The genetics along the Silk Road: structure and evolutionary history of the populations

MED/03 Genetica Medica

Ph.D. student
Massimo Mezzavilla

Ph.D. program Director
Prof.ssa Giuliana Decorti

Thesis Supervisor
Prof. Paolo Gasparini

ANNO ACCADEMICO 2013-2014
Abstract

The understanding of the genetic structure of a population is important to describe its population history, as well as designing studies of complex biomedical traits, including disease susceptibility. The Marco Polo expedition gave us the possibilities to explore several different populations in the Caucasus and Central Asia from Georgia to Kazakhstan, obtain information on taste, and smell perception and several other phenotypes in order to identify the genetic variants implicated. Considering the stratification issue in genetic association studies the aim is to fully characterize the genetic diversity in each population in order to provide a useful array of information for future association studies. Chemosensory phenotypes like olfactory perception, taste perception, are different between and within populations and are probably the results of the combination of gene-environment interactions, for this reason finding new variants could be challenging due to the small number of individuals sampled, in addition it is worth pointing out that a large number of individuals are needed to detect genetic variants with a modest effect on the variability of a phenotype. However, as a large sample size is not always a feasible option, especially in these countries, a population-based approach is needed to take into account the population history and the genetic structure in describing phenotype variation. This thesis defines the population genetic landscape and migration pattern of the population sampled during the Marco Polo expedition and an implementation of a methodology to describe it. In addition, we used an approach based on the population to interpret and analyze the observed pattern of olfactory perception, the role of genetic variation in eye color and finally the relationship between taste perception and food preferences across different countries.
Abstract (Italiano)

La conoscenza della struttura genetica di una popolazione è importante per descriverne la storia, e per la progettazione di studi di tratti biomedici complessi. La spedizione Marco Polo ci ha dato la possibilità di esplorare diverse popolazioni differenti in Caucaso e Asia Centrale, dalla Georgia al Kazakistan, e ottenere informazioni sul gusto, la percezione olfattiva e molti altri fenotipi, tutto questo al fine di esplorare le varianti genetiche implicate in questi ultimi Considerando il problema della stratificazione nell’analisi di associazione genetica, l'obiettivo è di caratterizzare la diversità genetica in ogni popolazione al fine di fornire una gamma di informazioni utili per gli studi futuri di associazione.

Fenotipi chemosensoriali come la percezione olfattiva, la percezione gustativa, sono diversi tra e all'interno delle popolazioni e sono probabilmente il risultato della combinazione d’interazioni gene-ambiente, e la ricerca di nuove varianti genetiche implicate in essi potrebbe essere difficile a causa del modesto numero di individui campionati nelle diverse popolazioni della Via della seta, in aggiunta è bene precisare che un grande numero di individui sono necessari a rilevare le varianti genetiche con un effetto modesto sulla variabilità di un fenotipo. Tuttavia ottenere un campione numeroso, non è sempre una possibilità fattibile, soprattutto in questi paesi; dunque un approccio basato sulla popolazione è necessario per tener conto della struttura della stessa nel descrivere la variabilità fenotipica. Questa tesi descrive la variabilità genetica delle popolazioni campionata durante la spedizione di Marco Polo e l'implementazione di una metodologia per descriverla. Inoltre è stato utilizzato un approccio basato sulla popolazione per comprendere la variazione nella percezione olfattiva, il ruolo di varianti genetiche in colore degli occhi e, infine, il rapporto tra percezione del gusto e le preferenze alimentari in diversi paesi.
# TABLE OF CONTENTS

CHAPTER 1. The Marco Polo scientific expedition ................................................. 5  
  1.1 The “Marco Polo” Expedition and the Silk Road ........................................ 6  
  1.2 The project ........................................................................................................ 8  
  1.3 Why studying the population genetic structure of these countries? .............. 8  

Aims of the present study .......................................................................................... 11  

CHAPTER 2. Genetic characterization and population history ........................ 12  
  2.1 Background ....................................................................................................... 13  
  2.2 Methodologies applied ...................................................................................... 16  
  2.3 Publications ...................................................................................................... 19  
     2.3.1. Article: Effective population size estimation ............................................... 20  
     2.3.2. Article: Genetic structure of the population of the Silk Road ..................... 29  

CHAPTER 3. Genetic history of the Armenians ............................................ 40  
  3.1 Introduction ...................................................................................................... 41  
  3.2 Results and discussion .................................................................................... 43  
     3.2.1. Structure of the Armenian population ..................................................... 48  
     3.2.2. Relationship to ancient Europeans ......................................................... 51  
  3.3 Conclusion ....................................................................................................... 54  
  3.4 Material and Methods ..................................................................................... 54  

CHAPTER 4. Genetic analysis of taste and olfactory perception ............. 59  
  4.1 Background ..................................................................................................... 60  
  4.2 Olfactory perception ....................................................................................... 61  
     4.2.1. Introduction .............................................................................................. 61  
     4.2.2. Material and Methods ........................................................................... 62  
     4.2.3. Results ................................................................................................... 63  
     4.2.4. Conclusion .............................................................................................. 65  
  4.3 Publications .................................................................................................... 68  
     4.3.1. Article: Genetic of the eye colour ............................................................. 69  
     4.3.2. Article: Bitter perception and food liking .............................................. 74  

Future developments .............................................................................................. 83  

Bibliography ........................................................................................................... 84  

CURRICULUM VITAE ............................................................................................ 89  

LIST OF PUBLICATIONS ...................................................................................... 91
CHAPTER 1. The Marco Polo scientific expedition
1.1 The “Marco Polo” Expedition and the Silk Road

From 17th July to 12th September 2010, a group of scientists, journalists and media professionals travelled along parts of the Central Asia and Caucasus (referred in this thesis as “Silk Road population”) in order to carry out a variety of genetics research projects. This expedition was called “Marco Polo”, as the Venetian explorer who travelled the Silk Road in the thirteenth century. The aim of this scientific expedition was to study three main topics: population genetics, taste genetics and food preference genetics. Moreover, analyses on (the genetics of) olfactory perception and hearing, has been executed during expedition, allowing a better understanding of the other genetic data.

During the October 2011 the same group of scientists travelled through Armenia and Crimea to continue collecting data along the “Silk Road”. Finally, in June 2012 the last country was sampled, that is Kirghizstan, marking the end of the sampling project started in 2010 along the Silk Road. The group travelled more than 14,000 km across 8 countries, visiting 22 communities, and took samples from more than 800 people. (See Figure 1.1)
Figure 1.1. Marco Polo Expeditions past and present, countries sampled during the expedition in 2010, 2011 and 2012
1.2 The project

The project aims to elucidate the genetic variants implicated in chemo sensory perception, like taste and smell, analysing if there are any genetic role in food preferences and also the relationship between all the different traits. During the expedition were collected:

- Bitterness and saltiness taste tests;
- A questionnaire about dietary preferences;
- A sense of smell test;
- An audiometric test to measure hearing;
- A brief general questionnaire on their general state of health and lifestyle choices;
- A collection of saliva samples - from which the DNA can be extracted, so that a genetic analysis of the population can be carried out.

1.3 Why studying the population genetic structure of these countries?

The Marco Polo project aims to study the complex relationship between genetic variants and phenotype through various methodologies like GWAS studies. Genome-wide association studies (GWAS) have success in identifying variants associated with medical traits and complex diseases. Correlating differences in disease/phenotype frequencies between groups with differences in allele frequencies at single nucleotide polymorphism (SNP) is the aim of GWAS studies. For this very reason the frequencies of the two alleles at one or more loci are of primary interest for identification of genes affecting diseases. Even the simplest study design carries the fundamental assumption that any differences in allele frequencies between groups (like case and controls) actually relate phenotype and there are no unobserved confounding effects (Schleselman and Schneiderman 1982). However allele frequencies also for causal variants are known to vary widely within and
between populations, regardless of disease status (Cavalli-Sforza et al. 1994).

Each population has specific genetic and social history, and thus ancestral patterns of migration, mating practices, expansions and bottlenecks, and also environmental characteristic that could affect natural selection this results in stochastic variation which yield differences in allele frequencies between individuals (Slatkin 1991), and probably none of them is necessarily associated with any particular disease. These population-frequency differences are common throughout the entire genome, and include many genes of known medical relevance (Goddard et al. 2000; Stephens et al. 2001). As results our assumption of no confounding effects in genetic applications of the genetic study design could be violated.

We should consider also the following aspect of population genetics: many outbred populations are confounded by genetic admixture at some level; the challenge is not merely to show that it exists, but to avoid making erroneous conclusions because of it. In addition the use of isolates population is becoming common in GWAS studies (Timpson et al. 2014) and isolation could shape the genetic texture of a population completely different from neighbour populations, only by genetic drift that belongs also to the same country and region (Esko et al. 2012). When cases and controls have different allele frequencies attributable to diversity in background population, unrelated to outcome status, a study is said to have population stratification. Two circumstances must be met for population stratification to affect genetic association studies: differences in disease prevalence must exist between cases and controls; and variations in allele frequency between groups must be present (Wacholder et al. 2000). Failure to replicate genetic association studies is an actual concern (Ioannidis et al. 2001; Cardon and Palmer 2003; Terwilliger and Göring 2009). For most complex human diseases, the reality of multiple disease- predisposing genes of modest individual effect, gene-gene interactions, gene-environment interactions, population heterogeneity of genetic and environmental determinants of disease, and the concomitant low statistical power mean that both detection and subsequent replication will probably be very challenging (Risch and Merikangas 1996; Cardon and Palmer 2003; Terwilliger and Göring 2009).
It becomes apparent that population stratification is one of many possible reasons for non-replication of association results. We must re-emphasize that, when substantial population substructure does exist, powerful methods for detection and correction for it are now available (ALEXANDER et al. 2009) (PATTERSON et al. 2006). These analyses will help to understand if population substructure was responsible for the initial evidence for association, or indeed, if it was accountable for masking the effect of some genetic variants in the replication sample.

The simplest way to obtain statistically strong results is to have a large sample of the same homogeneous population; however this solution is not always feasible and should be noted that there are several advantages in sampling many different populations, despite the low sample size. For example, some rare variants implicated in phenotype variation could be drifted at high frequency or the number of individuals with a particular rare phenotype could be higher in one population than in the others, furthermore using samples from different populations we could pinpoint the interaction between genes and environmental variables.

In addition, considering that diversity between members of the same population is very large (BARBUJANI et al. 2013), the population characterization of a sample is a requirement.

The Marco Polo project aims to find new genes and genetic variants, involved in chemosensory phenotypes in various populations from the Caucasus to the border of East Asia. However, despite common methods to detect stratification could be used, the population genetic history of each population is different, and despite no initial strong differences based on a first quick analysis, further analysis could elucidate a deeper history which could help to explain or better understand phenotypic differences between populations.

In this thesis, we compare different datasets of modern population in order to describe and characterize the genetic structure of the Silk Road providing a useful array of information for further studies that involve phenotype variations in these communities. The main question is not if the samples obtained from the Marco Polo expedition should be grouped but how we can characterize our individuals and how to use this information, to understand the relationship between genes, environment and phenotypes.
Aims of the present study

The project is divided into two parts:

1) The first part regards the genetic characterization of the Silk Road’s populations. As far we know population stratification is an important issue for association studies, the populations sampled during the Marco Polo expedition span for very different and distant geographical areas, and little is known about the population substructure of these countries. The aims are to provide a useful array of information regarding population stratification and population history, in order to better understand population differences and similarities at phenotype level (in the context of the phenotypes collected). For example, level of PROP perception, food preferences and smell perception are very different between and within countries.

2) The second part consists in the use of population-based approaches to understand phenotypic differences between populations. In this thesis, we highlight how the study of different and sometimes genetically and distant populations will help us to understand the relationship between several complex traits, like PROP perception and food preferences. In addition we will analyse a complex trait as smell perception in order to find variants implicated in olfactory perception, highlighting differences and similarities between populations.
CHAPTER 2. Genetic characterization and population history
2.1 Background

Today we have access to large amounts of genetic data from many different populations, especially coming from studies whose aims are to understand the relationship between genetic variants and phenotypic/biomedical traits. As well as stratification is a difficult issue, even the answer is not simple Each population has its history and characteristics and only by understanding all these aspects of a population, we can really try to solve the issue of stratification.

Population genetics try to answer the question alike: "How is structured the population I am observing? Which are the genetic characteristics of my population? Which was the history of this population and the relationship with others? How did a particular pattern of genetic diversity arise?" We can answer the above questions and address the stratification matter by combining several analytical approaches, starting from a description of the variation observed in the data, up to the use of inferential methods to estimate evolutionary and demographic parameters, could answer the previous questions and address the stratification matter. This procedure should not only help to avoid false positive results, but also to find and understand specific population variants that affect a particular phenotype/medical trait which otherwise could be lost in larger, less focused studies.

The study of population genetic structure of a single or several populations requires knowledge of many statistics within and between populations. In order to disentangle the problem of describing the genetics of a population, we should focus on different kind of analyses that can be divided into three types. Type I analyses regard the ones within the population, like the pattern of runs of homozygosity, inbreeding analysis, statistic of selection based on haplotype structure, estimation of effective population size through time and population substructure. Type II analyses refer to the relationship between populations, like time of divergence between them, admixture events, the proportion of admixture, and patterns of natural selection. Recently the availability of ancient DNA (aDNA) had made out the possibility to understand the relationship between several population and ancient genomes (GREEN et al.)
2010; RASMUSSEN et al. 2010; REICH et al. 2010), shading light into the ancient population history. This kind of analysis could be called as type III. (See Figure 2.1).

Figure 2.1. Workflow of population genetic characterization

In this chapter the following aspects will be explained:

• A useful set of concept used in population genetics (Box: Concepts)
• Main methodologies applied to understand the genetic structure of a population, subdivided into type I, II and III.
• An implementation of a method used to estimate the effective population size and the time of divergence between populations. Article I: Neon: An R Package to Estimate Human Effective Population Size and Divergence Time from Patterns of Linkage Disequilibrium between SNPS.
• The genetic background of the populations of the Silk Road. Article II: Genetic landscape of populations along the Silk Road: admixture and migration patterns.
Genetic drift: The stochastic process of sampling from one generation to another determines a random variation in allele frequencies over time and is called random genetic drift (Wright 1931). Genetic drift may cause allelic variants to disappear completely or to be fixed (reaching frequency of 1), and therefore reduces the population gene diversity.

Admixture: when individuals from two or several genetically distinct population begin interbreeding (due to migration), resulting in a new group of individual with mixed ancestry.

Effective population size (Ne): the effective population size is the size of an idealized population that experiences the same amount of genetic drift of the population under study (ref). The greater is the effective population size the smaller are the drift effects.

Bottleneck: instantaneous reduction of census size in a population, depending on the magnitude I could affect or not affect genetic diversity and the effective population size.

Gene flow: Migration is the movement of individuals from an occupied area to another, and differs from colonization since the latter regards a movement into a previously unoccupied territory. Gene flow is the outcome of the process of migration, when a migrant contributes to the next generation in the new location, and depends on the reproductive success of the migrants in the new area.

Local ancestry: considering that the genome of one individual is divided into chromosome segments of definite ancestral origin, the local ancestry tries to find the segment boundaries and assign each segment’s origin to a set of reference population.

Global ancestry: considering one individual global ancestry is the proportion of ancestry from each contributing reference population, considered as an average.

Fst: a measure of the extent of genetic differentiation among subpopulations can range from 0.0 (no differentiation) to 1.0 (complete differentiation, subpopulation fixed for different alleles).

Inbreeding coefficient: the mean reduction in heterozygosity of an individual due to non-random mating within a population.

Runs homozygosity (ROH): long stretches of homozygous SNP markers across the genome. Long ROH are common in recent isolates, short ROH are common in ancient isolated populations.
2.2 Methodologies applied

The first procedure when we analyse samples from one or several populations, is to apply a quality control of the data. That consists in first removing individuals with high missingness and then genetic loci with low genotyping rate. After this and depending on the subsequent analyses, loci with lower allele frequency (called rare alleles) and monomorphic loci should be removed. In population genetics studies the threshold commonly used for rare alleles is between 0.01-0.05.

