -an QD -an MQRankSum -anReadPosRankSum -an FS -an DP -an InbreedingCoeff. All other parameters were set to default values:

- dbsnp version 137: -resource:dbsnp, known=true, training=false, truth=false, prior=2.0.
- hapmap version 3.3: -resource:hapmap, known=false, training=true, truth=true, prior=15.0 (International Hapmap3 Consortium 2010).
- Omni genotyping array 2.5 million 1000G: -resource:omni, known=false, training=true, truth=true, prior=12.0.
- 1000G phase 1 high confidence: -resource:1000G, known=false, training=true, truth=false, prior=10.0.

Indels with the flags -- maxGaussians 4 -an FS -an ReadPosRankSum -an MQRankSum -an DP -an InbreedingCoeff. All other parameters were set to default values:

- Mills 1000G high confidence indels: -resource:mills, known=false, training=true, truth=true, prior=12.0.
- dbSNP version 137: -resource:dbsnp, known=true, training=false, truth=false, prior=2.0.

Supplementary Note 2. Quality control

Depth of Coverage and Fraction Covered

Genome coverage for each sample was estimated by DepthOfCoverage from GATK to check for bias in the probability of calling non-reference alleles due to different coverage between samples. The average coverage for autosomal chromosomes ranged from 12X to 21X, with an average of 15X (Supplementary Figure 1).

Sex determination

We also estimated the coverage for the X and Y chromosomes to determine the genetic sex of the samples by DepthOfCoverage from GATK. We calculate the ratio of the coverage on the X and Y chromosomes with respect to the coverage on autosomal chromosomes. In females, we expect the ratio of the coverage on the X chromosome and the coverage on the autosomes to be around one; whereas in males, it should tend to 0.5 (males only have one copy of the X). In males, we expect the ratio of the coverage on the Y chromosome and the coverage on the autosomes to tend to 0.5, whereas it should be zero in females (males only have one copy of the Y, and females none).

We observed a sample with ambiguous sex determination, sample RMN-17 (Supplementary Figure 2). This can indicate contamination of the sample and was further analyzed in the Estimation of heterozygosity and mitochondrial contamination sections. Four Romanian samples were identified as female (RMN-7, RMN-12, RMN-14, and RMN-31), the rest were classified as male. All Romani samples were classified as male.

Estimation of autosomal heterozygosity (inbreeding)

The inbreeding coefficient (F) was calculated for each sample by VCFtools (version 0.1.14) (Danecek et al., 2011). Individuals showing an outlier value of heterozygosity or F could be the result of contamination. Sample RMN-17 showed an extremely low value of F compared to any other sample (Supplementary Figure 3) and was removed from the main analysis (Supplementary Table 1).

Estimation of heterozygosity in males (X chromosome)

We estimated contamination levels based on the level of X-chromosome heterozygosity in male samples with ANGSD (Korneliussen, Albrechtsen, & Nielsen, 2014). As requested by the software we used a list of polymorphic sites and their frequency for the following populations from 1000 Genomes Project: CEU (Utah Residents (CEPH) with Northern and Western European Ancestry), CHB (Han Chinese in Bejing, China), PEL (Peruvians from Lima, Peru), YRI (Yoruba in Ibadan, Nigeria), and GHI (Gujarati Indian from Houston, Texas). As a recommendation samples with X chromosome contamination estimates higher than 2.5% should be classified as contaminated. Only RMN-17 appears affected with a 33-43% of contamination (Supplementary Table 2) and was removed from the main analysis (Supplementary Table 1).

Estimation of mtDNA contamination

Estimation of mitochondrial genome contamination was done by Rpackage contamMix (version 1.0-10) (Fu et al., 2013; Johnson, 2014) to identify mitochondrial heteroplasmy. First, a mitochondrial consensus sequence was constructed for each sample with SAMtools mpileup (version 1.2) filtering for reads with excessive mismatches (-C 50), minimum mapping quality (-q 20) and minimum base quality for a base (-Q 20). Then, the mitochondrial reads were mapped against the mitochondrial consensus sequence using BWA with the -w 50 flag (size of the band width) and the output converted to bam format. Second, we generated a multiple sequence alignment with the consensus genome and the 311potential contaminant mitochondrial genomes provided in contamMix package using Muscle (version 3.8.31) (Edgar, 2004). With these two inputs, the program estimates the proportion of endogenous (authentic) mitochondrial genome present in the sample (P.AUT). A P.AUT of 0.80 means there is 20% of contamination. A sample was classified as possibly contaminated if the proportion of reads that have a better match with the consensus sequence generated than with any of the 311 mitochondrial sequence provided is less than 95%, or if the 95% confidence lower bound of that proportion is less than 85%. There are 5 samples showing more than 5% of contamination in the mitochondrial genome (Supplementary Figure 4) even though only RMN-17 showed any sign of contamination in the other analysis. This could be due to different amplification of autosomal and mitochondrial reads, and samples RMN-17, S19, S25, S43, and S60 were flagged as

contaminated. This analysis was repeated adding the mitochondrial consensus sequence of RMN-17 to the set of 311 potential contaminant mitochondrial genomes as a potential source of contamination with the same results (data not shown). Samples marked as contaminated were removed from the main analysis (Supplementary Table 1).

Transition versus Transversion Ration

The transition vs. transversion ratio (Ts/Tv) can indicate whether there are problems with the variant calling of the samples. VariantEval from GATK was used to calculate Ti/Tv. In humans, the expected Ts/Tv ratios in whole-genome sequencing is around 2-2.1, within the range of our results: Ts/Tv = 2.18 for known variants and Ts/Tv = 1.91 for novel variants (Supplementary Table 3).

Supplementary Note 3. Population analysis

Relatedness

Kinship analysis was performed using KING (Manichaikul et al., 2010) with only bi-allelic autosomal SNPs. We calculated the kinship score within the individuals of each population and between the individuals belonging to different populations. We did not find any individual related to one from another population, but we detected several 2^{nd} and 3^{rd} degree relations within populations (Supplementary Figure 5). We removed one individual from each of the 2^{nd} and 3^{rd} degree relations until we were left with only unrelated individuals (Supplementary Table 1).

Principal component analysis

We performed a Principal Component Analysis (PCA) using 50 Romanies and 50 Romanians to detect possible mislabeled individuals or some unforeseen bias. We converted the VCF file to PED and MAP formats using PLINK 1.9 (Chang et al., 2015) keeping only bi-allelic autosomal SNPs, filtering by Minor Allele Frequency (MAF) (--maf 0.05), without missing information (--geno 0) and under Hardy-Weinberg Equilibrium (--hwe 0.000001 midp). The resulting dataset (5,216,078) was pruned (--indep 50 5 2). PCA was performed with Eigensoft (version 6.1) (Patterson, Price, & Reich, 2006) in the remaining 515,723 SNPs (Supplementary Figure 5). The first principal component (PC1)