In the context of cleaning the data, a common approach is to use principal component analysis to remove outliers that, you know, may reflect errors. If the hypothesis is that our sample comes from one population, outlier individuals appear in the Principal Component analysis (PCA) completely separated from other individuals, which instead tend to form a cluster. It is important to note that the concept of “outlier” is quite controversial because sometimes it is the population substructure that is responsible of their presence rather than sampling errors. In this way the concept of removing outliers in the sampled Silk Road’s populations could be considered an error because in this way we will lose precious information regarding the genetic structure of populations.

The final step is to remove very close related individuals, as could inflate Principal component analysis and several others statistics. However it is noteworthy that small populations with history of isolation, low effective population size and inbreeding could have exceptionally high levels of relatedness. So far the best approach to the question is to remove only duplicates and siblings, and get closer to understanding the genetic makeup of population under study.

Regarding the analyses of type I, the major aspects to investigate are the genetic signatures of the populations/samples. Inbreeding coefficient, level of relatedness, runs of homozygosity (see Box 1 for definitions) are common statistics used to describe a population, giving initial evidences if the
population is an isolate or an outbred one. Table 2.1 summarizes useful software/methodologies.

Furthermore an important parameter that can be estimated within a population is undoubtedly the effective population size (Ne). A detailed description of the importance of effective population size and an implementation of a method based on linkage disequilibrium to estimate effective population size will be showed in the article: Neon: An R Package to Estimate Human Effective Population Size and Divergence Time from Patterns of Linkage Disequilibrium between SNPS.

Regarding the analysis of type II, the Principal component analysis after the merge of dataset with several reference populations is the best approach to find evidence of substructure in a population.

However the question “How many populations are present in my dataset?” requires the conjunctions of the result of PCA with a clustering algorithm like Mclust (Fraley and Raftery 2006), methods based on the estimation of global ancestry like ADMIXTURE (Alexander et al. 2009), and also with methods based also on local ancestry, like SupportMix (Olberg et al. 2012). After the identification of the population/genetic cluster, more detailed analysis should be performed. TreeMix (Pickrell and Pritchard 2012) and f3-statistics (Reich et al. 2009) are useful formal test to detect gene flow between populations.

After the identification of the main possible event in the population history an estimation of when a specific admixture event happened in the past could be accomplished using linkage disequilibrium methods as ALDER (Loh et al. 2013).

The best approach is to combine several of these methods in order to find the best-supported results. A description and application of all the methodologies will be presented in the Article: Genetic landscape of populations along the Silk Road: admixture and migration patterns.

Regarding type III analysis with the availability of ancient DNA using an approach based on using outgroup f3 statistics (Reich et al. 2009; Raghavan et al. 2013) will help also to look deeper in the ancient history of the population. A description of the methodology will be explained in Chapter 3: Genetic history of the Armenians.
Table 2.1

<table>
<thead>
<tr>
<th>Analysis Type</th>
<th>Name of the software/methodology</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>PLINK</td>
<td>Estimate various individuals and population statistics: runs of homozygosty, inbreeding coefficient etc. Quality control parameters</td>
</tr>
<tr>
<td>I-II</td>
<td>EIGENSOFT.smartpca</td>
<td>Perform principal component analysis on genetic data</td>
</tr>
<tr>
<td>I-II</td>
<td>Mclust R package</td>
<td>Model-based population clustering</td>
</tr>
<tr>
<td>I-II</td>
<td>4P</td>
<td>Estimate various individual and population statistics, five different measure of genetic distance</td>
</tr>
<tr>
<td>I-II</td>
<td>NeON R package</td>
<td>Estimate effective population size (Ne) and time of divergence between populations</td>
</tr>
<tr>
<td>II</td>
<td>ADMIXTURE</td>
<td>Estimation of global ancestry</td>
</tr>
<tr>
<td>II</td>
<td>SupportMix</td>
<td>Estimation of local ancestry</td>
</tr>
<tr>
<td>II</td>
<td>F3-statistic</td>
<td>Estimation of admixture events</td>
</tr>
<tr>
<td>II</td>
<td>ALDER</td>
<td>Estimation and dating of admixture events</td>
</tr>
<tr>
<td>II</td>
<td>TreeMix</td>
<td>Ancestry graph analysis</td>
</tr>
<tr>
<td>III</td>
<td>Outgroup-f3 statistic</td>
<td>Estimate genetic affinity between population and ancient genomes</td>
</tr>
</tbody>
</table>
2.3 Publications
2.3.1. Article: Effective population size estimation
Neon: An R Package to Estimate Human Effective Population Size and Divergence Time from Patterns of Linkage Disequilibrium between SNPs

Massimo Mezzavilla1,2 and Silvia Ghirotto3
1Institute for Maternal and Child Health IRCCS "Burlo Garofolo", Trieste, Italy
2University of Trieste, Italy
3Department of Life Sciences and Biotechnologies, University of Ferrara, Italy

Abstract

Objective: Estimating the effective population size \( (Ne) \) is crucial to understanding how populations evolved, expanded or shrunken. One possible approach is to compare DNA diversity, so as to obtain an average Ne over many past generations; however as the population sizes change over time, another possibility is to describe this change. Linkage Disequilibrium (LD), and information about these changes, and therefore a large number of densely linked markers are available, can be used to monitor fluctuating population size through time. Here, we present a new R package, Neon that has been designed to explore population LD patterns to reconstruct two key parameters of human evolution: the effective population size and the divergence time between populations.

Methods: Neon starts by importing binary or pairwise-LD PLINK files, and allows (a) to assign a genetic map position using HapMap (NCBI release 36 or 37) (b) to calculate the effective population size over time exploiting the relationship between Ne and the average squared correlation coefficient of LD \( (r^2) \) within predefined recombination distance categories, and (c) to calculate the confidence interval about Ne based on the observed variation of the estimator across chromosomes; the outputs of the functions are both numerical and graphical. This package also offers the possibility to estimate the divergence time between populations given the Ne values calculated from the within-population LD data and a matrix of between-populations \( F_{st} \). These routines can be adapted to any species whenever genetic map positions are available.

Results and Conclusion: The functions contained in the R package Neon provide reliable estimates of effective population sizes of human chromosomes from LD patterns of genome-wide SNPs data, as it is shown here for the populations contained in the CEPH panel. The Neon package enables to accommodate variable numbers of individuals, populations and genetic markers, allowing analyzing those using standard personal computers.

Keywords: R package, Effective population size; Divergence time; Demographic parameter; Linkage disequilibrium; Recombination map; SNPs panel; Polymorphism data

Introduction

The effective population size \( (Ne) \) is at the same time one of the most important parameters of natural populations, and one of the most difficult to evaluate directly [1,2]. Common approaches to estimate Ne involve temporal methods [3,4] that require at least two samples, separated in time, of the same population. Other single-sample methods are based on the heterozygote excess [5,6], on the amount of linkage disequilibrium in neutral, unlinked loci [7], or on measures of the extent of current genetic variation [8,9]. Recently, the considerable progress in the field of population genetics, along with the development of methods based on the coalescent theory, have allowed to estimate the effective population size through time directly from a sample of gene sequences [10], or entire genomes [11,12]. Another way to study past populations dynamics exploits the information contained in the pattern of linkage disequilibrium (the non-random association between genetic loci) between densely spaced single nucleotide polymorphisms (SNPs) data. The Ne that is usually calculated from genomic variation represents an estimate of the long-term Ne, that is an average of the effective population size over many past generations, disregarding of past demographic fluctuations. By contrast, the extent and the strength of linkage disequilibrium between two genetic loci contains information about population dynamics such as changes in the effective population size through time [13,14]. Indeed, levels of LD increase due to random genetic drift and decays due to recombination, according to a recombination rate between pairs of genetic markers that are positively correlated with the distance between genetic markers. This means that LD decreases at increasing physical distance between loci. Consequently, if we consider that levels of LD depend on both Ne and on the recombination rate between markers [14], LD between loci separated by large distances along the chromosome reflects relatively recent Ne whereas LD over short recombination distances depends on relatively ancient Ne [13]. This relationship between LD and Ne, detailed in the further section, can be exploited to monitor fluctuating population size through time.

As well as offering a gold opportunity to follow the dynamics of demographic events, in addition the estimation of Ne from LD can be used to date the time since two populations diverged from one another.

Materials and Methods

All the functions contained in the Neon package have been developed for the free statistical R environment (http://www.r-project.org) and run under the major operating systems (UNIX and OSX).

*Corresponding author: Silvia Ghirotto, Department of Life Sciences and Biotechnologies, University of Ferrara, Italy. Tel: (+39) 0532-455312; Fax: (+39) 0532-249761; E-mail: ghirotto@unife.it

Received November 29, 2014; Accepted December 17, 2014; Published January 07, 2015


Copyright: © 2015 Mezzavilla M, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
Most of the NeON functions interact with the PLINK program [15] or relate with PLINK data files, a widely used data format in population genetic studies. The workflow of a complete NeON analysis consists of six steps as shown in (Figure 1). A detailed description of the functions available in NeON is reported below.

**Nemap (bim.file, map.file)**

Since the method implemented here to estimate Ne and divergence time is based on the recombination or genetic distance between SNPs, it is fundamental to have available correct genetic map information. Nemap actually prepares the file to update the genetic map information of the markers in your PLINK data file (binary format), based on the recombination rates and hotspots compiled of the markers in your NeMap file obtained from the NeMap website. Nemap requires as arguments the name of the .bim file data and the path to the recombination rates and hotspots compiled file that can be downloaded from the HapMap website http://hapmap.ncbi.nlm.nih.gov/downloads/recombination (e.g. NCBJ36/hg18 and GRCh37/hg19). The genetic map information is then extracted by matching the physical positions of the SNPs contained in the two files (the dataset and the recombination map). Nemap returns a list of SNP identifiers (SNP list) that can be used by the following function, NeUpdate, to actually update the genetic map information. Nemap requires as arguments the name of the .bim data file and the path to the recombination rates and hotspots compiled files (a single file for each chromosome), along with the prefix used in each file name before the chromosome number, paying attention to use the map that matches the build of your data. We provide the properly formatted recombination map files for the two last releases of human variation data, i.e. hg18 and hg19. Given so, to map your bim file to hg18 the right call of the Nemap function would be Nemap ("./mydata.bim";./genetic_map/genetic_map_b36_chr50), whereas to get genetic map information for hg19 it would be Nemap ("./mydata.bim";./genetic_map/genetic_map_GRCh37).

**Ne Update (plink.file, snp.list, outfile)**

This function relies on the snp.list file created by the previous function (Nemap) to update your PLINK .bim data file with the correct genetic map information. NeUpdate requires as arguments the prefix of your PLINK data files (i.e. without the .bim, bed or .fam extension), the name of the file obtained from the Nemap function, and the prefix of the updated PLINK data files that will be created. The PLINK executable has to be in the same folder of the data files.

![Figure 1: Workflow of a complete NeON analysis.](image)

Ne LD (plink.file, geno=0.02, mind=0.9, ld.window.kb=500, ld.window=9999, outfile="output.ld")

This function estimates the squared correlation coefficient of linkage disequilibrium \((r^2)\) between markers. The default parameters of the function are a genotyping rate higher than 98% \((\text{geno}=0.02)\), a rate of individual missing data lower than 10% \((\text{mind}=0.9)\), a window of 500 kilobases \((\text{ld.window.kb}=500)\) and 9999 SNPs \((\text{ld.window}=9999)\). These parameters are also detailed in the PLINK tutorial online at http://pngu.mgh.harvard.edu/~purcell/plink/ld.shtml#ld1, and can be changed to fit your purpose. Other than these parameters, NeLD requires the prefix of your PLINK data files (i.e. without the .bim, .bed or .fam extension) and the name of the output file. The PLINK executable has to be in the same folder of the data files.

**Nestimate (file.ld, sample.size, min.R2=0.001, max.R2=0.999, method="MG", min.cfr=5)**

This function estimates the effective population size. It requires the output.ld obtained from the NeLD function and applies the formula \(N_{\text{e}} = 1/(4c^2) \times [(1/r^2) - 2]\), where \(c\) is the distance between genetic markers in Morgan. Nestimate creates several categories of recombination distance, with incremental upper boundaries of 0.005 centiMorgan \((\text{cM})\) up to 0.25 \(\text{cM}\), and calculates the \(r^2\) for each pair of markers in each recombination distance category. To do this, we implemented two different methods: one (method="McEvo") is the same method that has been used in [17], with 50 not overlapping bin sizes from 0.005 up to 0.25 \(\text{cM}\); the other (method="MG", the default) is the Mezzavilla-Ghirotto method, which consider 250 overlapping bins with a step of 0.001 \(\text{cM}\) from 0.005 to 0.25 \(\text{cM}\). Nestimate calculates a value of effective population size, according to the formula above, within each of the 50 or 250 identified bins. The Ne value calculated in each bin corresponds to the effective population size at a specific moment in the past, i.e.1/2 generation ago [13], with \(c\) calculated as the mean value in each recombination distance category. Other parameters of the function are: sample size, that is the size of your sample to allow the \(r^2\) value to be corrected according to the formula \(r^2 = r^2 - 1/n\) (where \(n\) is the sample size); \(\text{min.R2}\) and \(\text{max.R2}\) that are the minimum and the maximum \(r^2\) allowed for the Ne estimation (very high and very low \(r^2\) values, e.g. equal to 0 and 1, may lead to untreatable results), and \(\text{min.cfr}\) that is the minimum number of comparisons in each recombination category to allow the bin to be considered. Nestimate returns a data frame with the values of the effective population size and the correspondent time in the past (in generations), for each bin, for each chromosome.

**Ne_CI (Nestimate.output, ci=c(0.05,0.5,0.95))**

This function estimates the long-term \(N_{\text{e}}\) and its confidence intervals. The long term \(N_{\text{e}}\) is calculated as the harmonic mean [18] of the effective population sizes along the generations in the past. The confidence interval of the long term \(N_{\text{e}}\) is calculated using each chromosome as a replicate (default 5%, 50% and 95% percentile of the distribution of the \(N_{\text{e}}\) over each chromosome). Ne_CI requires as input the output of the previous function (Nestimate).

**Ne_Med (Nestimate.output, method="MG", ci=FALSE, ci.int=c(0.05,0.5,0.95))**

This function calculates the demographic function (effective population size over time) of a population along with its confidence interval, calculated as above, for each bin. Ne_Med requires as input the output of the Ne_Med function and the method used to bin the data in recombination distance categories ("McEvoy" or "MG", MG as
match the population labels reported in the Ff generation ago). This function returns a data frame with the ci.int parameter of the function. The default confidence interval is the 90%; once again, it can be modified changing the values of the ci.int parameter of the function. The function returns a data frame with the first three columns indicating the quantiles of the distribution of the effective population size for each bin over all chromosomes (the default values are the 90% confidence interval and the median value) and the last column with the moment in time to which the effective population size is referred (1/2c generation ago).

\[
\text{Ne\_Plot (Ne, file, approx=TRUE, ylim=c(0, 15000), xlim=c(200, 6000), main="Ne from linkage disequilibrium", xlab="Generation ago", ylab="Ne", ci=TRUE)}
\]

This function is useful to obtain a graphical representation of the changes in the effective population size of a population over time. Ne\_Plot takes as input the data frame with the effective size and temporal information obtained from the Ne\_Med function for each bin, and plots the demographic function of the population. We indicated some default values for the standard R graphical parameters; obviously they can be modified, if needed.

\[
\text{Tdverg (Fst, All\_H)}
\]

This function returns a matrix of the time of divergence between populations in generation following: \( T=\ln (1-F_{ST})/\ln (1-2Ne) \). Tdverg requires a matrix of pairwise \( F_{ST} \) between populations (Fst) and a text file with a list of the long-term Ne for each population, with header that match the population labels reported in the \( F_{ST} \) matrix (All\_H).

We validated the efficiency of the Tdverg function in correctly estimating the time of split between two populations using a forward simulation-based approach through the python library simuPOP [19]. A detailed description of the parameters used in the simulations is reported in (Table 1). Each scenario was replicated 100 times, and we evaluated the power in the divergence time estimation every 100 generations sampling 25 individuals per population.

**Table 1:** Simulation parameters used in the forward simulations.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Scenario 1</th>
<th>Scenario 2</th>
<th>Scenario 3</th>
<th>Scenario 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ne ancestral</td>
<td>10000</td>
<td>10000</td>
<td>10000</td>
<td>10000</td>
</tr>
<tr>
<td>Ne_1</td>
<td>5000</td>
<td>2500</td>
<td>5000</td>
<td>2500</td>
</tr>
<tr>
<td>Ne_2</td>
<td>5000</td>
<td>5000</td>
<td>5000</td>
<td>5000</td>
</tr>
<tr>
<td>Sample size</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>N SNPs</td>
<td>2200</td>
<td>2200</td>
<td>2200</td>
<td>2200</td>
</tr>
<tr>
<td>N chromosome</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>SNP mutation rate</td>
<td>2.00E-08</td>
<td>2.00E-08</td>
<td>2.00E-08</td>
<td>2.00E-08</td>
</tr>
<tr>
<td>Recombination rate between SNPs</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Migration</td>
<td>0</td>
<td>0</td>
<td>1 migrant per generation</td>
<td>1 migrant per generation</td>
</tr>
</tbody>
</table>

To show how to perform a complete analysis using NeON, we analyzed the populations contained in the CEPH panel [21]. Since the method implemented in NeON to calculate Ne relies on LD patterns, to avoid any bias from small sample size, we decided to consider only the CEPH populations with a sample size >20 individuals. We started from the rough data, represented by the PLINK binary files that can be downloaded at http://www.hapsc.org/hgdp/files.html. The map files we used to build the genetic map were downloaded at http://hapmap.ncbi.nlm.nih.gov/downloads/recombination/ and needed a little editing to reduce his size to 2,500 individuals 3) the two populations have the same effective size but including bi-directional migration, and 4) one population underwent a bottleneck (as in case 2) and including bi-directional migration (Figure 2A). As for the scenarios with migration [3,4], we considered the two populations exchanging one migrant per generation; this value has been used as the lowest gene flow to avoid panmixia [20]. A detailed description of the parameters used in the simulations is reported in (Table 1). Each scenario was replicated 100 times, and we evaluated the power in the divergence time estimation every 100 generations sampling 25 individuals per population.

**Results**

Figure 2: Forward simulation analysis. A) Simulated scenarios B) Results of the simulations: the black line represents the real values, the grey lines represent a range spanning 200 generations around the real value and the dots represent the values we estimated.
fit the requirement of the Neomap function: the recombination map has to be in a single folder, with separate files for each chromosome, and each of these files needs to be structured with five columns with an header (Chromosome; Position (bp); Rate (cM/Mb); Map (cM)). It is fundamental to pay attention to consider the genetic map corresponding to the correct release of the data. Within the package we already provide the edited genetic map corresponding to the human genome reference NCBI36/hg18 and GRCh37/hg19. If your .bin file already contains information about the genetic map, you should skip this step and proceed with the effective population size estimation using NeLD and Nestimate. For each population we ran NeLD function that exploit the PLINK program to estimate the strength of the linkage disequilibrium between markers \( r^{2} \) and to arrange the data in a specific output file for the subsequent analysis (output.ld). Through Nestimate, the \( r^{2} \) values were binned into distance categories, averaged, and related to Ne as \( E \left( r^{2} \right) = 1 \left( 1 - 2N_{e}c \right) \), where \( c \) is the genetic distance between loci in Morgans [14]. The so calculated Ne for each bin corresponds to an estimate of the effective population size \( N_{e}c \) generations ago [13]. This function returns a data frame with the estimate of Ne for each bin for each chromosome separately, along with the time to which the estimate is referred (in generations). For each bin, and hence for a specific moment in the past, a single estimate of the value of \( N_{e}c \) is obtained using the Ne_Med function, that calculates the median of the distribution of the Ne estimates across all chromosomes, together with user specified quantiles of this distribution (default: 0.05-0.95). The demographic functions describing the variation of the effective population size through time were visualized by means of Ne_Plot, a function that takes the output of Ne_Med and plots the demography of the corresponding population. Figure 3 shows the functions of the effective population size through time for the CEPH populations as resulting from Ne_Plot. The x-axis represents the time (in generations) from the present (on the left) to the past (on the right). The time depth depends on the minimum distance categories chosen to bin the data, here corresponding to the default value of 0.005-0.01 cm (~6,500 generation ago). This range of values also represents the lower boundary allowed by our marker distances which may have been particularly affected by gene conversion, for which the presented method does not account for [14]. The y-axis represents the effective population size values. It is possible to follow the demographic history of a population going from left to right, with solid lines representing the median value of effective population size over all chromosomes in each bin, and dotted lines representing the value of the 5th and 95th quantile of the distribution of the effective population size over all chromosomes in each bin. As it is shown in (Figure 3), the demographic history of populations clearly reflects their geographical localization: African populations have had maintained stable and larger over time (with a slight decrease in recent times, especially for Biaka Pygmies), whereas most non-African populations started to expand around 1,000 generations ago (corresponding to 25,000 years ago, considering a generation time of 25 years). This is particularly evident for European (e.g. French) and Asian (e.g. Han) populations. Other populations, like Yakut in Siberia or Maya in South America, show relatively low and constant population sizes over time. With the Ne_CI function we calculated the long term Ne for each considered population, along with its confidence interval estimated as the harmonic mean of Ne at the 5th and 95th quantile of the distribution of the effective population sizes for each bin. Figure 4 shows the long term Ne for the considered populations. The range estimated spans from 10,000 (Africans) to 4,000 (Maya and Yakut), consistent with previous estimates obtained by different methods [22-24]. We also compared, by means of a Mantel test, our CEPH Ne estimates with those reported in a previous study [25] for the same populations, obtaining a correlation coefficient of 0.946 (p-value<2.2e-16).

As well as offering the possibility to study fluctuating population size across time and space, LD pattern can be used to explicitly date population divergence times (T). Under neutrality, the level of population differentiation is determined by genetic drift, the extent of which depends on Ne and on the time since the populations diverged. Having an estimate of Ne, and knowing the amount of differentiation (measured by \( F_{st} \)) between a pair of populations, it is possible to estimate their separation time in generations according to the formula \( T=ln \left( 1- F_{st} \right)/ln(1-1/2N_{e}) \) embedded in the Tdiyberg function. We estimated the pairwise Weir and Cockham \( F_{st} \) [27], using the software 4P [28], available online at (www.unife.it/dipartimento/biologia-evoluzione/ricerca/evoluzione-e-genetica/software). The output of the function is a matrix where each value represents the divergence time of a specific pair of populations. To visualize the evolutionary relationships among populations, we calculated an unrooted UPGMA from the divergence time matrix exploiting the upgma function of the phangorn R package (Figure 5). From the tree it is clear that separations happened more recently for populations from the same geographical area (e.g. between Central South Asian, European, and East Asian populations), whereas a long branch (namely a longer separation time) separate Africans from non-African populations. The Mozabite, a population from North Africa, falls next to Near East populations (Druze, Palestinian and Bedouin), highlighting a genetic resemblance already reported in previous studies [29]. A distinctive pattern of separation arises within Africa, with Biaka Pygmy that separated in ancient time from Yoruba and Mandenka. This is interesting because, even though Mandenka, Yoruba and Baika Pygmy experienced quite different historical dynamics and lifestyle, when compared with other worldwide populations, they cannot be genetically distinguished from each other [29]. The separation pattern that is shown here for African populations depicts what have already been reported in previous works using simulations methods [30-32]. The divergence times we obtained are also in agreement with what have been estimated by Gronau et al. [11]; they estimated indeed a separation time between Europe and Africa of 38-64 Kya (our estimate is –62 Kya) and between Europe and Asia of 31-40 Kya (our estimate is ~36 Kya).

The results of the simulation framework we developed to test the power of this method in correctly estimating the divergence time between populations are shown in (Figure 2B). In general, the forward simulations show that the method implemented in NeON exhibit a reasonably good power in estimate the real time of divergence. This is particularly true when the two populations have the same size through time (scenario 1); in this case indeed all the estimates (dots) fall in a range spanning 200 generations (grey lines) around the real value (black line), with an extremely precise estimation for recent splits. When one of the two populations experienced a bottleneck (scenario 2) the divergence time is a bit overestimated, except for recent and ancient separations (namely below 500 and above 1700 generations). On the contrary, when migration is taken into account (scenarios 3 and 4) the divergence time is underestimated, but only for splits more ancient than 1,000 generations.

Discussion

Human effective population size represents the average effect of drift across generation and so it is related with the level of population differentiation, and allows one to understand how populations evolved through time, whether they expanded or experienced drastic
Figure 3: Plots of the effective population size trough time. The x-axis represents the time (measured in generations) from the present (on the left) to the past (on the right). The y-axis represents the effective population size values. The continuous lines correspond to the median values of the Ne, dashed lines correspond to the 5th and 95th percentile of the Ne distribution.
reductions [33]. Because the effect of drift accumulated through time, a direct measure of Ne from census data is problematic. Advances in genome technology have facilitated the extensive genome-wide survey of densely spaced single nucleotide polymorphisms (SNPs), now available for many human populations [17,34,35]. This high-density genetic information can be used to estimate population genetics and evolutionary parameters that played a role in shaping today’s genome variation (including recombination rate [36], level of population differentiation [37], both useful to infer past populations dynamics or demographic events [38]. A way to study these historical processes exploits the information contained in the pattern of LD between markers, which depends both on intrinsic cellular factors as mutation, recombination or gene conversion and on extrinsic evolutionary aspects of populations as selection, migration, and effective population size [39]. In this paper we have introduced a new tool to infer the history of effective population size from patterns of linkage disequilibrium of genome-wide single nucleotide polymorphisms data. Using the functions contained in the NeON package we showed how it is possible to estimate the past demographic dynamics and the divergence time of the populations genotyped in the CEPH-panel; moreover, our simulation framework showed that the divergence times so calculated can be generally considered well estimated, especially when the two populations diverged quite recently.

Other than clarify aspects of the biological evolution, the inference
of populations’ demographic parameters as the degree of relatedness between human populations can also help to assess the presence and the extent of the interaction between biological and phenotypic or cultural variables. To give some examples, the so calculated divergence times can be correlated with those estimated from polymorphisms and cranial shape variables of human populations, to find evidence supporting a specific process of dispersal of early modern humans out of Africa [40], or compared with an estimation of linguistic split times, to test whether the parallelism between biological evolution and language diversification, firstly proposed by Charles Darwin [41] and subsequently verified with empirical data [42,43], has been originated by the same demographic dynamics.

The method embedded in the functions of this R package can improve and/or integrate other methods developed to estimate the effective population size or the degree of relatedness among populations from genomic data (e.g. Treemix [44], and the PSMC [12] or its extension MSMC [45]). Respect to Treemix, NeON has the advantage of estimate the time of separation between populations other than their relationships, but does not take into account migrations, whereas, respect to the sophisticated MSMC, the method presented here does not require phased data from multiple genomes, and can hence be suitable when only SNP panels are available.

Although being aware that the SNP panels suffer of ascertainment bias (resulting even in biased estimates of 18% downward [14]), the method we proposed here has already been successfully applied to the study of human evolution [17]. However, the lack of user-friendly programs strong limited its application to the study of real populations. With the NeON package, developed for the widely used R environment, this method can now be easily applied to analyze past population dynamics, giving the opportunity to shed light to different aspects of human population history [40]. The package NeON, together with tutorial and examples, is available for download and installation from CRAN website (http://www.r-project.org) with the license of GPL (>2), or from University of Ferrara, Population Genetics group’s website (http://www.unifi.it/dipartimento/biologia-evoluzione/ricerca/evoluzione-e-genetica/software/). It requires the package psych (http://CRAN.R-project.org/package=psych Version=1.3.10) and PLINK executable [15].

Acknowledgement

This study was supported by the European Research Council ERC-2011-AdG_295733 grant (Langelin). We than GUIDO BARBUJANI for valuable suggestions and for critically reading the manuscript, Serena Tucci that helped us finding a name for the package, and the Population Genetics group of the University of Ferrara.

References

An early divergence of KhoeSan ancestors from those of other modern humans is supported by an ABC-based analysis of autosomal resequencing data. Mol Biol and evol 29: 617-639.


2.3.2. Article: Genetic structure of the population of the Silk Road
Genetic landscape of populations along the Silk Road: admixture and migration patterns

Massimo Mezzavilla1,2, Diego Vozzi1, Nicola Pirastu2, Giorgia Girotto2, Pio d’Adamo2, Paolo Gasparini1,2 and Vincenza Colonna3*

Abstract

Background: The ancient Silk Road has been a trading route between Europe and Central Asia from the 2nd century BCE to the 15th century CE. While most populations on this route have been characterized, the genetic background of others remains poorly understood, and little is known about past migration patterns. The scientific expedition “Marco Polo” has recently collected genetic and phenotypic data in six regions (Georgia, Armenia, Azerbaijan, Uzbekistan, Kazakhstan, Tajikistan) along the Silk Road to study the genetics of a number of phenotypes.

Results: We characterized the genetic structure of these populations within a worldwide context. We observed a West-East subdivision albeit the existence of a genetic component shared within Central Asia and nearby populations from Europe and Near East. We observed a contribution of up to 50% from Europe and Asia to most of the populations that have been analyzed. The contribution from Asia dates back to ~25 generations and is limited to the Eastern Silk Road. Time and direction of this contribution are consistent with the Mongolian expansion era.

Conclusions: We clarified the genetic structure of six populations from Central Asia and suggested a complex pattern of gene flow among them. We provided a map of migration events in time and space and we quantified exchanges among populations. Altogether these novel findings will support the future studies aimed at understanding the genetics of the phenotypes that have been collected during the Marco Polo campaign, they will provide insights into the history of these populations, and they will be useful to reconstruct the developments and events that have shaped modern Eurasians genomes.

Keywords: Silk Road, Admixture, Migration patterns, Central Asia, Effective population size, Georgia, Armenia, Azerbaijan, Uzbekistan, Kazakhstan, Tajikistan

Background

The ancient Silk Road has been a trading route for several centuries in the past (2nd century BCE - 15th century CE) serving as a main connection between Europe and Asia. The route traverses a geographical region that was central during the human expansion from Africa [1,2], however its role during the stages of human evolution and its recent genetic history have not been fully clarified. At any rate, the complicated demographic history of the Silk Road populations must have left a trace on the patterns of genetic variation in a very unique way. Previous studies based on uniparental markers revealed extensive admixture between Europeans and eastern Asians in central Asia, increased genetic diversity and correlation with linguistic patterns [3-5]. Similar results were observed in a study based on genomic data from twenty-seven microsatellites genotyped in 767 individuals from Uzbekistan Kyrgyzstan and Tajikistan [6]. This study reported a pattern of diffused admixture and migration among populations in the area, resulting in great genetic diversity correlated with patterns of linguistic diversity, as it is often the case [7]. Recently a more comprehensive study disentangled the different levels of gene flow in worldwide populations but did not include a number of populations form Central Asia, for which migration patterns are still not fully understood [8].

With the aim of studying the genetics of several phenotypes (e.g. food preferences, hearing and taste [9-12]),
the Marco Polo scientific campaign, collected in 2010 genetic and phenotypic information of 411 individuals from six countries dwelling along the Silk Road (Georgia, Armenia, Azerbaijan, Uzbekistan, Kazakhstan and Tajikistan). Some of them have not been extensively described in population genetics literature, and the work we present here aims to investigate for the first time the population structures and the admixture patterns by using hundreds of thousands of genome-wide markers and 53 other worldwide populations as a reference for comparison. We confirmed results of previously mentioned studies, and extended the analysis to quantify admixture proportions and migration contributions. Here, we are providing insightful information about the differences of gene flow in these populations, highlighting the differences in population history and genetic structure.

Results and discussion

Great genetic diversity and east-west blocks in the Silk Road populations

We compared genetic data at 299,899 single nucleotide variants for 441 individuals from six populations along the Silk Road (abbreviated in SR from now on) with 943 individuals from 53 populations worldwide from the Human Genome Diversity Project (HGDP) panel [13]. To make the discussion of the results easier we grouped HGDP populations by their geographical area, and we will refer to the populations form Central Asia newly genotyped in this study as Silk Road, even if they are not exhaustively representative of the Silk Road (Additional file 1: Table S1).