separates a tight cluster formed by most of the Romanian individuals from a more spread cluster of Romani individuals. We see several admixed individuals that do not clearly belong to one cluster or another; and a sample labeled as Romani clustered with Romanians (Supplementary Figure 6). Mislabeled samples and samples that could not be clearly assigned to a cluster were removed from the main analysis (Supplementary Table 1). After removing samples indicated in Supplementary Table 1, we were left with 40 Romanies and 40 Romanians. We merged these 80 samples with worldwide populations from 1000 Genomes Project Phase 3 (Auton et al., 2015): CEU (Utah Residents (CEPH) with Northern and Western European Ancestry), TSI (Toscani in Italia), FIN (Finnish in Finland), GBR (British in England and Scotland), IBS (Iberian Population in Spain), CHB (Han Chinese in Beijing, China), JPT (Japanese in Tokyo, Japan), CHS (Southern Han Chinese), CDX (Chinese Dai in Xishuangbanna, China), KHV (Kinh in Ho Chi Minh City, Vietnam), GIH (Gujarati Indian from Houston, Texas), PJL (Punjabi from Lahore, Pakistan), BEB (Bengali from Bangladesh), STU (Sri Lankan Tamil from the UK), ITU (Indian Telugu from the UK), YRI (Yoruba in Ibadan, Nigeria), LWK (Luhya in Webuye, Kenya), GWD (Gambian in Western Divisions in the Gambia), MSL (Mende in Sierra Leone), and ESN (Esan in Nigeria). From each population we randomly selected 40 unrelated individuals. We also added the following populations from continental India: 10 Uttar Pradesh Upper Caste Brahmins (UBR), 10 Rajput (RAJ), nine Vellalar (VLR), 10 Irula (ILA), nine Birhor (BIR), and 10 Riang (RIA) (Mondal et al. 2016). We applied the same filters as before resulting in a dataset of 938 individuals and 4,574,497 SNPs. We performed a PCA on the pruned dataset without the African populations (YRI, LWK, GWD, MSL, and ESN) (738 individuals and 467,592 SNPs). Romani are differentiated from the rest in PC3 whereas PC4 is created by two tribal Indian populations: BIR and ILA (Supplementary Figure 7).

Admixture analysis

To infer the ancestral populations of the Romani individuals, we run ADMIXTURE (version 1.3.0) (Alexander et al. 2009) in the pruned dataset of 738 worldwide individuals. We tested values of K from 2 to 9 with 5-fold cross-validation (Supplementary Figure 8a). Each run was run 25 times with different seeds and the run with the lowest CV was selected. The best supported model was K = 4

(Supplementary Figure 8b). As we are not including African populations, the first split is between Asian and European components (K = 2). In K = 3 appears an Indian component that is also seen in the Roma. Indian populations show both European and Asian components along with the Indian genetic component. The Roma populations show their own component in K = 4. In K = 5 we see the distinction between the Japanese (JPT) and the Han Chinese (CHB and CHS) from other Asian populations. Finnish (FIN) show their own component in K = 6. In K = 7 the Indian component is separated in tribal (ILA and BIR) and non-tribal. In K = 8, JPT separates from the other Asian populations. In K = 9, VLR show their own component. No admixture was detected in Romanies using the 3-Population test (Supplementary Figure 9). However, the f3-statistic will not detect the test population as admixed if after the admixture event the population has undergone strong population-specific drift (Patterson et al., 2012), as is the case with the Roma people that suffered a series of bottlenecks (Fraser, 1992).

Supplementary Note 4. Prioritization of candidate variants and genes

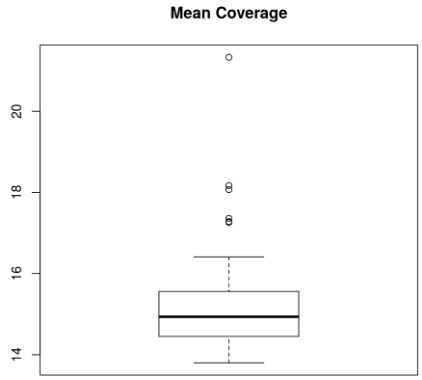
SNPs in regions that were inferred to be under positive selection were annotated with ANNOVAR (Wang, Li, & Hakonarson, 2010) in GRCh37 (hg19) using RefSeqGene, dbSNP 147, and CADD (Combined Annotation Dependent Depletion) version 13 (Kircher et al., 2014). We then identified highly differentiated variants linked to the inferred selection signals by comparing those populations that share a given selection signal with the population in which the signal is absent. That is, for Recent Shared Signals, the Derived Allele frequency (DAF) differences between Romanies and Rajput and between Romanians and Rajput were computed, and variants were subsequently identified as strongly differentiated when the corresponding average DAF difference to Rajput was greater than 0.25. Similarly, for Old Shared Signals the DAF differences between Romanies and Romanians and between Rajput and Romanians were computed and variants were identified as strongly differentiated when the corresponding average DAF difference to Romanians was greater than 0.25. Subsequently, those highly differentiated SNPs in both Recent and Old Shared Selection Signals that are either non-synonymous, annotated as cis-eQTLs in

the Genotype-Tissue Expression Project (Release V6p), present CADD values greater than 10 (meaning they are predicted to be among the 10% most deleterious in the human genome), or that appear clustered in exonic/splicing regions/ncRNAs/UTRs were classified as potential candidate variants for adaptation. In the Recent Shared Signals, 6,452 SNPs out of 28,640 markers are highly differentiated and of those, 293 were predicted to be among the 10% most deleterious changes in the human genome, 28 were non-synonymous changes (including 9 SNPs with CADD values ≥ 10), 18 implied synonymous changes, and one stopgain change (with a CADD value = 40). As for the Old Shared Signals, 1,296 SNPs out of 7,205 are highly differentiated and of those, 64 were predicted to be among the 10% most deleterious changes in the human genome, six cause synonymous changes and four imply non-synonymous changes (three of them with CADD values ≥ 15). Whereas in the Recent Shared Signals up to 16 candidate genes presented highly differentiated nonsynonymous or stop gain variants, in the Old Shared Signals we only detected 4 highly differentiated nonsynonymous changes in 4 candidate genes (Supplementary Table 4). Overall, three candidate genes related to the immune system presented highly differentiated non-synonymous SNPs with CADD values ≥ 10 in the recent (*ELF1*, *SETX*) and old (*DOCK8*) shared signals detected.

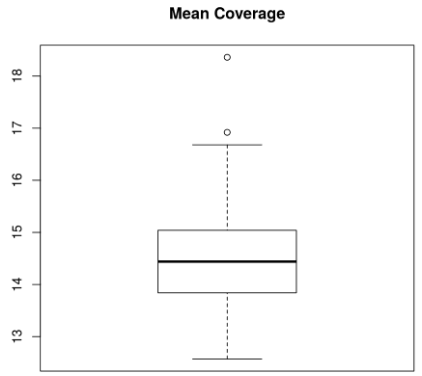
Supplementary Figures

Supplementary Figure 1. Distribution of the average coverage for autosomal chromosomes in a) Roma and b) Romanian samples. Fraction of the genome that is covered by at least X reads in c) Roma and d) Romanian samples.

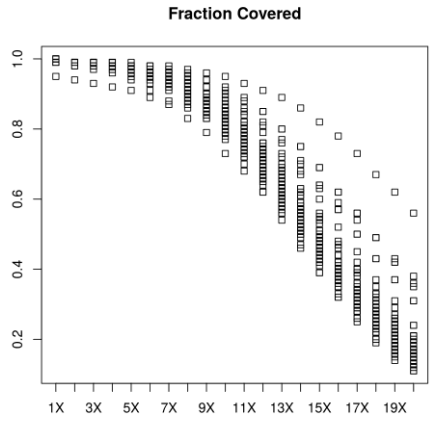
a)



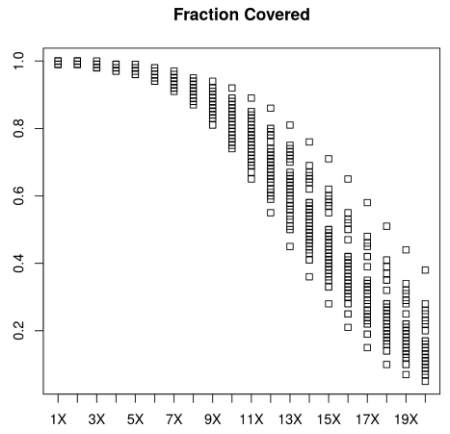
b)



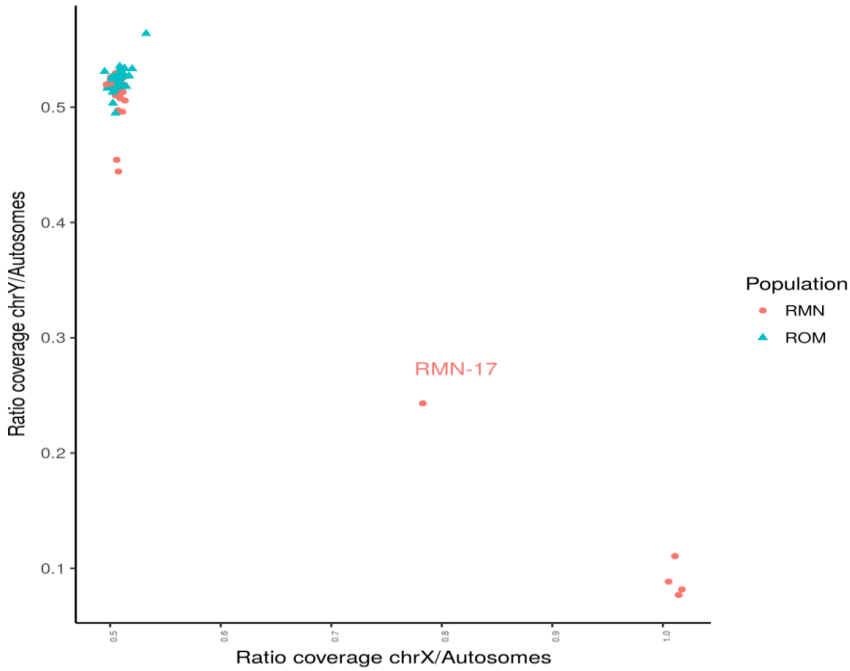
c)



d)

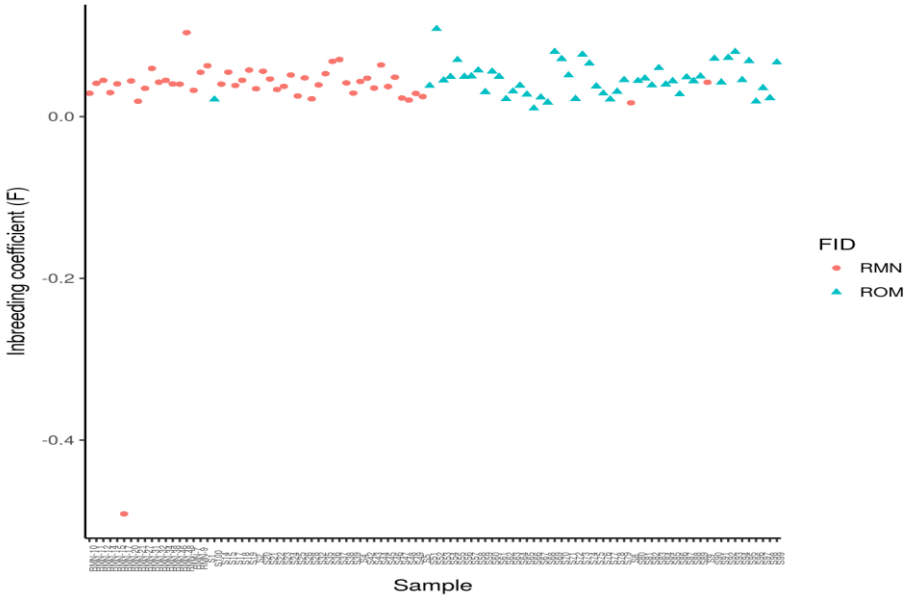


Supplementary Figure 2. Ratio of the coverage on the X and Y chromosomes with respect to the coverage on autosomal chromosomes for Romani and Romanian samples. Sample RMN-17 with ambiguous sex determination is labelled. For Romanian samples are classified as female: RMN-7, RMN-12, RMN-14, and RMN-31.

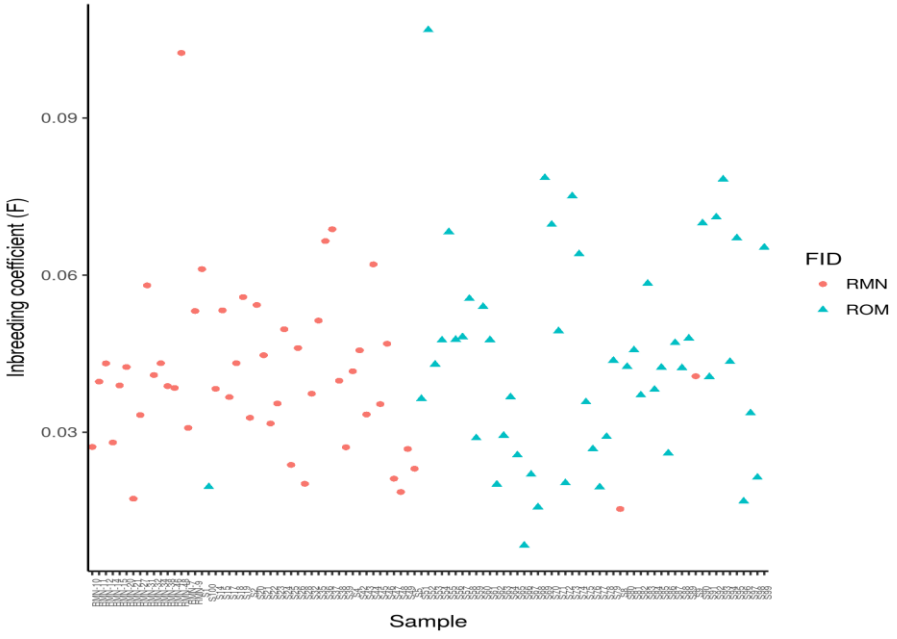


Supplementary Figure 3. a) Inbreeding coefficient (F) for Romani (ROM) and Romanian (RMN) samples. b) Same plot as a) after removing outlier sample RMN-17.