First, we explored global patterns of genetic variation in our study populations. In the principal component analysis (PCA) of all 53 populations (Additional file 2: Figure S1), both first and second components distribute SR populations along the Asia-Europe gradient, and close to the nearby populations from Central Asia. SR populations, especially Uzbekistan and Kazakhstan, span over a region of the plot only slightly smaller in size than the one separating Europe from East Asia, suggesting great genetic diversity in these populations. ADMIXTURE analysis [14] on the same set of populations (Additional file 2: Figure S2), revealed, under the most likely hypothesis of ten clusters, very little or no contribution from Africa, Oceania and America and thus, we repeated analyses after removing these populations. The resulting PCA plots (Figure 1A and Additional file 2: Figure S3) show two different patterns for the Western and the Eastern SR. The Western SR (WSR) including Armenia, Georgia, and Azerbaijan shows proximity with Europe and the Near East, while the Eastern Silk Road (ESR) including Uzbekistan and Tajikistan and Kazakhstan shows a proximity generally closer to Asia, except for a number of individuals from Uzbekistan and Kazakhstan who are closer to Europe and the WSR. This pattern does not show a single cluster for populations from the Near East (Figure 1A). As expected, the Silk Road populations form Central Asia newly genotyped in this study as Silk Road, even if they are not exhaustively representative of the Silk Road (Additional file 1: Table S1).

First, we explored global patterns of genetic variation in our study populations. In the principal component analysis (PCA) of all 53 populations (Additional file 2: Figure S1), both first and second components distribute SR populations along the Asia-Europe gradient, and close to the nearby populations from Central Asia. SR populations, especially Uzbekistan and Kazakhstan, span over a region of the plot only slightly smaller in size than the one separating Europe from East Asia, suggesting great genetic diversity in these populations. ADMIXTURE analysis [14] on the same set of populations (Additional file 2: Figure S2), revealed, under the most likely hypothesis of ten clusters, very little or no contribution from Africa, Oceania and America and thus, we repeated analyses after removing these populations. The resulting PCA plots (Figure 1A and Additional file 2: Figure S3) show two different patterns for the Western and the Eastern SR. The Western SR (WSR) including Armenia, Georgia, and Azerbaijan shows proximity with Europe and the Near East, while the Eastern Silk Road (ESR) including Uzbekistan and Tajikistan and Kazakhstan shows a proximity generally closer to Asia, except for a number of individuals from Uzbekistan and Kazakhstan who are closer to Europe and the WSR. This pattern does not show a single cluster for populations from the Near East (Figure 1A). As expected, the Silk Road populations form Central Asia newly genotyped in this study as Silk Road, even if they are not exhaustively representative of the Silk Road (Additional file 1: Table S1).

First, we explored global patterns of genetic variation in our study populations. In the principal component analysis (PCA) of all 53 populations (Additional file 2: Figure S1), both first and second components distribute SR populations along the Asia-Europe gradient, and close to the nearby populations from Central Asia. SR populations, especially Uzbekistan and Kazakhstan, span over a region of the plot only slightly smaller in size than the one separating Europe from East Asia, suggesting great genetic diversity in these populations. ADMIXTURE analysis [14] on the same set of populations (Additional file 2: Figure S2), revealed, under the most likely hypothesis of ten clusters, very little or no contribution from Africa, Oceania and America and thus, we repeated analyses after removing these populations. The resulting PCA plots (Figure 1A and Additional file 2: Figure S3) show two different patterns for the Western and the Eastern SR. The Western SR (WSR) including Armenia, Georgia, and Azerbaijan shows proximity with Europe and the Near East, while the Eastern Silk Road (ESR) including Uzbekistan and Tajikistan and Kazakhstan shows a proximity generally closer to Asia, except for a number of individuals from Uzbekistan and Kazakhstan who are closer to Europe and the WSR. This pattern does not show a single cluster for populations from the Near East (Figure 1A). As expected, the Silk Road populations form Central Asia newly genotyped in this study as Silk Road, even if they are not exhaustively representative of the Silk Road (Additional file 1: Table S1).
not reflect the sampling strategy (rural communities versus cities, see Methods) but rather true genetic diversity; indeed samples closer to Europeans are from the same sampling locations of samples in the other block. Admixture analysis on the same reduced set of populations, revealed more details. The most likely hypothesis of seven clusters (Figure 1B, see results for different number of clusters in Additional file 2: Figure S4) shows that, overall, populations of the SR have a similar amount of European ancestry (red in Figure 1B) suggesting a common genetic legacy between Europe and the SR. An exception is found for some individuals in Uzbekistan and Kazakhstan where this component is >50%; these individuals cluster with Europe in the PCA. There are instead two opposite gradients in the WSR and the ESR for components shared with Near East and Central-South/East Asia. In the WSR we observe predominance of the Near Eastern component (pink and light green in Figure 1B), whereas in the ESR there is a predominance of Asian components (shades of blue in Figure 1B). Notably the dark blue component, which is diffused within the ESR, is predominant in the Kalash isolate [15] (Additional file 2: Figure S4), suggesting a common ancestral origin. The overall picture suggests that, beside a shared genetic pattern, there are two blocks of populations, and the ESR populations tend to be more admixed than the WSR ones.

The hypothesis of the two blocks, is further corroborated by the positions that the SR populations occupy in the tree of the populations’ divergence times estimates, as described in [16] (Figure 2) with the WSR close to Europeans, and the ESR between Central-South Asia and East Asia, with a large split time difference.

Next we explored the history of isolation and consanguinity in our set of populations. When calculating inbreeding coefficient, we did not observe particularly high values for the SR populations (Additional file 2: Figure S5). We evaluated the prevalence of stretches of homozygous regions (ROH) by comparing the cumulative distribution of the total length of ROH per individual in each population or group of populations, assuming a

![Figure 2 Populations divergence times.](image-url)
minimum ROH length of 2 Mb. Among the Silk Road populations, Tajiks have the longest stretches of homozygosity (Figure 3), higher than Europeans and East Asians (Additional file 2: Figure S6) and only lower than other founder populations, e.g. Surui in America, Papuan in Oceania and Kalash in Central-South Asia. In 80% of the Tajiks, the stretches of homozygosity sum up to 100 Mb whereas for other populations this value is below 30%. Because the length and the extension of ROH reflect the effective population size \(N_e\) [17,18], we confirmed our findings by evaluating \(N_e\) as described in [19-21] Tajikistan has one of the lowest \(N_e\) among SR populations, with a declining trend in the last 10,000 years (Additional file 2: Figure S7). However, for all SR populations, long-term estimates \(N_e\) is comparable to other reference populations (Additional file 2: Figure S8), suggesting high level of genetic variation in that area.

Migrations patterns revealed unidirectional gene flow from East Asia to Eastern Silk Road

We used two approaches to evaluate the role and effects of gene flow among populations in this study. Firstly, we assessed admixture for each population from any pair of other two populations using the three-population test for admixture [22,23]. All possible combinations of populations were performed, but we reported only results with Z-score < -5 in Additional file 1: Table S2. Overall, Z-score results indicate little or no gene flow in WSR, except for a number of contributions to Azerbaijan (mainly for Armenia and East Asia, see Additional file 1: Table S2), including contributions from American populations that we interpret more as a shared origin rather than an introgression. On the contrary, many migration events in the ESR are observed. Uzbekistan and Kazakhstan show the highest level of mixing in accordance with results from ADMIXTURE and PCA analyses (Figure 1), whereas Tajikistan received contributions mainly from Central-South Asian populations. We concluded that, as for admixture, we observed clearly different patterns of migrations in the WSR compared to the ESR, with more migration events taking place in the ESR, consistently to the degree of admixture that we inferred.

In order to look in more detail at migrations, we built a tree of all populations through ancestry graph analysis [23] (Figure 4). The tree structure recapitulates the known (existing) relationships among these populations, and highlights the great complexity of this geographical area, as previously shown by the neighbor joining tree, based on the time of divergence (Figure 2). The tree built with all the populations from HGDP reaches the maximum variance explained (99.7%) with the lowest standard error of residuals, with nine migration edges (Figure 4A). The tree shows the WSR populations compacted near Europe and the Near East, whereas the ESR populations are scattered (Figure 4A). Tajikistan falls with Central-South Asia; Kazakhstan with East Asia with

Figure 3 Cumulative distribution of runs of homozygosity. The x-axis indicates the total homozygosity in the genome in Megabases (Mb). Minimum ROH length was set to 2Mb.
a migration from Adygei; Uzbekistan is the target of a migration from the branch which connects East Asian and Americans.

To focus only on migrations relevant to the Silk Road populations, we built a tree removing Africa, Oceania and America (Figure 4B). The resulting tree with 9 migration edges explains 99% of the variance that was observed. Six migration edges connect East Asia with the SR. Indeed, with this analysis we pinpoint the gene flow from East Asia to the ESR and we estimate that the fraction of alleles that East Asia donated as source to other populations, ranges from 0.14 to 0.44, major recipient being Kazakhstan (Table 1). Tajikistan also received a consistent contribution (0.42) from Central-South Asia, thus highlighting the different history of this population within the ESR. As expected, we found gene flow from Europe to the WSR (0.39 towards Georgia and Azerbaijan) and a very low contribution of East Asia to Azerbaijan (0.05), but not to the other WSR populations. With these analyses we identified and quantified gene flow within the SR. Our results suggest unidirectional gene flow from East Asia to the ESR, and apparently this gene flow never reached the WSR, except for Azerbaijan, that was affected in a minor mode.

**To what extent populations admixed?**

Having assessed the general pattern of genetic relationships between populations, we next quantified admixture between populations. We first estimated the fraction of reciprocal contributions using ALDER [24] with 1 reference curve. In Figure 5 we summarized fractions of contributions from source populations to target. We estimated that the WSR received contributions from Europe and Near East in a measure of 40-50%. Armenia shows the highest level of Near East ancestry (48%) and Azerbaijan is the only population receiving significant contributions from the ESR (Kazakhstan 18%, Uzbekistan 35%). Among the ESR populations and in general, Uzbekistan and Kazakhstan are the more admixed populations with contributions of respectively 47% and 48% from Europe, and 49% and 47% from Central-South Asia. For these two populations, admixture within the SR is also high, ranging from 40% to 80%, and they also have the highest score of admixture (~80%) between themselves. Finally, East Asian components are higher in Kazakhstan compared to Uzbekistan and Tajikistan, as reported previously [25]. These results are consistent with ADMIXTURE analyses that we illustrated in a previous section. We also corroborated these observations by evaluating haplotype sharing among populations through chromosome painting. We used SupportMix [26] to assign haplotypes of individuals in SR populations to reference populations in the HGDP panel. In Additional file 2: Figure S9 we summarized the fraction of loci in target SR populations assigned to reference HGDP populations and

---

**Table 1 Migration edges estimates from Treemix**

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
<th>Allele sharing (standard deviation) in %</th>
</tr>
</thead>
<tbody>
<tr>
<td>East Asia</td>
<td>Kazakhstan</td>
<td>44 (0.8)</td>
</tr>
<tr>
<td>East Asia</td>
<td>Uzbekistan</td>
<td>27 (0.1)</td>
</tr>
<tr>
<td>East Asia</td>
<td>Tajikistan</td>
<td>14 (0.2)</td>
</tr>
<tr>
<td>Central-South Asia</td>
<td>Tajikistan</td>
<td>42 (2.3)</td>
</tr>
<tr>
<td>Europe</td>
<td>Georgia/Azerbaijan/ Near East/Armenia</td>
<td>39 (1)</td>
</tr>
<tr>
<td>East Asia</td>
<td>Azerbaijan</td>
<td>5.1 (0.2)</td>
</tr>
<tr>
<td>Europe</td>
<td>Near East</td>
<td>18 (0.9)</td>
</tr>
<tr>
<td>East Asia</td>
<td>Central-South Asia</td>
<td>18 (0.05)</td>
</tr>
<tr>
<td>East Asia</td>
<td>Near East</td>
<td>1.5 (0.3)</td>
</tr>
</tbody>
</table>
we observe that these are consistent with admixture trends estimated from ALDER and ADMIXTURE. Details of SupportMix analyses are reported in Additional file 1: Table S3.

To sum up, our data demonstrated poor admixture between the WSR and the ESR: only Azerbaijan seems to have a connection with the ESR. Within the WSR there is a non-negligible contribution of the Near East to Azerbaijan, but not to Georgia and Armenia, despite the three populations clustering together in the PCA analysis. Within the ESR, Tajikistan is generally less admixed than the other two populations. This could reflect a history of isolation and related low gene flow within the ESR, as suggested by the pattern of long stretches of homozygosity (Figure 3). Finally, Uzbekistan and Kazakhstan are the more admixed, both between themselves, and with respect to others, as hypothesized from PCA analyses (Figure 1).

When did admixture events take place?

We inferred the time in generations during which main migration events took place from patterns of linkage disequilibrium decay [24], and reported results that remain significant after Bonferroni’s correction in Additional file 1: Table S4. Within the WSR, we observed quite ancient events: a flow from Georgia and the Near East to Armenia ~210-220 generations ago (solid arrows in Figure 6), and a flow from the ESR and East Asia to Georgia ~60 generations ago (dot and line arrows in Figure 6). Among recent events (~25 generations ago, dashed lines in Figure 6) in the WSR, we observed contributions of East Asia to Azerbaijan (consistent with the results in the previous section); ESR populations instead are characterized by several introgressions from Central-South and East Asia.

The overall picture is of a flux around 25 generations ago from Central-South and East Asia towards the West that only reached Azerbaijan through the WSR, and this is consistent with the gradient of contributions of Central-South and East Asia described in Figure 5. Considering a generation time of 25 years, this scenario is compatible with the Mongolian Empire expansion, as can be observed in [8]. Finally, Tajikistan shows an additional event, even more ancient, involving Central-South Asia ~40 generations ago, which could explain the high level of haplotype sharing obtained using SupportMix compared to Uzbekistan and Kazakhstan, thus confirming a previous study based on Y chromosome [27].

Having demonstrated that indeed admixture took place among the populations that were taken into account for this study, and having assessed directions of migrations, thanks to these analyses, we now add extra information.

Figure 5 Admixture between pairs of populations. Heat map showing admixture contributions of source populations to targets inferred from linkage disequilibrium patterns. Rows are the target populations, columns are the source populations. Admixture with the same population is set at 100% and coded in red. Full lines indicate exact values and dashed line mark the 50%.
on times during which these events might have occurred. Time estimates should be interpreted with caution, but in combination with historical information might be useful to understand past events.

**Conclusions**

We provided a detailed description of some demographic events that shaped genetic diversity in some of the populations along the Silk Road. We observed high genetic heterogeneity in patterns of admixture and genetic structure. We identified a main subdivision in Western and Eastern Silk Road and an East-West gradient of East Asia contribution reaching only Azerbaijan within the West and taking place ~25 generations ago, a time compatible with the Mongolian expansion. A fine-scale resolution of these admixture events will be possible with whole genome sequencing of these populations. Our findings provide a catalogue of genetic variation that will help to better understand the phenotypic variations that we previously described in these countries [9-12,28] and provide information on time and extent of past migration events that can be integrated with historical studies.

**Methods**

**Ethic statement**

All communities belong to the Terra Madre organization (www.terramadre.org). Approval for information collection and processing has been obtained by the Ethical committee of the Maternal and Child Health Institute IRCCS-Burlo Garofolo Hospital (Trieste, Italy). All individuals signed an appropriate consent form (written in their local language) after they were instructed about the project, and all samples were completely anonymous.

In addition to the Italian Ethical committee approval we obtained the approval from the local Ethical committee and from the National Council on Bioethics in Georgia. For other countries where Ethical Committee were not present at the time of sampling we received authorization through official direct approval letters by National Authorities (Ministry of Research or Ministry of Health).