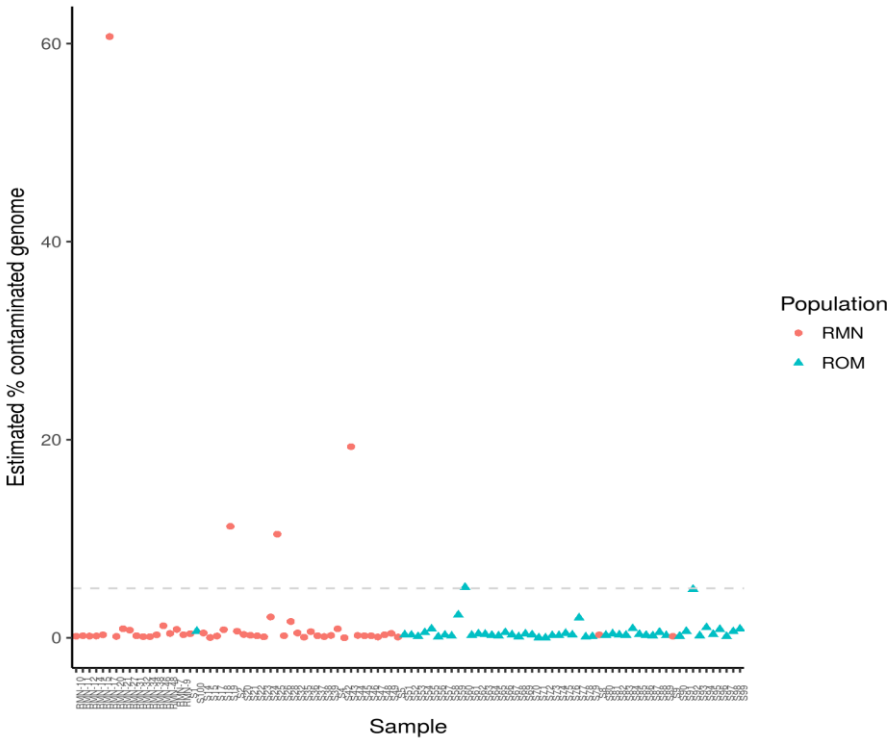
a)



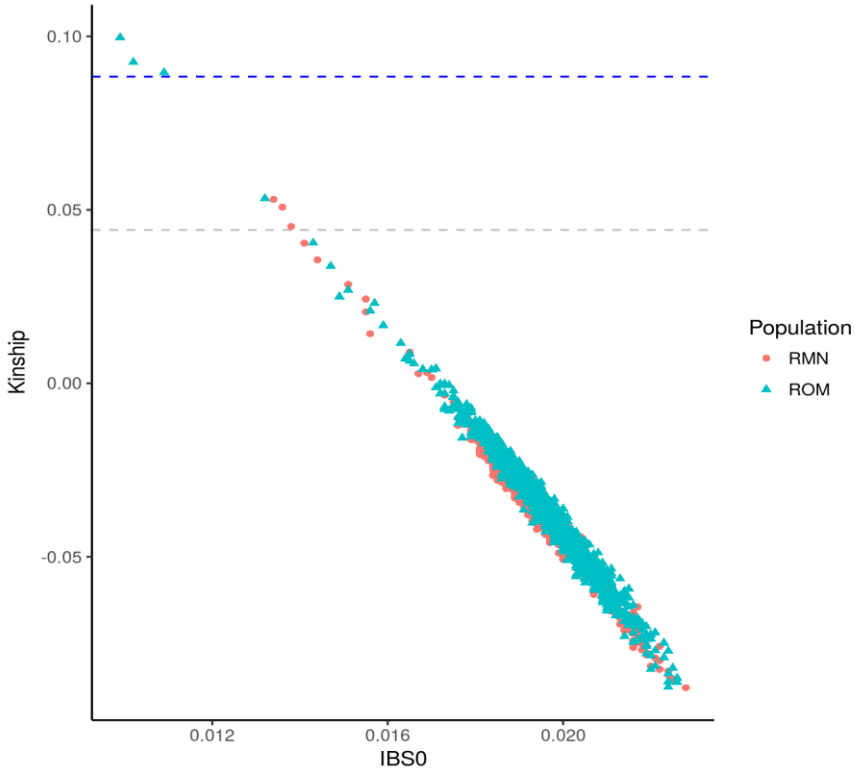
b)



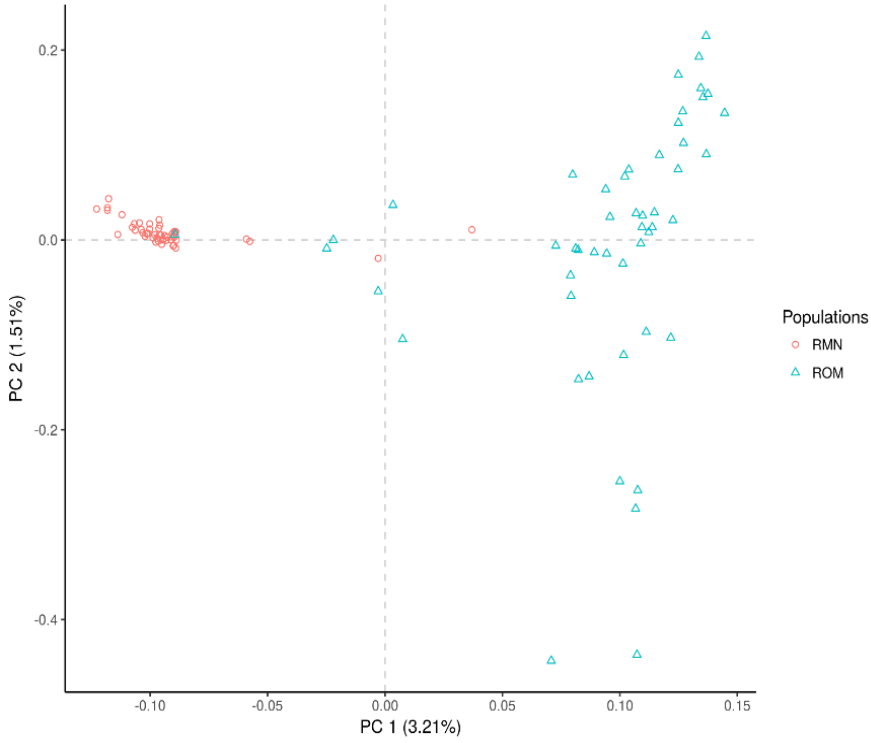
Supplementary Figure 4. Estimates of mtDNA contamination in Romani and Romanian samples. Samples with an estimated percentage of contaminant genome over 5% were classified as contaminated.