**Samples and genotypes**

Subjects in this study have been enrolled as part of the scientific campaign "Marco Polo", aimed at collecting social, genotype and phenotype information from populations of central Asia and Caucasus along the Silk Road. Populations in this study originate from West to East: Georgia, Armenia, Azerbaijan, Uzbekistan, Kazakhstan, Tajikistan and for the sake of simplicity we will collectively refer to them as the Silk Road populations, even if they are not fully representative of the entire range of countries along the Silk Road (Additional file 1: Table S1). Samples of saliva were collected in these countries from healthy donors from rural communities (Georgia, Azerbaijan,
Armenia, Tajikistan or form both rural communities and cities (Uzbekistan, Kazakhstan). DNA was extracted from samples that were collected, and it was used to determine genotypes at 624,851 single nucleotide polymorphic sites (SNPs) as described in a previous study [28]. For all subsequent analyses we used Plink v.1.07 [29] unless differently specified. Genomic coordinates refer to the GRCh37/hg19 version of the human genome sequence. Relatedness among pairs of individuals within populations was calculated as the proportion of genome that is identical by descent (\(\text{pi}_h\) in Plink) and one random individual per pair related above first or second cousins (\(\text{pi}_h > 0.25\)) was kept for subsequent analyses. After removing individuals that failed genotype quality controls and/or were related, the final sample consisted of 441 individuals. As reference populations, we used publicly available data from HGDP [13], after removing elated subjects as suggested in [30]. We merged genotype information from the Silk Road populations with HGDP genotypes at the 22 autosomal chromosomes filtering for minor allele frequency >1%, and genotyping success rate >97%. After quality controls, the merged data set consisted of 299,899 SNPs genotypes.

Population structure, effective population size and time of divergence estimates

Shared ancestry between populations was evaluated using ADMIXTURE v 1.22 and the number of clusters that better represent the data was established by cross-validation as described in [14]. Each ADMIXTURE run was replicated 5 time using different random seeds. Principal Component Analysis (PCA) analysis was performed using EIGENSOFT [31], considering the difference in sample size between our populations and HGDP ones. Our samples from the Silk Road were projected using the axis obtained from the HGDP populations. PCA and ADMIXTURE were performed on a linkage disequilibrium (LD) pruned dataset (\(r^2 < 0.4\)), consisting of 186267 SNPs.

Runs of homozygosity (ROH) were calculated using 299,899 SNPs. A ROH is defined as a chunk of genome \(\geq 2000\) kb long, containing at least 100 SNPs, with SNP density \(\geq 1\) per 50 kb. Within a ROH one heterozygous SNP and five missing calls are allowed. Two consecutive ROHs are considered as a single unit if their distance is \(\leq 1\) Mb. Coefficient of inbreeding was estimated using the function \(\text{het}\) implemented in PLINK. Demography and long-term effective population size were estimated by using linkage disequilibrium information as described in [16]. Since this method requires the genetic map position to be known, we used HapMap recombination map to assign a genetic position. Time of divergence between populations was estimated as described in [16] implemented in the R package NeON [32], a unrooted neighbour joining tree was built using the R package ape [33]. The generation time considered is of 25 years per generation.

Time of admixture and gene flow

Treemix v1.12 [23] was used to estimate maximum likelihood tree of our populations. Two different dataset were used: the first one with all the population and the second with population merged as continental groups without Africa, America and Oceania (see Additional file 1: Table S1 for group labeling). Zero to 12 migration edges were modeled, using blocks of 500 SNPs and the -se option was used to calculate standard errors. The tree was rooted with Yoruba when using the dataset with all the populations. Migration edges were added until the model explained 99% of the variance and also the migration edges were significant. Migration rate was estimated from Treemix run for the best tree selected. A Three-population test [22] was used to assess evidence of admixture in each of the six Silk Road population, values with a Z-score of less than -5 were collected. Level and time of admixture events were estimated using ALDER v.1.0.3 [24], considering the high number of possible comparisons and to avoid bias in allele frequencies due to small samples, we merged each reference population in continental groups.

Significance of the admixture was assessed after Bonferroni’s correction and consisted in the decay of linkage disequilibrium. In addition the level of admixture was assessed for each population using as reference for the population of the Silk Road only one population at the time. As an additional approach, we applied a machine learning method, SupportMix [26], to estimate genome-wide ancestry. This method allows to infer loci-specific genomic ancestry when analyzing different reference populations. SupportMix was run on phased datasets using a sliding window of 500 Kb for each chromosome and each individual. The mean percentage of the assigned chunk of genomes to each continental group was collected as well. Phasing was obtained using BEAGLE [34]. A mean value of assigned loci was calculated for each population of the Silk Road.

**Additional files**

**Additional file 1:** Table S1. Population description and sample size.  
**Table 2.** Three population test statistic. Only results with Z-score < -5 are reported and only when populations belong to different continental groups. Table S3. SupportMix mean percentage of assigned loci across all 22 chromosomes. Rows are the references, columns are the targets.  
**Table S4.** Time of admixture estimate using ALDER.  
**Additional file 2:** Figure S1. Principal component analysis of populations from the Silk Road newly genotyped in this study (in color) with populations in the Human Genome Diversity Project panel (in grey). Silk Road populations, and especially Uzbekistan and Kazakhstan, span.
over a large region of the plot, suggesting great genetic diversity among individuals. The first axis explain 5.6% of the variance, the second axis 4.3% of the variance. Figure S2. Admixture plots for the hypotheses of 8, 9 and 10 clusters (K). The most likely hypothesis (K=10) shows little or no contribution from Africa, America and Oceania to populations newly genotyped in this study. Figure S3. Principal Component Analysis of a subset of worldwide populations. Representation of components 3 to 6. Figure S4. Admixture plots for the hypotheses of 4.6 and 8 clusters (K) in a subset of populations. Figure S5. Inbreeding coefficients. Averages of individuals inbreeding coefficients (y-bars represent standard deviation) per population or group of populations. Figure S6. Cumulative distribution of runs of homozygosity. The x-axis indicate the total homozygosity in the genome in Megabases (Mb). We set the minimum ROH length to 2 Mb. Figure S7. Effective population size through time. The y-axis represent the effective population size (Ne), the x-axis the time in the past. Generation time is 25 years. Figure S8. Long term estimate of the effective population size. Figure S9. Proportions of admixture between pairs of populations inferred from haplotype sharing. On the y-axis the mean percentage (estimated across 22 autosomes) of genomic segment assigned to different continental groups (see colors in the legend) per each population on the x-axis.

Abbreviation
SR: Silk Road; N: Effective population size; ROH: Runs of homozygosity; WSR: Western Silk Road; ESR: Eastern Silk Road.

Competing interest
The authors declared that they have no competing interests.

Authors’ contributions
MM, VC, PG conceived and designed the study. GG, PG, NP, DV, PDA collected samples, MM analyzed the data. MM, VC wrote the paper. All authors read and approved the final manuscript.

Acknowledgments
We wish to thank all the subjects from the populations of Georgia, Armenia, Azerbaijan, Kazakhstan, Uzbekistan and Tajikistan who donated saliva samples for DNA analyses. We also wish to thank Irene Dall’Ara and Guido Barbujani for their contributions in the early stage of the project and Qasim Ayub and Chris Tyler-Smith for comments and suggestions on the manuscript. Finally, we thank two anonymous reviewers for helpful comments on the manuscript. The study was funded by the sponsors of the scientific expedition Marco Polo 2010, and the Ministry of Health of the region of Friuli Venezia Giulia.

Author details
1. Institute for Maternal and Child Health – IRCCS “Burlo Garofolo,” Trieste, Italy.
2. Department of Medical Sciences, University of Trieste, Trieste, Italy.
3. Institute of Genetics and Biophysics “A. Buzzati-Traverso,” National Research Council (CNR), via Pietro Castellino 111, 34131 Trieste, Italy.
4. Department of Medical Genetics, University of Trieste, Trieste, Italy.

Received: 17 June 2014 Accepted: 19 November 2014
Published online: 05 December 2014

References
29. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker P, Daly MJ: PLINK: a tool set for whole-genome


33. ape. [http://cran.r-project.org/web/packages/ape/index.html]


doi:10.1186/s12863-014-0131-6

Cite this article as: Mezzavilla et al: Genetic landscape of populations along the Silk Road: admixture and migration patterns. BMC Genetics 2014 15:131.
CHAPTER 3. Genetic history of the Armenians
3.1 Introduction

Observed patterns of human genetic diversity can be compared with models that include past population processes such as migration, divergence and admixture. These models require representatives of the ancestral populations and mostly consider present-day populations as direct descendants of the ancient inhabitants of a region. However, archaeological and genetic data reveal that human history has often been shaped by recent regional or localized population movements that can confound simple demographic models (VALLADAS et al. 2013; RAGHAVAN et al. 2014). Ancient DNA (aDNA) studies have also shown that the genetic landscape of the sites studied has been continuously shifting (RAGHAVAN et al. 2013; SKOGLUND et al. 2014), possibly triggered by environmental and cultural transitions. aDNA research is useful for understanding past demographic events; however, samples are limited and obtaining aDNA from warm climates remains a challenge. We have previously shown that studying genetic isolates also provides insights into human genetic variation and past demographic events (HABER et al. 2013). For example, by studying Jews, Druze, and Christians from the Near East, we showed the region had more genetic affinity to Europe 2,000 years ago than at present (HABER et al. 2013).

In this study, we investigate the Armenians, a population today confined to the Caucasus but who occupied Eastern Anatolia, reaching as far as the Mediterranean coast, up until the start of the 20th century (Figure 3.1).
Figure 3.1. Map of the Near East and surrounding regions. Map of the Near East and surrounding regions. The map shows location of modern Armenia and neighbouring countries. Blue lozenges show the recruitment sites for Armenian samples used in this study. Political turmoil during World War One resulted in the displacement of the East Turkey Armenian population (orange lozenge) to modern-day Armenia or to several other nearby countries such as Lebanon.

Political turmoil in the region during World War One resulted in the displacement of the Armenian population and to its restriction today to an area in Caucasus between the Black and Caspian seas. Armenians are an ethno-linguistic-religious group distinct from their surrounding populations. They have their own church, the Armenian Apostolic Church, which was founded in the 1st century CE, and became in 301 CE the first branch of Christianity to become a state religion. They have also their own alphabet and language which is classified as an independent branch of the Indo-European language family. The Armenian language is a subject of interest and debate among linguists for its distinctive phonological developments within the Indo-European languages and for its affinity to Balkan languages such as Greek and Albanian. The historical homeland of the Armenians sits north of the Fertile Crescent, a region of substantial importance to modern human
evolution. Genetic and archaeological data suggest farmers expanding from this region during the Neolithic populated Europe and interacted/admixed with pre-existing hunter-gatherer populations (Lazaridis et al. 2014). Furthermore, Armenia's location may have been important for the spread of Indo-European languages, since it is believed to encompass or be close (Anatolia or Pontic Steppe) to the Proto-Indo-European homeland from which the Indo-Europeans and their culture spread to Western Europe, Central Asia and India.

Previous genetic studies on Armenians are scarce and genome-wide analysis is limited to a few Armenian samples in broad surveys without any detailed analysis. Armenians were found to have genetic affinity to populations including the Jews, Druze and Lebanese Christians, in addition to showing genetic continuity with the Caucasus (Behar et al. 2010; Yunusbayev et al. 2012; Haber et al. 2013).

In this study, we analyse newly generated genome-wide data from Armenians as well as individuals from 78 other worldwide populations. We seek genetic signatures of past events such as the emergence of Armenians as an ethnic group, cultural changes in the Near East, and the expansions of ancient populations in this region.

### 3.2 Results and discussion

To study Armenians' genetic relationship to worldwide populations, we computed principal components using 78 populations (Table S3.1) and projected the Armenians onto the plot in a procedure called “PCA projection” (Patterson et al. 2006) (Figure 3.2A) which ensures that the PCA patterns are not affected by the large number of Armenians used in the analysis. We observe that Armenians form a distinctive cluster bounded by Europeans, Near Easterners, and the Caucasus populations. More specifically, Armenians are close to 1) Spaniards, Italians and Romanians from Europe; 2) Lebanese, Jews, Druze and Cypriots from the Near East; and 3) Georgians and Abhkassians from the Caucasus (Figure 3.2B).
Figure 3.2. Principal component analysis of >240,000 SNPs showing the top two components. A) The position of Armenians in a global genetic diversity sample based on 78 populations from 11 geographical regions. Armenians (173 individuals) were projected to the plot and therefore did not contribute to the observed global structure. B) A magnification shows that the Armenians (red) demonstrate genetic continuity with the Near East, Europe, and the Caucasus.

The position of the Armenians within global genetic diversity is unique and appears to mirror the geographical location of Anatolia (mostly modern day Turkey), which forms a bridge connecting Europe, the Near East and the Caucasus. Anatolia’s location and history have placed it at the centre of several modern human expansions in Eurasia: it has been inhabited continuously since at least the early Upper Palaeolithic (Kühn et al. 2009), and has the oldest known monumental complex built by hunter-gatherers in the 10th millennium BCE (Steadman and McMahon 2011). It is believed to have been the origin and/or route for migrating Near Eastern farmers towards Europe during the Neolithic (Pinhasi et al. 2005), and has also played a major role in the dispersal of the Indo-European languages (Renfrew 1990). Previous genetic studies have generally used Turks as representatives of ancient Anatolians. Our results show that Turks are genetically shifted towards Central Asians, a pattern consistent with a history of mixture with populations from this region. Turkish history is closely associated with the Seljuk Turks who originated north of the Caspian and Aral seas and moved to
Anatolia during the 11th century CE and probably admixed with the local population. These diversity patterns observed in the PCA motivated formal testing of admixture in Armenians and other regional populations.

To formally test for population mixture in Armenians we performed a 3-population test (PATTERSON et al. 2012) in the form of $f_3$ (Armenian; A, B) where a significantly negative value of the $f_3$ statistic implies that Armenians descend from a mixture of the populations represented by A and B, chosen from the 78 global populations. We found signals of mixture from several African and Eurasian populations (Table 3.1, Figure 3.3).

Table 3.1. Source populations and admixture time for Armenians

<table>
<thead>
<tr>
<th>Source 1</th>
<th>Source 2</th>
<th>$f_3$-statistics†</th>
<th>$z$-score</th>
<th>Time ± se</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>West Europeans</td>
<td>Sub-Saharan Africans</td>
<td>-0.0010508</td>
<td>5.0100</td>
<td>672.84</td>
<td>0.00E-08</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-</td>
<td>5554.64</td>
<td>4.10E-05</td>
<td></td>
</tr>
<tr>
<td>Caucasus populations</td>
<td>Sardinians</td>
<td>-0.0002515</td>
<td>4.8946</td>
<td>361.76</td>
<td>0.00E-36</td>
</tr>
<tr>
<td>Central and South</td>
<td></td>
<td>-</td>
<td>4886.28</td>
<td>3.00E-34</td>
<td></td>
</tr>
<tr>
<td>Asians</td>
<td>Sardinians</td>
<td>-0.0011018</td>
<td>16.261</td>
<td>271.32</td>
<td>0.00E-34</td>
</tr>
<tr>
<td></td>
<td>Arabic Peninsula</td>
<td>-</td>
<td>4812.08</td>
<td>8.10E-22</td>
<td></td>
</tr>
<tr>
<td>Caucasus populations</td>
<td>populations</td>
<td>-0.0004673</td>
<td>7.3205</td>
<td>381.64</td>
<td>0.00E-22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-</td>
<td></td>
<td></td>
<td>9.10E-05</td>
</tr>
<tr>
<td>Caucasus populations</td>
<td>North Levantines</td>
<td>-0.0002370</td>
<td>5.9960</td>
<td>4673.2 ± 343</td>
<td>0.00E-19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-</td>
<td>4376.12</td>
<td>7.60E-07</td>
<td></td>
</tr>
<tr>
<td>South Levantines</td>
<td>Caucasian populations</td>
<td>-0.0002351</td>
<td>5.0367</td>
<td>426.16</td>
<td>0.00E-19</td>
</tr>
<tr>
<td>Central and South</td>
<td></td>
<td>-</td>
<td>4233.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asians</td>
<td>West Europeans</td>
<td>-0.0003456</td>
<td>5.4263</td>
<td>1111.6</td>
<td>0.0029</td>
</tr>
<tr>
<td></td>
<td>Arabic Peninsula</td>
<td>-</td>
<td>3939.88</td>
<td>1.70E-17</td>
<td></td>
</tr>
<tr>
<td>North Europeans</td>
<td>populations</td>
<td>-0.0004331</td>
<td>5.1784</td>
<td>239.96</td>
<td>0.00E-32</td>
</tr>
<tr>
<td>Central and South</td>
<td></td>
<td>-</td>
<td>3908.52</td>
<td>2.10E-07</td>
<td></td>
</tr>
<tr>
<td>Asians</td>
<td>North Levantines</td>
<td>-0.0004840</td>
<td>6.9569</td>
<td>682.08</td>
<td>0.00E-07</td>
</tr>
</tbody>
</table>

† lowest $f_3$ result from each source region
Neolithic stone circles appear in Metsamor, Armenia.

Spread of agriculture throughout Europe from Near East.

Cycladic culture in the Aegean combining Anatolian and Greek elements.

Development of the wheel in Mesopotamia.

The trans-Caucasian culture spread, first evidence on the Armenian plateau.

The Halk Calendar begins (Legendary foundation of Armenia).

Otzi the Iceman dies

First cities developed in Mesopotamia

Babylonian influence predominant in Mediterranean regions

First writings, wheel carts appear

Domestication of the horse and appearance of the chariots.

Indo-European expansion to Anatolia, Europe and Central Asia

Late Bronze Age collapse 1206 and 1150 BCE

Greek dominance of the Aegean and the rise of the Hittite Empire.

Hellenistic period

Greco-Persian Wars

Armenia incorporated in Sassanid Empire

Armenia divided between the Ottoman Empire and Safavid empire in Iran 16th Century CE

Russia Empire annex Eastern Armenia 1828 CE

Armenian genetic structure 1429-1541 CE

Kingdom of Armenia 321 BCE to 428 CE
The most significantly negative f3 statistics are for mixture of populations related to Sardinians and Central Asians, followed by several mixtures of populations from the Caucasus, Arabian Peninsula, the Levant, Europe, and Africa. We sought to date these mixture events using exponential decay of admixture-induced linkage disequilibrium (LD). The oldest mixture events appear to be between populations related to Sub-Saharan Africans and West Europeans occurring ~3,800 BCE, followed closely by mixture of Sardinian and Caucasus-related populations. Later, several mixture events occurred from 3,000-1,200 BCE involving diverse Eurasian populations (Table 1, Figure 3). This period spans the Bronze Age, characterized by extensive use of metals in farming tools, chariots and weapons, accompanied by development of the earliest writing systems and the establishment of trade routes and commerce. Many civilizations such as in Ancient Egypt, Mesopotamia, and the Indus valley grew to prominence. Major population expansions followed, triggered by advances in transportation technology and the pursuit of resources. Our admixture tests show Armenians genomes carry signals of extensive population mixture during this period. We note that these mixture dates also coincide with the legendary establishment of Armenia in 2492 BCE. Admixture signals decrease to insignificant levels after 1,200 BCE, a time when Bronze Age civilizations in the eastern Mediterranean world suddenly collapsed, with major cities being destroyed or abandoned and most trade routes disrupted. This appears to have triggered the Armenians’ isolation from their surroundings, sustained later by the cultural/linguistic/religious distinctiveness that persists until today.

We compared patterns of admixture in Armenians to other regional populations. Unlike the Armenians, most Near Easterners have a recent
history of admixture with diverse populations. For example, we detect 7.9% (±0.4) East Asian ancestry in Turks from admixture occurring 812 (±168) years ago coinciding with the arrival of the Seljuk Turks in Anatolia from their homelands near the Aral sea. We also detect sub-Saharan African gene flow 840 (±84) years ago in Syrians, Palestinians and Jordanians, consistent with previous reports on recent gene flow from Africans to Levantine populations after the Arab expansions (MOORJANI et al. 2011; HABER et al. 2013)

The admixture pattern in Armenians appear close to patterns we have observed in genetic isolates in the region such Sephardic Jews and Lebanese Christians that show limited admixture with culturally different neighbouring populations in the last two millennia (HABER et al. 2013). Our tests suggest that Armenians had no significant mixture with other populations in their recent history and have thus been genetically isolated since the end of the Bronze Age, 3,000 years ago.

3.2.1 Structure of the Armenian population

To investigate the presence of genetic structure within the Armenian population, we performed model-based clustering on the values of the Armenian samples from the global PCA. We observe the following: 1- Armenians in the diaspora that trace their origin to historical Western Armenia (modern-day East Turkey) form one group (Figure 3.4, Cluster 1). 2- Armenians in modern-day Armenia (historical Eastern Armenia) are split in two major groups: 33% are in Cluster 1 and 57% form Cluster 2 (Figure 3.4). This structure could be the result of the Western Armenians' migration to the East after the events of 1915 CE that displaced the entire Western Armenian population. 3- A few Armenians recruited from Chambarak and Maykop (Republic of Adygea, Russia), form an outlier to the two major Armenian clusters (Figure 3.4, Cluster 3).
Figure 3.4. Genetic structure in Armenians. MCLUST classifies Armenians using a Bayesian Information Criterion into three clusters. Cluster 1 (blue) includes 95% of the Armenians that trace their origin to Western Armenia (East Turkey) (labelled W). Cluster 1 also includes 33% of the general Armenians (recruited from modern Armenia). Cluster 2 includes 57% of the general Armenians. Cluster 3 includes six Armenians recruited from Chambarak (modern-day Armenia) or Maykop (Republic of Adygea, Russia).

We investigated Armenian structure further by a procedure called “chromosome painting” (LAWSON et al. 2012) which reconstructs the haplotype of every individual (receiver) in our dataset using the haplotypes of other individuals (donors) in the dataset. We then constructed a tree that infers population relationships and similarities (Figure 3.5).
Figure 3.5. Population relationships from genome-wide haplotypes. Each tip of the tree corresponds to an individual; numbers of individuals are shown next to their population name at the tip of the branches. Numbers on branches show partition posterior probability. Armenians are shown in blue, forming two major clusters in a Near Eastern branch.
We found, similarly to our previous clustering methods results, fine genetic structure that splits Armenians into two major groups that are more similar to each other than to any other global population. We estimate from the LD patterns that divergence between the two major Armenian groups started 450-525 years ago. The dates coincide with the start of the Ottoman-Persian wars and the split of Armenia into West and East between the Ottoman Empire in Turkey and the Safavid Empire in Iran.

3.2.2 Relationship to ancient Europeans

One of the most studied demographic processes in population genetics is the Neolithic expansion of the Near Eastern farmers into Europe beginning ~8,000 years ago. Armenians’ location at the northern tip of the Near East suggests a possible relationship to the expanding Neolithic farmers. We merged our dataset with the genome of the Tyrolean Iceman, a 5,300-year-old individual discovered on the Italian part of the Ötztal Alps. We used TreeMix (Pickrell and Pritchard 2012) to construct a tree of genetic relationships using representative regional populations plus Armenians and Turks from the Near East. TreeMix uses a model that allows for both population splits and gene flow to better capture historical relationships between populations. We obtained a tree that recapitulates the known relationships among population groups. Furthermore, the tree shows the Iceman shared drift with Sardinians, as previously reported (Keller et al. 2012). We then ran TreeMix allowing it to infer only one migration event, and revealed gene flow from the Iceman to Armenians accounting for about 29% of their ancestry. The graph structure appears robust in 100 bootstrap replicates with the first migration (highest weight and lowest p-value) always going from the Iceman to Armenians (Figure 3.6).
Figure 3.6. Inferred population tree with one mixture event. The graph was inferred by TreeMix allowing one migration event. The migration arrow is coloured according to its weight; the weight is correlated with the ancestry fraction and shows that 29% of Armenian ancestry is derived from a population related to ancient Europeans. The graph is stable in 100 bootstrap replicates.

This structure was further investigated using outgroup f3 statistics (REICH et al. 2009; RAGHAVAN et al. 2014]). The expected value of f3 (Yoruba; Iceman, X) in the absence of admixture with Yoruba, will be a function of the shared genetic history of the Iceman and X (non-African populations). Most shared ancestry with the Iceman is with Sardinians and other Europeans (Figure 3.7). This is directly followed by shared ancestry with some Near Eastern populations: Cypriots, Sephardic Jews, Armenians, and Lebanese Christians. Other Near Easterners such as Turks, Syrians, and Palestinians show less shared ancestry with the Iceman. These results suggest that genetic isolates in the Near East - Cypriots (an island population), Near Eastern Jews and Christians (religious isolates), and Armenians (Ethno-linguistic isolate) - probably retain features of an ancient genetic landscape in the Near East that had more affinity to Europe than the present populations do.
Figure 3.7. Shared genetic drift between worldwide populations and the Tyloean Iceman, a 5,300-year-old European.
3.3 Conclusion

Armenians’ adoption of a distinctive culture early in their history resulted in their genetic isolation from their surroundings. Their genetic resemblance today to other genetic isolates in the Near East, but not to most other Near Easterners, suggests that recent admixture has changed the genetic landscape in most populations in the region. Armenians’ genetic diversity reveals that the ancient Near East had higher affinity to Neolithic Europe than it does now, and that Bronze Age demographic processes had major impact on the genetics of populations in this region.

The importance of populations like the Armenians is not only limited to the study of past demographic processes; isolated populations are emerging as a powerful tool for many different genetic investigations such as rare variant associations with complex phenotypes and the characterization of gene-environment interactions. Armenians emerged from founders in the Bronze Age accompanied by a long period of isolation that may have enriched rare disease alleles and therefore merit future medical exploration.

3.4 Material and Methods

Samples were collected from Lebanon (39), Chambarak (30), Deprabak (18), Gavar (12), Martumi (19), Yegvard (11), and Yerevan (9). Armenian individuals recruited in Lebanon traced their ancestry to East Turkey; they signed informed consent approved by the IRB of the Lebanese American University and were genotyped on Illumina 610K or 660K bead arrays. Data on 521,595 markers from 39 individuals are available at: Armenian subjects recruited from the present-day republic of Armenia signed consents approved by the ethical committee of the Maternal and Child Health Institute IRCCS-Burlo Garofolo Hospital (Trieste, Italy). Samples were genotyped on the Illumina HumanOmniExpress and the data are available upon request. Additional Armenian samples (35) were added along with 1509
samples from the literature that represent 78 worldwide populations (Li et al. 2008; Behar et al. 2010; Yunusbayev et al. 2012; Haber et al. 2013). PLINK (21) was used for data management and quality control. The required genotyping success rate was set to 99%, sex-linked and mitochondrial SNPs removed, leaving 300,899 SNPs. Principal components were computed with EIGENSOFT v 4.2 (Patterson et al. 2006) using 78 global populations, and the Armenian samples were projected onto the plot. The Bayesian Information Criterion (BIC) was computed by mclust (http://www.stat.washington.edu/mclust) over the first 10 principal components of the projected Armenian samples on the global PCA. The best model to classify the Armenians according to BIC is with three components (clusters) (Figure S1). Inference of population relations from haplotypes was assessed using Chromopainter (Lawson et al. 2012) with 10,000,000 burnin and runtime and 10,000 MCMC samples. A bifurcating tree of relationships amongst these populations was built using fineSTRUCTURE (Lawson et al. 2012) (Figure S2). Effective population size was estimated from linkage disequilibrium and time of divergence between populations was calculated using NeON with default parameters. Genetic distance (Fst) between populations was calculated using the software 4P (Benazzo et al. 2014). The generation time used was 28 years.

We used f3 statistics (24) f3(A; B,C) where a significantly negative statistic provides evidence that A is derived from admixture of populations related to B and C. We tested all possible f3 statistics in our dataset and calculated standard errors using blocks of 500 SNPs (Pickrell and Pritchard 2012). To date the time of admixture, we used ALDER (Lox et al. 2013) which computes the weighted LD statistic to make inferences about population admixture. The reference populations consisted of 1300 samples and 53 populations reduced from the original dataset after removing populations that are themselves highly admixed (Table S1). We collected results that are significant (z-score >|5|) and summarize the findings in Table 1 after pooling populations into respective geographical groups. Sardinians appear to have a distinctive
admixture pattern from other West Europeans and therefore are shown separately.

For tests of genetic affinity to Neolithic Europeans, we merged our samples with the genome of the Tyrolean Iceman (REICH et al. 2009). We downloaded BAM files mapped to hg18 and called all variants using GATK (MCKENNA et al. 2010). liftOver (http://genome.ucsc.edu) was used to convert the coordinates to hg19. The final dataset consisted of 91,115 SNPs.

We applied TreeMix (PICKRELL and PRITCHARD 2012) rooting the tree with a Denisovan genome, and standard errors were estimated using blocks of 500 SNPs. We generated 100 bootstrap replicates by resampling blocks of 500 SNPs to assess the stability of the tree topology. We used outgroup f3 statistics (REICH et al. 2009; RAGHAVAN et al. 2014) in the form of f3 (Yoruba; Iceman, X) to assess the shared genetic history of the Iceman with the modern populations. In the absence of admixture with Yoruba, deviation from 0 will be a function of the shared genetic history of the Iceman and the non-African population.