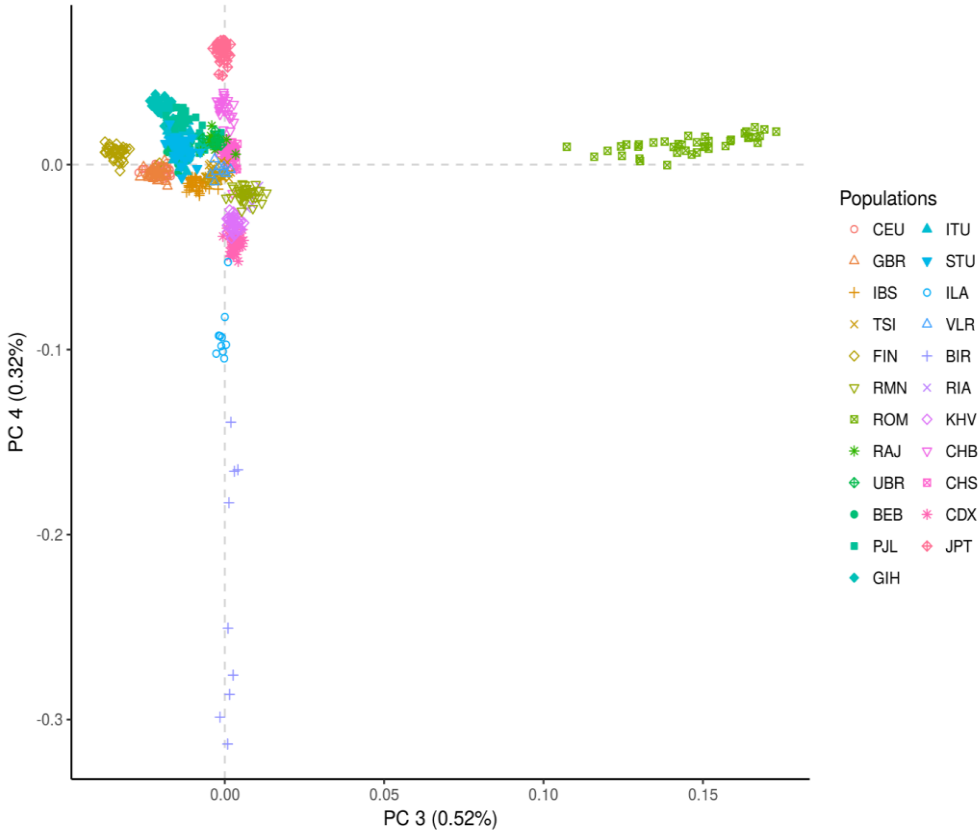
Supplementary Figure 5. Estimated kinship coefficient versus the proportion of SNPs with zero Identical-by-state (IBS0) in 50 Romanies (ROM) and 49 Romanian (RMN; without contaminated sample RMN-17). Blue dashed line indicates threshold of 2nd degree relation (Kinship range = [0.0884, 0.177]); grey dashed line indicates threshold of 3rd degree relation (Kinship range = [0.0442, 0.0884]).



Supplementary Figure 6. Principal component analysis of 50 Romani (ROM) and 50 Romanian samples (RMN). Principal component (PC) 1 and PC2. The percentage of variance explained by each component is added in the labels.

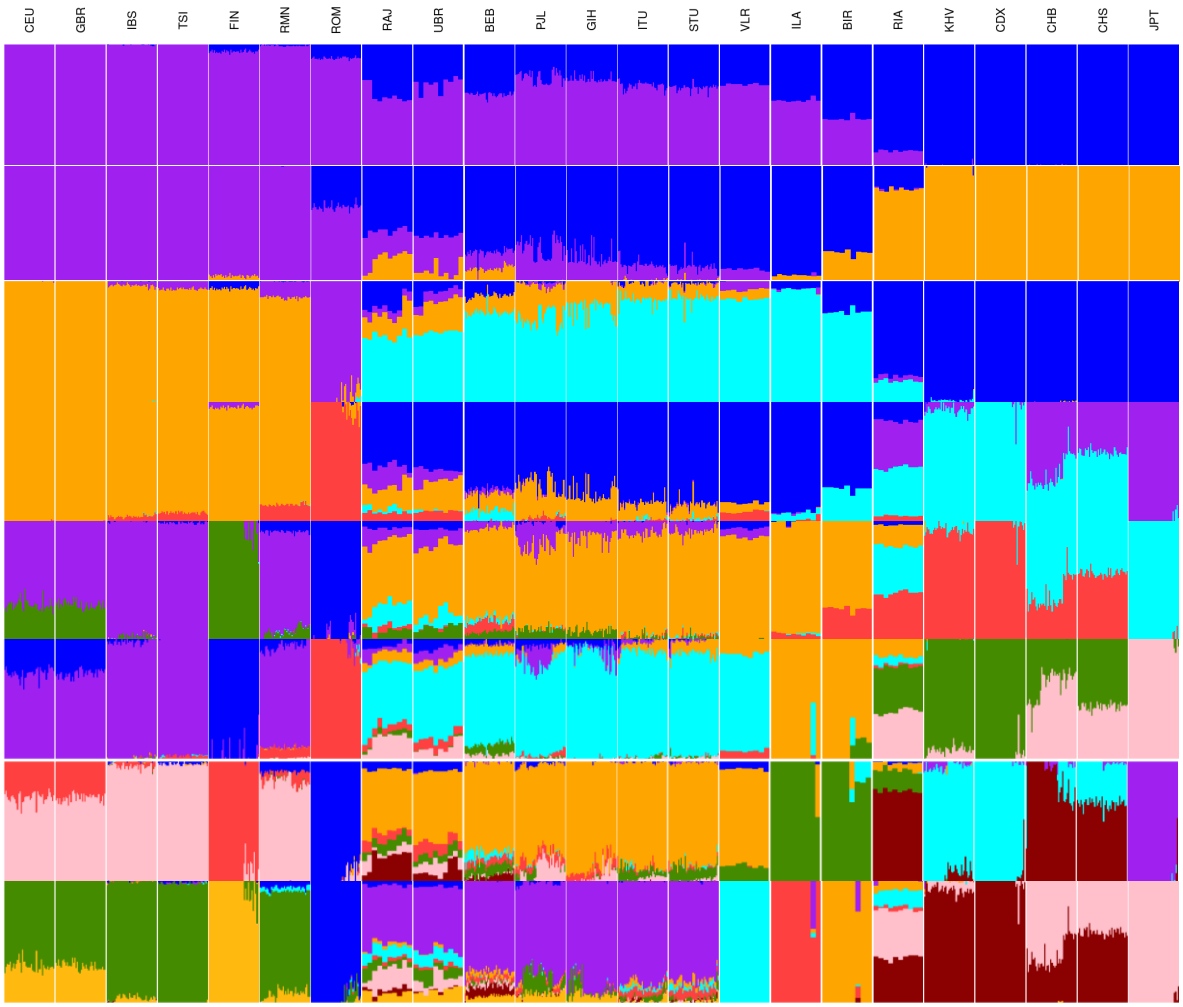


Supplementary Figure 7. Principal component analysis of 40 Romani (ROM) and 40 Romanian (RMN) samples with 1000 Genomes Project Phase 3 and mainland Indian populations from (Mondal et al., 2016). Principal component (PC) 1 and PC2 are shown in the main text. PC 3 and PC4. The percentage of variance explained by each component is added in the labels.