### Table S3.1: Populations selected for this study

<table>
<thead>
<tr>
<th>Geographic region</th>
<th>Population</th>
<th>Location</th>
<th>Reference for region/population</th>
<th>N</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>Bantu</td>
<td>Africa</td>
<td></td>
<td>18</td>
<td>Li et al., 2008</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Central African</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Biaka Pygmy</td>
<td>Republic</td>
<td>Sub-Saharan</td>
<td>22</td>
<td>Li et al., 2008</td>
</tr>
<tr>
<td></td>
<td>Egyptian</td>
<td>Egypt</td>
<td></td>
<td>12</td>
<td>Behar et al., 2010</td>
</tr>
<tr>
<td></td>
<td>Ethiopian</td>
<td>Ethiopia</td>
<td></td>
<td>19</td>
<td>Behar et al., 2010</td>
</tr>
<tr>
<td></td>
<td>Mandenka</td>
<td>Senegal</td>
<td>Sub-Saharan</td>
<td>22</td>
<td>Li et al., 2008</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Democratic Republic of</td>
<td>Sub-Saharan</td>
<td>13</td>
<td>Li et al., 2008</td>
</tr>
<tr>
<td></td>
<td>Mbuti Pygmy</td>
<td>Congo</td>
<td>Sub-Saharan</td>
<td>10</td>
<td>Li et al., 2008</td>
</tr>
<tr>
<td></td>
<td>Moroccan</td>
<td>Morocco</td>
<td></td>
<td>29</td>
<td>Li et al., 2008</td>
</tr>
<tr>
<td></td>
<td>Mozabite</td>
<td>Algeria (Mzab)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>San</td>
<td>Namibia</td>
<td>Sub-Saharan</td>
<td>6</td>
<td>Li et al., 2008</td>
</tr>
<tr>
<td></td>
<td>Yoruba</td>
<td>Nigeria</td>
<td>Sub-Saharan</td>
<td>21</td>
<td>Li et al., 2008</td>
</tr>
<tr>
<td>Near East</td>
<td>Armenian</td>
<td>Armenia, Lebanon</td>
<td>Armenian</td>
<td>173</td>
<td>This study, Behar et al., 2010, Yunusbayev et al., 2012</td>
</tr>
<tr>
<td></td>
<td>Bedouin</td>
<td>Israel (Negev)</td>
<td>Arabic Peninsula</td>
<td>46</td>
<td>Li et al., 2008</td>
</tr>
<tr>
<td></td>
<td>Carmel Druze</td>
<td>Israel (Carmel)</td>
<td></td>
<td>42</td>
<td>Behar et al., 2010</td>
</tr>
<tr>
<td></td>
<td>Cypriot</td>
<td>Cyprus</td>
<td>North Levant</td>
<td>12</td>
<td>Behar et al., 2010</td>
</tr>
<tr>
<td></td>
<td>Iranian</td>
<td>Iran</td>
<td></td>
<td>19</td>
<td>Behar et al., 2010</td>
</tr>
<tr>
<td></td>
<td>Jordanian</td>
<td>Jordan</td>
<td>South Levant</td>
<td>20</td>
<td>Behar et al., 2010</td>
</tr>
<tr>
<td>Region</td>
<td>Country</td>
<td>Continent</td>
<td>Population</td>
<td>Reference</td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>-----------</td>
<td>-----------------</td>
<td>------------</td>
<td>------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Lebanese</td>
<td>Lebanon</td>
<td>North Levant</td>
<td>25</td>
<td>Haber et al., 2013</td>
<td></td>
</tr>
<tr>
<td>Christian</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lebanese Druze</td>
<td>Lebanon</td>
<td>North Levant</td>
<td>25</td>
<td>Haber et al., 2013</td>
<td></td>
</tr>
<tr>
<td>Lebanese Muslim</td>
<td>Lebanon</td>
<td></td>
<td>25</td>
<td>Haber et al., 2013</td>
<td></td>
</tr>
<tr>
<td>Palestinian</td>
<td>Palestine</td>
<td>South Levant</td>
<td>45</td>
<td>Li et al., 2008, Behar et al., 2010</td>
<td></td>
</tr>
<tr>
<td>Saudi</td>
<td>Saudi Arabia, Bulgaria</td>
<td></td>
<td>20</td>
<td>Behar et al., 2010</td>
<td></td>
</tr>
<tr>
<td>Sephardic Jew</td>
<td>Portugal, Turkey</td>
<td>North Levant</td>
<td>19</td>
<td>Behar et al., 2010</td>
<td></td>
</tr>
<tr>
<td>Syrian</td>
<td>Syria</td>
<td>South Levant</td>
<td>16</td>
<td>Behar et al., 2010</td>
<td></td>
</tr>
<tr>
<td>Turk</td>
<td>Turkey</td>
<td>Turk</td>
<td>19</td>
<td>Behar et al., 2010</td>
<td></td>
</tr>
<tr>
<td>Yemenese</td>
<td>Yemen</td>
<td>Arabic Peninsula</td>
<td>8</td>
<td>Behar et al., 2010</td>
<td></td>
</tr>
<tr>
<td>Caucasus</td>
<td>Abhkasian</td>
<td>Caucasus, Caucasus</td>
<td>20</td>
<td>Yunusbayev et al., 2012</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Adygei</td>
<td>Russia</td>
<td>17</td>
<td>Li et al., 2008, Yunusbayev et al., 2012</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Balkar</td>
<td>Caucasus, Caucasus</td>
<td>19</td>
<td>Yunusbayev et al., 2012</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chechen</td>
<td>Caucasus, Caucasus</td>
<td>20</td>
<td>Behar et al., 2010</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Georgian</td>
<td>Georgia, Caucasus</td>
<td>20</td>
<td>Yunusbayev et al., 2012</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kumyk</td>
<td>Russia</td>
<td>14</td>
<td>Behar et al., 2010</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lezgin</td>
<td>Russia</td>
<td>18</td>
<td>Yunusbayev et al., 2012</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nogais</td>
<td>Russia</td>
<td>16</td>
<td>Yunusbayev et al., 2012</td>
<td></td>
</tr>
<tr>
<td></td>
<td>North Ossetian</td>
<td>Russia, Caucasus</td>
<td>15</td>
<td>Yunusbayev et al., 2012</td>
<td></td>
</tr>
<tr>
<td>Europe</td>
<td>Basque</td>
<td>France</td>
<td>West Europe</td>
<td>24</td>
<td>Li et al., 2008, Behar et al., 2010</td>
</tr>
<tr>
<td></td>
<td>Belorussian</td>
<td>Belorussia, North Europe</td>
<td>9</td>
<td>Yunusbayev et al., 2012</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bulgarian</td>
<td>Bulgaria, Central Europe</td>
<td>13</td>
<td>Behar et al., 2010</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chuvash</td>
<td>Russia</td>
<td>17</td>
<td>Yunusbayev et al., 2012</td>
<td></td>
</tr>
<tr>
<td></td>
<td>French</td>
<td>France</td>
<td>West Europe</td>
<td>28</td>
<td>Li et al., 2008, Behar et al., 2010</td>
</tr>
<tr>
<td></td>
<td>Hungarian</td>
<td>Hungary</td>
<td>Central Europe</td>
<td>19</td>
<td>Yunusbayev et al., 2012</td>
</tr>
<tr>
<td></td>
<td>Italian</td>
<td>Italy (North)</td>
<td>West Europe</td>
<td>21</td>
<td>Behar et al., 2010</td>
</tr>
<tr>
<td></td>
<td>Lithuanian</td>
<td>Lithuania</td>
<td>North Europe</td>
<td>10</td>
<td>Yunusbayev et al., 2012</td>
</tr>
<tr>
<td></td>
<td>Mordovian</td>
<td>Russia</td>
<td>North Europe</td>
<td>15</td>
<td>Yunusbayev et al., 2012</td>
</tr>
<tr>
<td></td>
<td>Orcadian</td>
<td>Orkney Islands</td>
<td>West Europe</td>
<td>15</td>
<td>Yunusbayev et al., 2012</td>
</tr>
<tr>
<td></td>
<td>Romanian</td>
<td>Romania</td>
<td>Central Europe</td>
<td>16</td>
<td>Yunusbayev et al., 2012</td>
</tr>
<tr>
<td></td>
<td>Russian</td>
<td>Russia</td>
<td>North Europe</td>
<td>27</td>
<td>Yunusbayev et al., 2012</td>
</tr>
<tr>
<td></td>
<td>Sardinian</td>
<td>Italy</td>
<td>Sardinian</td>
<td>28</td>
<td>Yunusbayev et al., 2012</td>
</tr>
<tr>
<td></td>
<td>Spaniard</td>
<td>Spain</td>
<td>12</td>
<td>Yunusbayev et al., 2012</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ukranian</td>
<td>Ukraine</td>
<td>North Europe</td>
<td>20</td>
<td>Yunusbayev et al., 2012</td>
</tr>
<tr>
<td>Central and South Asia</td>
<td>Balochi</td>
<td>Pakistan</td>
<td>Central and South Asia</td>
<td>24</td>
<td>Li et al., 2008</td>
</tr>
<tr>
<td>Region</td>
<td>Language</td>
<td>Country</td>
<td>Sample Size</td>
<td>Reference</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>----------</td>
<td>---------</td>
<td>-------------</td>
<td>----------------------------</td>
<td></td>
</tr>
<tr>
<td>South Asia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kurd</td>
<td></td>
<td>Kazakhstan</td>
<td>6</td>
<td>Yunusbayev et al., 2012</td>
<td></td>
</tr>
<tr>
<td>Makrani</td>
<td></td>
<td>Pakistan</td>
<td>25</td>
<td>Li et al., 2008</td>
<td></td>
</tr>
<tr>
<td>Malayan</td>
<td></td>
<td>India</td>
<td>2</td>
<td>Behar et al., 2010</td>
<td></td>
</tr>
<tr>
<td>North Kannadi</td>
<td></td>
<td>India</td>
<td>8</td>
<td>Behar et al., 2010</td>
<td></td>
</tr>
<tr>
<td>Paniya</td>
<td></td>
<td>India</td>
<td>4</td>
<td>Behar et al., 2010</td>
<td></td>
</tr>
<tr>
<td>Pathan</td>
<td></td>
<td>Pakistan</td>
<td>23</td>
<td>Li et al., 2008</td>
<td></td>
</tr>
<tr>
<td>Sakilli</td>
<td></td>
<td>India</td>
<td>4</td>
<td>Behar et al., 2010</td>
<td></td>
</tr>
<tr>
<td>Sindhi</td>
<td></td>
<td>Pakistan</td>
<td>24</td>
<td>Li et al., 2008</td>
<td></td>
</tr>
<tr>
<td>Tajik</td>
<td></td>
<td>Tajikstan</td>
<td>15</td>
<td>Yunusbayev et al., 2012</td>
<td></td>
</tr>
<tr>
<td>Turkmen</td>
<td></td>
<td>Turkmenistan</td>
<td>10</td>
<td>Yunusbayev et al., 2012</td>
<td></td>
</tr>
<tr>
<td>Uzbek</td>
<td></td>
<td>Uzbekistan</td>
<td>15</td>
<td>Behar et al., 2010</td>
<td></td>
</tr>
<tr>
<td>East Asia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Han</td>
<td></td>
<td>China</td>
<td>44</td>
<td>Li et al., 2008</td>
<td></td>
</tr>
<tr>
<td>Japanese</td>
<td></td>
<td>Japan</td>
<td>29</td>
<td>Li et al., 2008</td>
<td></td>
</tr>
<tr>
<td>North East Chinese</td>
<td>China</td>
<td>37</td>
<td>Li et al., 2008</td>
<td></td>
<td></td>
</tr>
<tr>
<td>North West Chinese</td>
<td>China</td>
<td>29</td>
<td>Li et al., 2008</td>
<td></td>
<td></td>
</tr>
<tr>
<td>South Chinese</td>
<td></td>
<td>China</td>
<td>65</td>
<td>Li et al., 2008</td>
<td></td>
</tr>
<tr>
<td>Yakut</td>
<td></td>
<td>Siberia</td>
<td>25</td>
<td>Li et al., 2008</td>
<td></td>
</tr>
<tr>
<td>Native America</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colombian</td>
<td></td>
<td>Columbia</td>
<td>5</td>
<td>Li et al., 2008</td>
<td></td>
</tr>
<tr>
<td>Karitiana</td>
<td></td>
<td>Brazil</td>
<td>7</td>
<td>Li et al., 2008</td>
<td></td>
</tr>
<tr>
<td>Maya</td>
<td></td>
<td>Native America</td>
<td>19</td>
<td>Li et al., 2008</td>
<td></td>
</tr>
<tr>
<td>Pima</td>
<td></td>
<td>Native America</td>
<td>9</td>
<td>Li et al., 2008</td>
<td></td>
</tr>
<tr>
<td>Surui</td>
<td></td>
<td>Native America</td>
<td>4</td>
<td>Li et al., 2008</td>
<td></td>
</tr>
<tr>
<td>Oceania</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Papuan</td>
<td></td>
<td>New Guinea</td>
<td>17</td>
<td>Li et al., 2008</td>
<td></td>
</tr>
<tr>
<td>Melanesian</td>
<td></td>
<td>Melanesia</td>
<td>8</td>
<td>Li et al., 2008</td>
<td></td>
</tr>
</tbody>
</table>
CHAPTER 4. Genetic analysis of taste and olfactory perception
4.1 Background

The main scope of the Marco Polo Project is to determinate the variants implicated in phenotype diversity, especially the chemosensory one. The main limitation of the Silk Road expedition is indeed the low sample size. Therefore there is a risk of detecting limited effects of variants on phenotype and, including also the need to stratify the sampled individuals, the result may be a small number of samples to be analyzed. However, to combine population informations and to use population-based approaches can be useful to understand phenotypic differences between populations. The population-based approach consists in considering all previous informations about genetic structure of the populations investigated. A population cannot be characterized only by PCA and ancestry methods, but also by its history and relationships due to migration patterns. Association analysis should be performed in the light of all these informations. Just think that even in the Silk Road population there are some individuals from the same country (e.g Kazakhstan) who have completely different ancestries (Mezzavilla et al. 2014). In addition the recent analyses of Armenian population give us hints to consider it as a separate population with a peculiar history, despite the first analyses gave us the idea that could be similar to Western Silk Road populations.

In this chapter the following studies will be investigated:

- A study of olfactory perception between different populations.
- A study on genetic variants implicated in the eye colour. Article: *Genetics of eye colours in different rural populations on the Silk Road.*
- A study on the impact of bitter taste perception in food liking in different countries across the Silk Road. Article: *A Population-Based Approach to Study the Impact of PROP Perception on Food Liking in Populations along the Silk Road.*
4.2 Olfactory perception

4.2.1 Introduction
Humans vary in their ability to smell numerous odors (BLAKESLEE 1918; AMOORE 1967), including those associated with food (AMOORE 1977; WYSOCKI and BEAUCHAMP 1984). Odor sensitivity is heritable (WHISSELL-BUECHY 1973; LISON et al. 1980; GROSS-ISSEROFF and LANCET 1988; KNAAPILA et al. 2012), with examples linking genetic variation for sensitivity to specific odors, typically located near olfactory receptor (OR) genes (KELLER et al. 2007; MENASHE et al. 2007; JAEGGER et al. 2010). The process of smell recognition could be divided in two steps: The first step is the recognition (using ~400 olfactory receptors (ORs) proteins (BUCK and AXEL 1991; SAIITO et al. 2009; NARA et al. 2011). After that a signal transduction cascade leads to the generation of a stimulus that reaches the brain (SCHILD and RESTREPO 1998; KLEENE 2008; BOZZA et al. 2009; MORI and SAKANO 2011). Considering a specific or a set of odorants we can observe hyposmia (diminished sensitivity) and specific hyperosmia (enhanced sensitivity) and a complete incapacity to sense a given odorant, that is called specific anosmia. Recent studies have shown a genetic association of sensitivity to specific odorants with identified OR genes (KELLER et al. 2007; MENASHE et al. 2007). As examples: a single nucleotide polymorphism (SNP) in the potassium voltage-gated channel KCNA3 (MIM# 176263) is associated with the altered general olfactory function (GUTHOFF et al. 2010). As previously shown odorant sensitivity average is a legitimate quantitative phenotype, with a potential genetic basis and is referred to as “General Olfactory Sensitivity” (GOS) (KEYDAR et al. 2013). GOS decrements are also associated with complex diseases such as Alzheimer's disease, Parkinson’s disease (MORLEY and DUDA 2010), schizophrenia (ATANASOVA et al. 2008), and autism (DUOVA et al. 2011), and in some cases serve as a biomarker. Recent works show as some genes are involved in olfactory compensation, making this phenotype very difficult to analyze, therefore as we have some variants, which could lower the GOS, many others could compensate for them.
We took advantage of the samples collected during the “Marco Polo” scientific expedition during which phenotypes related to smell perception were also collected. Smell perception is one of the most interesting and intriguing phenotypes. Perception of odors, intricately linked with taste, is of crucial evolutionary importance because it allows humans to distinguish between nonpoisonous and poisonous food, to relate to other individuals in order to identify potential mates (JACOB et al. 2002) or to facilitate mother recognition in newborn individuals (BUSH and VOSSHALL 2012). The ability of humans to recognize and discriminate among odorants displays differences between and within populations (KELLER et al. 2012; SOROKOWSKA et al. 2013). A number of studies describe associations between odor perception and OR genes in Caucasian populations (McRAE et al. 2012; McRAE et al. 2013). This study presents the first results obtained by from a smell-perception survey in populations living in Central Asia and Caucasus region.

4.2.2 Material and Methods
The olfactory test consisted in blind detection of nine odorants in the form of substances created by perfumers aiming at imitating natural scents as close as possible. Odorants identification was assessed using the "Sniffin' Sticks" kit (Burghart Medical Technology, Tinsdaler, Germany). Odorants in the kits resemble Lemon, Peppermint, Orange, Cinnamon, Fish, Coffee, Rose, Pineapple, Banana, Liquorice, Cloves and Leather. The last three (Liquorice, Cloves and Leather) were removed from the analysis because in all the populations more than 60% of individuals failed to identify them. Test participants were asked to smell odorants and choose among four possible answers; every correct answer counted as a score. This test served as a quick and easy first diagnostic test for odor identification related diseases. Smell perception was tested in 561 individuals divided in 8 countries. Thus we set a threshold for those individuals who failed to correctly identify any four out of nine odorants and we classified them as to as hyposmic in the following text as opposed to normosmic. Population clustering was performed combing Principal component analysis and Mcust algorithm taking the first 20 components. Association analysis was performed using PLINK (PURCELL et al. 2007), Polyphen and SIFT (FLANAGAN et al. 2010) were used to assess
the deleteriousness of the missense variation. Regression analysis on the significant loci was performed using the R package party.