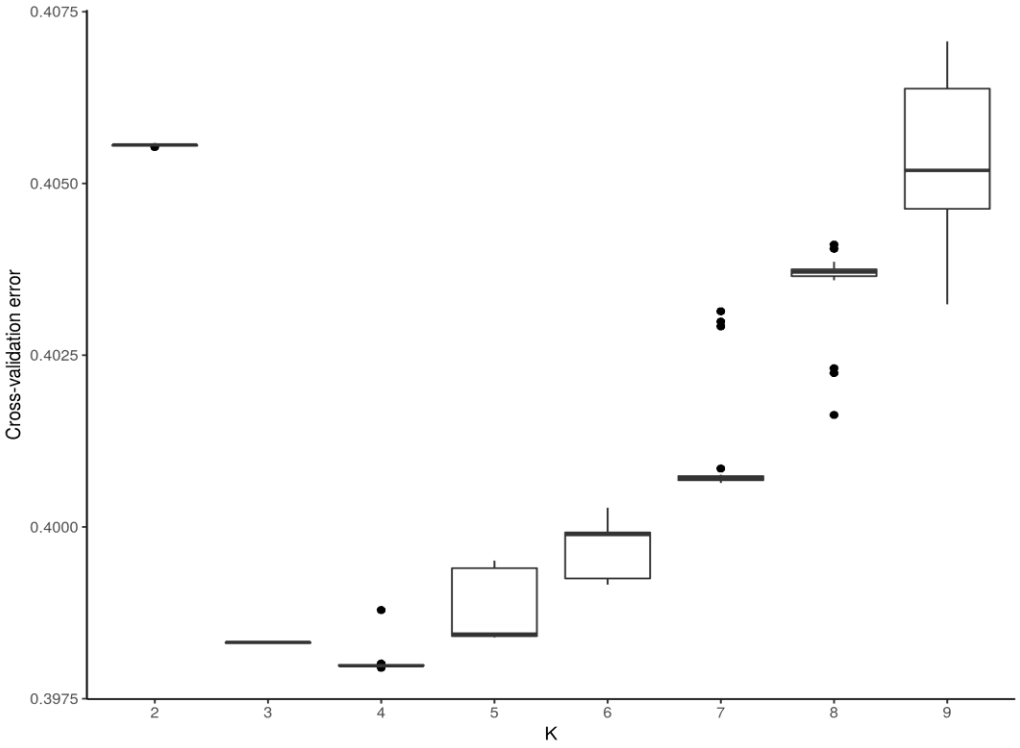


Supplementary Figure 8. a) Admixture plot of 40 Romani (ROM) and 40 Romanian (RMN) with 1000 Genomes Project Phase 3 and mainland Indian populations from (Mondal et al., 2016). a) Admixture plot showing runs of K from 2 to 9; b) Cross-validation error of runs with K values from 2 to 9.

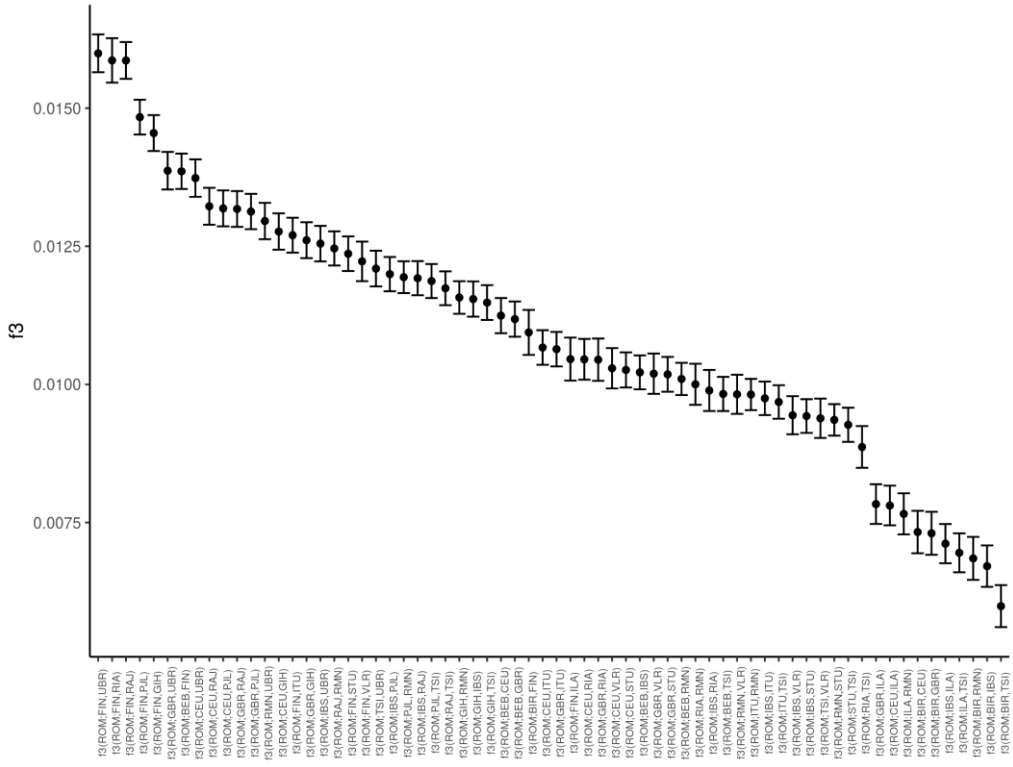
a)



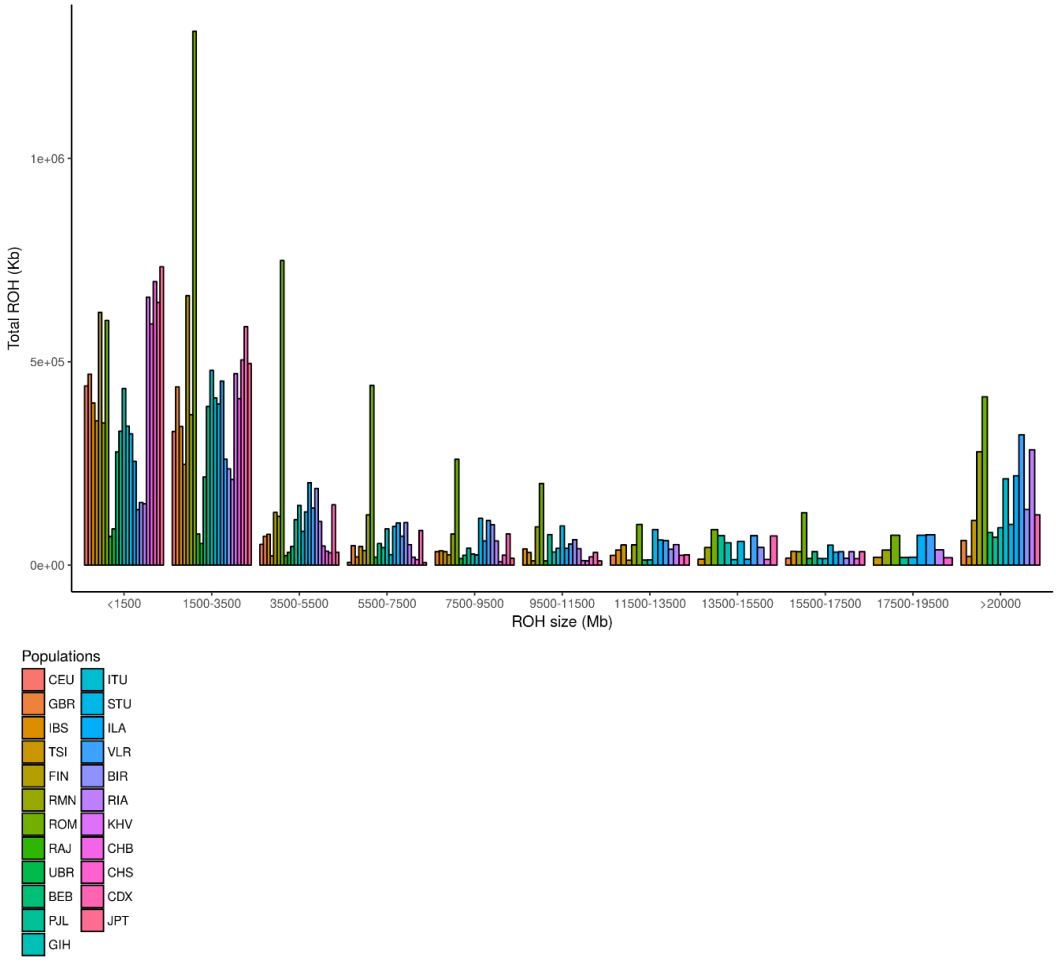
b)


0.4075
0.4050
0.4025
0.4000
0.3975
Cross-validation error
2
3
4
5
6
7
8
9
K

Supplementary Figure 9. F3-statistic in the form f3(Romani; European, Indian), where European and Indian are populations from the 1000 Genomes Project (Auton et al., 2015) and from (Mondal et al., 2016). The analysis was repeated indicating that the Romanies were an inbreed population (flag inbreed = YES), obtaining the same results (data not shown).

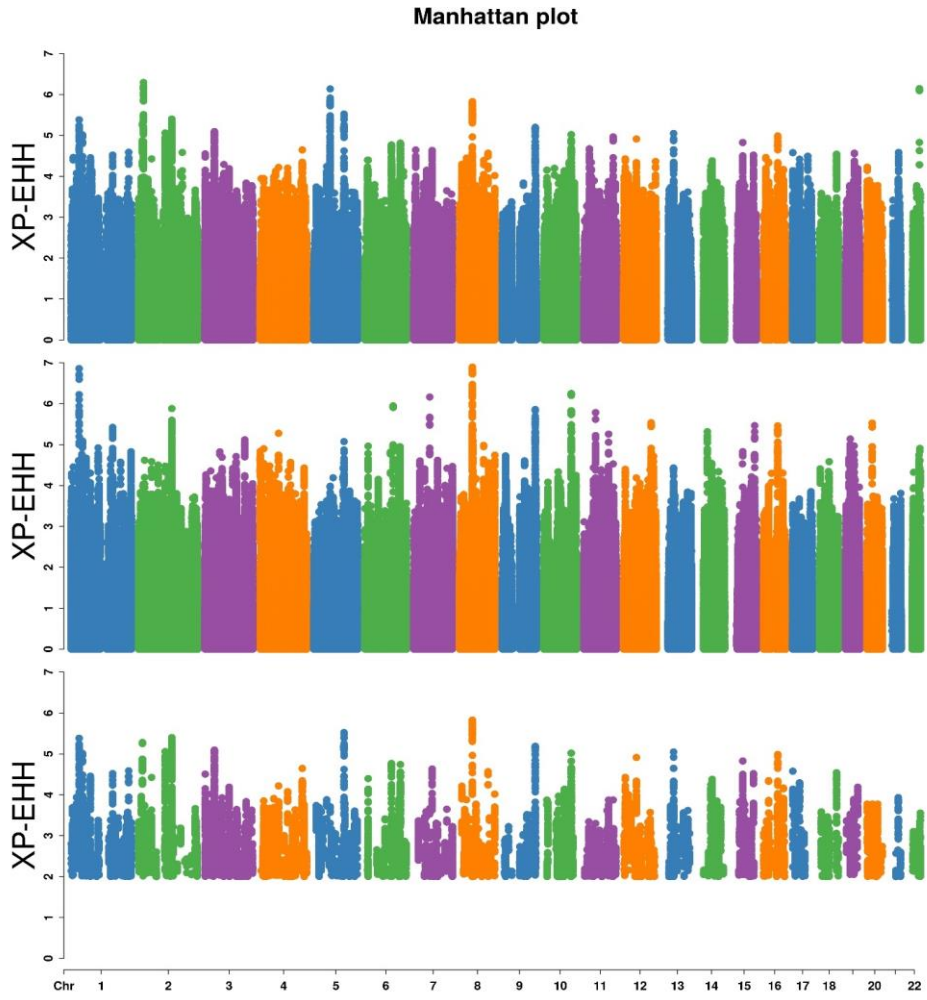


Supplementary Figure 10. Distribution of the total length of runs of homozygosity (ROHs) classified by length categories in worldwide populations.

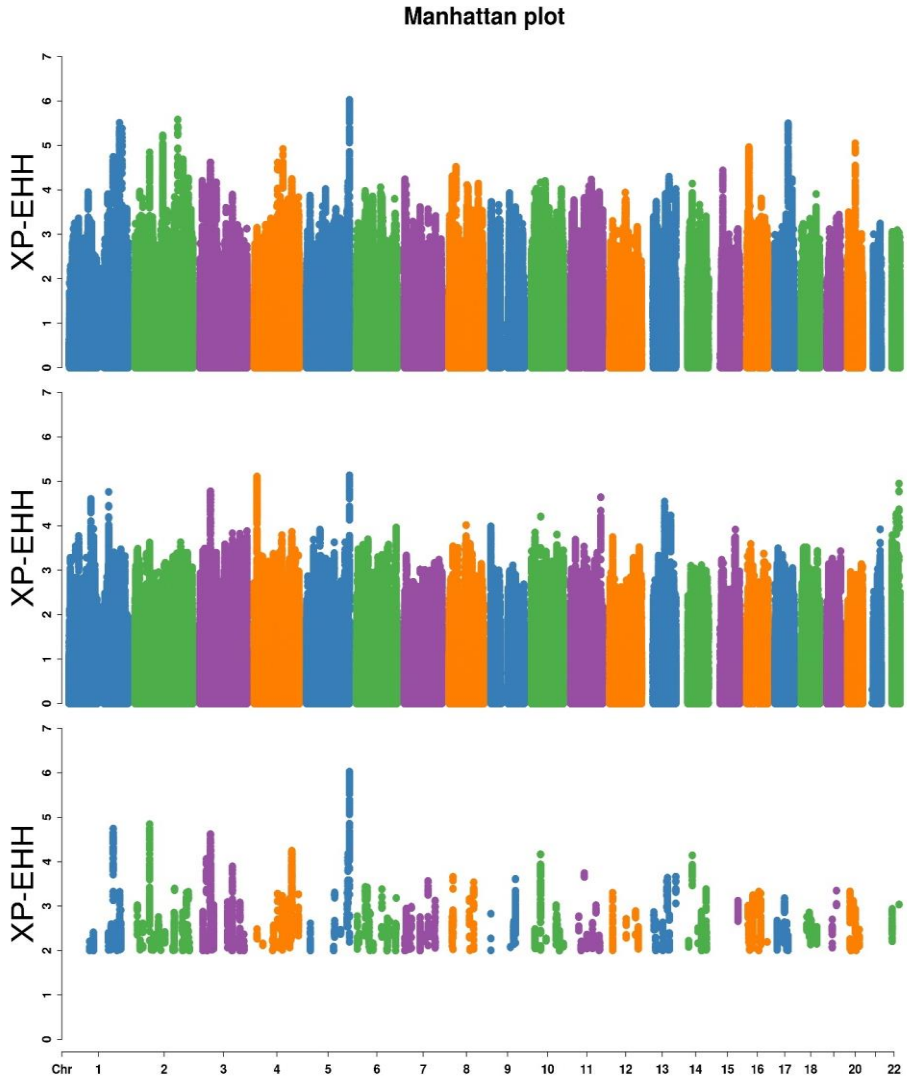


Supplementary Figure 11. a) Manhattan plot of XP-EHH test between Roma compared to Rajput, Romanians compared to Rajput, and the Recent Shared Signals (from top to bottom); b) Manhattan plot of XP-EHH test between Roma compared to Romanians, Rajput compared to Romanians, and the Old Shared Signals.

a)



b)


Manhattan plot
XP-EHH
XP-EHH
XP-EHH
0
1
2
3
4
5
6
7
Chr
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
20
22

Supplementary Tables

Supplementary Table 1. Samples removed from the main analysis and reason for exclusion.

Sample	Population	Reason
RMN-17	Romanian	mtDNA contamination; outlier X heterozygosity; outlier autosomal heterozygosity
S19	Romanian	mtDNA contamination
S25	Romanian	mtDNA contamination
S43	Romanian	mtDNA contamination
S60	Roma	mtDNA contamination
S92	Roma	2nd-degree relation to S9 (Romani)
S85	Roma	2nd-degree relation to S87 (Romani)
S71	Roma	2nd-degree relation to S72 (Romani); Admix individual
S72	Roma	2nd-degree relation to S71; Admix individual
S79	Roma	Discordant self-identification
S74	Roma	3rd-degree relation to S70 (Romani)
S21	Roma	3rd-degree relation to S9 and S17 (Romanies)
S8	Roma	3rd-degree relation to S14 (Romanies)
S66	Roma	Admix individual
S96	Roma	Admix individual
S100	Roma	Admix individual
RMN-11	Romanian	Admix individual
RMN-14	Romanian	Admix individual
S28	Romanian	Admix individual
RMN-21	Romanian	Admix individual

Supplementary Table 2. Estimation of X-chromosome heterozygosity in male samples due to contamination. ML = Maximum likelihood contamination estimate; SE = standard error estimated using jackknife.

Sample	Putative source of contamination	ML	SE(ML)
RMN-17	CEU	0.438325	0.003455738
RMN-17	CHB	0.359364	0.00285054
RMN-17	GIH	0.392264	0.003129929

RMN-17	PEL	0.343802	0.002722593
RMN-17	YRI	0.335109	0.00266473

Supplementary Table 3. Transition versus transversion ratio for Roma and Romanian individuals. Reported values for all, known and novel variants. Novel variants are defined by using dbSNP137.

Marker	**nTs**	**nTv**	**Ts/Tv**
All	14516088	6988721	2.08
Known	9552023	4382784	2.18
Novel	4964065	2605937	1.91

Supplementary Table 4. Candidate genes and nonsynonymous candidate variants in Shared Selection Signals. Locations are in GRCh37 (hg19).

Marker	**Chr**	**Position**	**Ancestral**	**Derived**	**Gene**	**Type**
rs614486	chr1	47138819	T	G	TEX38	nonsyn
rs2056899	chr1	47607851	A	T	CYP4A22	nonsyn
rs1056820	chr13	41515286	T	A	ELF1	nonsyn
rs7799	chr13	41533052	T	C	ELF1	nonsyn
rs2287679	chr19	33600764	T	C	GPATCH1	nonsyn
rs10416265	chr19	33605300	A	G	GPATCH1	nonsyn
rs10421769	chr19	33605312	T	C	GPATCH1	nonsyn
rs1402467	chr2	108994808	C	G	SULT1C4	nonsyn
rs59900519	chr2	135988127	T	A	ZRANB3	nonsyn
rs935615	chr2	135988416	C	T	ZRANB3	nonsyn
rs1112438	chr3	39152345	G	A	TTC21A	nonsyn
rs1453241	chr3	130103709	G	A	COL6A5	nonsyn
rs11917356	chr3	130110550	A	G	COL6A5	nonsyn
rs12488457	chr3	130116696	A	C	COL6A5	nonsyn
rs1497312	chr3	130125116	G	C	COL6A5	nonsyn
rs16827497	chr3	130134492	T	C	COL6A5	nonsyn
rs3762672	chr3	132218623	G	T	DNAJC13	nonsyn
rs34358	chr5	74965122	G	A	ANKDD1B	stopgain

rs2307111	chr5	75003678	T	C	POC5	nonsyn
rs1550526	chr6	13295515	A	C	LOC100130357	nonsyn
rs2305473	chr7	158536267	T	C	ESYT2	nonsyn
rs2305475	chr7	158536345	A	G	ESYT2	nonsyn
rs2788478	chr7	158672619	A	G	WDR60	nonsyn
rs7019716	chr9	26116150	G	T	LOC100506422	nonsyn
rs1056899	chr9	135139901	T	C	SETX	nonsyn
rs2296871	chr9	135173685	T	C	SETX	nonsyn
rs543573	chr9	135202829	T	C	SETX	nonsyn
rs1183768	chr9	135203231	C	T	SETX	nonsyn
rs1185193	chr9	135203409	A	C	SETX	nonsyn
rs7006	chr10	103368654	T	C	DPCD	nonsyn
rs9284879	chr3	44284584	G	A	TOPAZ1	nonsyn
rs2272044	chr3	44692564	C	G	ZNF35	nonsyn
rs529208	chr9	286593	C	A	DOCK8	nonsyn

References

Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. a, Durbin, R. M., Handsaker, R. E., … McVean, G. a. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*(7422), 56–65. https://doi.org/10.1038/nature11632

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., … Schloss, J. A. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1), 7. https://doi.org/10.1186/s13742-015-0047-8

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with

high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Fraser, A. (1992). *The Gypsies*. Blackwell Publishers.

Fu, Q., Mittnik, A., Johnson, P. L. F., Bos, K., Lari, M., Bollongino, R., … Krause, J. (2013). A revised timescale for human evolution based on ancient mitochondrial genomes. *Current Biology*, *23*(7), 553–559. https://doi.org/10.1016/j.cub.2013.02.044

Johnson, P. (2014). contamMix: Mitochondrial genome contamination estimation.

Kircher, M., Witten, D. M., Jain, P., O'roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), 310–315. https://doi.org/10.1038/ng.2892

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*, *15*(1), 356. https://doi.org/10.1186/s12859-014-0356-4

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv Preprint ArXiv*, *00*(00), 3. https://doi.org/arXiv:1303.3997 [q-bio.GN]

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W. M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, *26*(22), 2867–2873. https://doi.org/10.1093/bioinformatics/btq559

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, et al. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Science (80-) [Internet]. 2012 Oct 12;338(6104):222 LP-226. Available from: http://science.sciencemag.org/content/338/6104/222.abstract

Mondal, M., Casals, F., Xu, T., Dall'Olio, G. M., Pybus, M., Netea, M. G., … Bertranpetit, J. (2016). Genomic analysis of

Andamanese provides insights into ancient human migration into Asia and adaptation. *Nature Genetics*, *48*(9), 1066–1070. https://doi.org/10.1038/ng.3621

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., … Reich, D. (2012). Ancient admixture in human history. *Genetics*, *192*(3), 1065–1093. https://doi.org/10.1534/genetics.112.145037

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, *2*(12), e190. https://doi.org/10.1371/journal.pgen.0020190

Sherry, S. T. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, *29*(1), 308–311. https://doi.org/10.1093/nar/29.1.308

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164. https://doi.org/10.1093/nar/gkq603

3.2. Is there adaptation in the human genome for taste perception and phase I biotransformation?

Dobon B, Rossell C, Walsh S, Bertranpetit J. Is there adaptation in the human genome for taste perception and phase I biotransformation? BMC Evol Biol. 2019 Dec 31;19(1):39. DOI: 10.1186/s12862-019-1366-7

3.3. Influence of network topology on the evolution of metabolic enzymes in humans and mammals

Montanucci L, Laayouni H, Dobon B, Keys KL, Bertranpetit J, Peretó J. Influence of pathway topology and functional class on the molecular evolution of human metabolic genes. PLoS One. 2018 Dec 1;13(12). DOI: 10.1371/journal.pone.0208782