4.2.3. Results
We found that hyposmic individuals represent a minority in all populations (range is 0.12-0.26; Figure S4.2.1) with the exception of Tajikistan where the incidence of hyposmia was double (0.48) and comparable to that of normal perception (t-test p-value 0.66). After dividing the population in different cluster (Figure 4.2.1 A-B). We found strong substructure in several of our populations. In addition to the identification a sub region of Tajikistan, Georgia and Armenia belong to different cluster.

![Figure 4.2.1 Genetic clustering of the Silk Road populations.](image)

Figure 4.2.1 Genetic clustering of the Silk Road populations. A) Principal component analysis using only population sampled during the Marco Polo expedition. Circles refer to the cluster identified using Mclust. B) Results of clustering and population subdivision, the dimension of the square refers to the number of individual that from one country (rows) falls into a Cluster (columns).

The distribution of the phenotype in each cluster is shown in Figure 4.2.2 Cluster 6 represents the Pamir region, an isolated area in the Tajikistan mountains and showed the highest frequency for hyposmic individuals.
Figure 4.2.2 Mosaic plot of the distribution of hyposmic individuals across clusters. The dimension of the box is proportional to the number of samples in each cluster.

Association analysis was performed after genetic clustering analysis for each cluster separately, after annotation we selected only the variants with $p$-value $< 1E^{-5}$ in at least two clusters, and with coherent effect among all of them. After this first step we selected the variants that are missense variation or belongs to the list of genes present in the Gosdb database (a database of GOS candidate genes based on both data mining and experimental research). Figure 4.2.3 shows our procedure.

Figure 4.2.3. Procedure used for collecting results
We found several missense variations associated but only two of them were found damaging/deleterious by SIFT and Polyphen predictors, interestingly, these two missense that were in linkage disequilibrium (r2=1) were found in the olfactory gene OR1L3, maps on chromosome 9, at 9q33.2. Regarding the Gosdb genes we found several markers inside these genes, but may of them are intronic, so their role is unknown. (Table S4.2.1).

Regression tree analysis show the interaction of the variants inside the most associated genes regarding the hyposmic phenotype. Among all the genes found OR1L3, CANCNG3 (Calcium Channel, Voltage-Dependent, Gamma Subunit 3) and ALDHB3 (aldehyde dehydrogenase 3 family, member B1) are the one with highest and significant p-value in the regression tree (See Figure 4.2.4)

![Figure 4.2.4](image)

**Figure 4.2.4. Regression tree analysis.** As show by the tree OR1L3 have a dominant effect, in which the majority of the carriers of C missense variation are hyposmic. In addition there are evidences than of and additive effects of CANCNG3 and ALDHB3.

4.2.4. Conclusion

This analysis show as GOS phenotype is a complex trait, in which the interaction of several genes could lead a different response. The novel aspect of this work is that exploring a wide set of population both admixed and isolated and using a population based approach we found two deleterious
variations in the gene OR1L3 that seems to be implicated in hyposmia. In addition, our analysis shows as other genes (present also in Gosdb database) that are implicated in hyposmic phenotype, providing more hints of the role of these genes. These results could be obtained only combining population genetic information with simple association methods.

Table S4.2.1 Association results

<table>
<thead>
<tr>
<th>Variation Name</th>
<th>PolyPhen prediction</th>
<th>SIFT prediction</th>
<th>Ancestral allele (ALL)</th>
<th>Minor allele</th>
<th>Type</th>
<th>genesymbol</th>
<th>chr</th>
<th>chrpos</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs2273863</td>
<td>benign</td>
<td>deleterious</td>
<td>G</td>
<td>T</td>
<td>rs2273863</td>
<td>LGALS8</td>
<td>1</td>
<td>236542978</td>
</tr>
<tr>
<td>rs6661946</td>
<td>benign</td>
<td>tolerated</td>
<td>T</td>
<td>T</td>
<td>rs6661946</td>
<td>HEATR1</td>
<td>1</td>
<td>236559808</td>
</tr>
<tr>
<td>rs4859571</td>
<td>benign</td>
<td>tolerated</td>
<td>T</td>
<td>T</td>
<td>rs4859571</td>
<td>NAAA</td>
<td>4</td>
<td>75936156</td>
</tr>
<tr>
<td>rs72773968</td>
<td>benign</td>
<td>tolerated</td>
<td>G</td>
<td>A</td>
<td>rs72773968</td>
<td>ERAP1</td>
<td>5</td>
<td>96803892</td>
</tr>
<tr>
<td>rs2294066</td>
<td>benign</td>
<td>deleterious</td>
<td>C</td>
<td>T</td>
<td>rs2294066</td>
<td>MYOM2</td>
<td>8</td>
<td>2098870</td>
</tr>
<tr>
<td>rs73571128</td>
<td>benign</td>
<td>deleterious</td>
<td>G</td>
<td>T</td>
<td>rs73571128</td>
<td>OR1L1</td>
<td>9</td>
<td>122661910</td>
</tr>
<tr>
<td>rs16912096</td>
<td>benign</td>
<td>tolerated</td>
<td>T</td>
<td>A</td>
<td>rs16912096</td>
<td>OR1L3</td>
<td>9</td>
<td>122675446</td>
</tr>
<tr>
<td>rs16912099</td>
<td>benign</td>
<td>tolerated</td>
<td>G</td>
<td>G</td>
<td>rs16912099</td>
<td>OR1L3</td>
<td>9</td>
<td>122675490</td>
</tr>
<tr>
<td>rs74634130</td>
<td>possibly damaging</td>
<td>deleterious</td>
<td>T</td>
<td>C</td>
<td>rs74634130</td>
<td>OR1L3</td>
<td>9</td>
<td>122675965</td>
</tr>
<tr>
<td>rs79636164</td>
<td>damaging</td>
<td>deleterious</td>
<td>C</td>
<td>T</td>
<td>rs79636164</td>
<td>OR1L3</td>
<td>9</td>
<td>122675965</td>
</tr>
<tr>
<td>rs2272094</td>
<td>benign</td>
<td>tolerated</td>
<td>A</td>
<td>A</td>
<td>rs2272094</td>
<td>CRTAM</td>
<td>11</td>
<td>122867553</td>
</tr>
<tr>
<td>rs17109924</td>
<td>benign</td>
<td>tolerated</td>
<td>T</td>
<td>C</td>
<td>rs17109924</td>
<td>LGR5</td>
<td>12</td>
<td>71584007</td>
</tr>
<tr>
<td>rs11541286</td>
<td>benign</td>
<td>tolerated</td>
<td>T</td>
<td>C</td>
<td>rs11541286</td>
<td>ZFC3H1</td>
<td>12</td>
<td>71614641</td>
</tr>
<tr>
<td>rs17855933</td>
<td>benign</td>
<td>tolerated</td>
<td>A</td>
<td>G</td>
<td>rs17855933</td>
<td>VEZT</td>
<td>12</td>
<td>95266406</td>
</tr>
<tr>
<td>rs17855934</td>
<td>benign</td>
<td>tolerated</td>
<td>T</td>
<td>G</td>
<td>rs17855934</td>
<td>VEZT</td>
<td>12</td>
<td>95300335</td>
</tr>
<tr>
<td>rs77215854</td>
<td>benign</td>
<td>tolerated</td>
<td>G</td>
<td>A</td>
<td>rs77215854</td>
<td>AHNK2</td>
<td>14</td>
<td>104940990</td>
</tr>
<tr>
<td>rs76611205</td>
<td>benign</td>
<td>tolerated</td>
<td>G</td>
<td>A</td>
<td>rs76611205</td>
<td>AHNK2</td>
<td>14</td>
<td>104941188</td>
</tr>
<tr>
<td>rs2586514</td>
<td>benign</td>
<td>tolerated</td>
<td>G</td>
<td>A</td>
<td>rs2586514</td>
<td>SLFN12</td>
<td>17</td>
<td>35422527</td>
</tr>
<tr>
<td>rs3803824</td>
<td>benign</td>
<td>tolerated</td>
<td>C</td>
<td>T</td>
<td>rs3803824</td>
<td>KIF2B</td>
<td>17</td>
<td>53823368</td>
</tr>
<tr>
<td>rs8135914</td>
<td>benign</td>
<td>tolerated</td>
<td>C</td>
<td>C</td>
<td>rs8135914</td>
<td>MICAL3</td>
<td>22</td>
<td>17818927</td>
</tr>
<tr>
<td>rs6433044</td>
<td>NA</td>
<td>NA</td>
<td>C</td>
<td>A</td>
<td>rs6433044</td>
<td>STK39</td>
<td>2</td>
<td>168208697</td>
</tr>
<tr>
<td>rs2268837</td>
<td>NA</td>
<td>NA</td>
<td>C</td>
<td>T</td>
<td>rs2268837</td>
<td>DGKG</td>
<td>3</td>
<td>186258459</td>
</tr>
<tr>
<td>rs17059082</td>
<td>NA</td>
<td>NA</td>
<td>C</td>
<td>C</td>
<td>rs17059082</td>
<td>GALNTL6</td>
<td>4</td>
<td>173009697</td>
</tr>
<tr>
<td>rs6832095</td>
<td>NA</td>
<td>NA</td>
<td>T</td>
<td>T</td>
<td>rs6832095</td>
<td>GALNTL6</td>
<td>4</td>
<td>173013905</td>
</tr>
<tr>
<td>rs10020676</td>
<td>NA</td>
<td>NA</td>
<td>G</td>
<td>G</td>
<td>rs10020676</td>
<td>GALNTL6</td>
<td>4</td>
<td>173024579</td>
</tr>
<tr>
<td>rs4331761</td>
<td>NA</td>
<td>NA</td>
<td>G</td>
<td>A</td>
<td>rs4331761</td>
<td>GALNTL6</td>
<td>4</td>
<td>173027973</td>
</tr>
<tr>
<td>rs888781</td>
<td>NA</td>
<td>NA</td>
<td>G</td>
<td>A</td>
<td>rs888781</td>
<td>SLIT3</td>
<td>5</td>
<td>169282063</td>
</tr>
<tr>
<td>rs2784923</td>
<td>NA</td>
<td>NA</td>
<td>C</td>
<td>C</td>
<td>rs2784923</td>
<td>SLIT1</td>
<td>10</td>
<td>97150097</td>
</tr>
<tr>
<td>rs17410931</td>
<td>NA</td>
<td>NA</td>
<td>C</td>
<td>T</td>
<td>rs17410931</td>
<td>SLIT1</td>
<td>10</td>
<td>97152419</td>
</tr>
<tr>
<td>rs4646812</td>
<td>NA</td>
<td>NA</td>
<td>G</td>
<td>G</td>
<td>rs4646812</td>
<td>ALDH3B1</td>
<td>11</td>
<td>68021094</td>
</tr>
<tr>
<td>rs105147</td>
<td>NA</td>
<td>NA</td>
<td>C</td>
<td>C</td>
<td>rs105147</td>
<td>ALDH3B1</td>
<td>11</td>
<td>68025546</td>
</tr>
<tr>
<td>rs9926562</td>
<td>NA</td>
<td>NA</td>
<td>T</td>
<td>C</td>
<td>rs9926562</td>
<td>CACNG3</td>
<td>16</td>
<td>24283998</td>
</tr>
<tr>
<td>rs8043733</td>
<td>NA</td>
<td>NA</td>
<td>G</td>
<td>G</td>
<td>rs8043733</td>
<td>ZNF423</td>
<td>16</td>
<td>49643601</td>
</tr>
</tbody>
</table>
Figure S4.2.1. Mosaic plot of the distribution of hyposmic individuals across countries. The dimension of the box is proportional to the number of samples in each country.
4.3. Publications
4.3.1. Article: Genetic of the eye colour
Genetics of eye colours in different rural populations on the Silk Road

Sheila Ulivi*, Massimo Mezzavilla and Paolo Gasparini

Eye colour is a highly transmissible and discernible trait in humans. A genome-wide association scan for variants associated to eye pigmentation was carried out on a large group of individuals coming from the Silk Road. Significant associations were detected not only with HERC2 \( (P\text{-value} = 4.99 \times 10^{-37}) \) and OCA2 \( (P\text{-value} = 4.51 \times 10^{-9}) \) genes but also with CTNNA2 gene \( (P\text{-value} = 4.06 \times 10^{-8}) \). Moreover, the multifactor dimensionality reduction analysis clearly showed the effect of HERC2 haplotype over OCA2 mostly associated with SNP, thus enabling a highly accurate eye-colour prediction. Finally, the regression tree analysis showed that individuals carrying a given combination of haplotypes have a significant probability to show a blue or green/grey iris colour as compared with brown, with a gradient from west to east.

**European Journal of Human Genetics (2013) 21, 1320–1323; doi:10.1038/ejhg.2013.41; published online 13 March 2013**

**Keywords:** eye colours; GWAS; silk road

**INTRODUCTION**

Eye irises’ colour is a common polygenic phenotype, which can vary between individuals.\(^1\) The iris is a thin diaphragm composed mostly of connective tissue and smooth muscle fibres regulating the amount of light entering the pupil.\(^2\) The colours, texture, and patterns of each person’s eyes are unique and are due to different factors, such as the density and structure of the iris stroma, the pigment epithelium, the pigment content within the melanocytes of the iris stroma, and the light-absorption properties of the melanin colouring skin, hair, and eyes.

Furthermore, also genetic factors have a relevant role with well-known genes such as OCA2 (OMIM: 203200), HERC2 (OMIM: 605837), which regulates OCA2 expression, and several other genes.\(^3\)

In this paper, we analysed this polygenic trait in several communities scattered along the Silk Road, to fill a gap of knowledge on genes underlying iris colour among people living in this part of the world.

**MATERIALS AND METHODS**

An overall number of 1015 individuals randomly selected, ranging from 8 to 84 years of age, were recruited during the ‘MARCO POLO’ scientific expedition (www.marcopolo2010.it) and divided as follow: (a) Western Silk Road (WSR); Crimea (peninsula of Ukraine) (102), Georgia (147), Armenia (174), and Azerbaijan (73) and (b) Eastern Silk Road (ESR); Uzbekistan (122), Tajikistan (118), Kazakhstan (60), and Kyrgyzstan (219). Sociodemographic information, as well as data on professional activity, lifestyle, eating habits, and family history, was obtained. A colour eye chart was used to classify iris colour. Phenotypes were defined as previously described\(^4\) (ie, blue = 0, intermediate (green, amber) = 1, and brown = 2).

Only 710 individuals out of 1015 gave a saliva sample. Before imputation, individuals with more than 5% missing genotypes and SNPs missing in more than 5% of samples were excluded. After doing the default filtering (call rate \( \geq 97\% \), \( R_{\text{QQ}} \geq 0.3 \), Hardy–Weinberg Equilibrium \( P\text{-value} \leq 1 \times 10^{-4} \), and minor allele frequency \( \geq 0.01 \)), 657 samples (WSR: 385; ESR: 272) and 2157485 directly genotyped and imputed SNPs passed quality controls. After genotyping (HumanOmniExpress; Illumina, Inc., San Diego, CA, USA), data were imputed using MACH (www.sph.umich.edu) to a common set of ~2.3 million autosomal SNPs based on LD patterns observed in Hap Map release 22 CEU samples (http://www.hapmap.org). In order to avoid the presence of close relatives in our data set, statistical analyses were performed using a mixed model regression, in which the kinship matrix is the random effect, as implemented in GenABEL for genotyped SNPs and ProbABEL for imputed data. Haplovew 4.2 (www.broadinstitute.org) and Plink v1.07 (http://pngu.mgh.harvard.edu) were used for haplotype construction, haplotype phases, and genetic-association studies. For studies on selection, a spatial ancestry analysis (SPA) was carried out to check for signal selection;\(^5\) in addition, we performed BAYESCAN\(^4\) to detect SNPs in the genomic region of CTNNA2, which show a higher level of population divergence than neutral loci; the decisive criterion that we adopted corresponds to a posterior probability \( > 0.99 \), which is indicative that a locus is under selection. A recursive partitioning tree-based analysis, using the R party package, was also performed to predict eye colour on the basis of HERC2 diploptypes and country of origin of each individual. For each node, we calculated the \( P\text{-value} \) associated after Bonferroni correction. Gene–gene interactions between analysed SNP and haplotype positions were tested using the multifactor dimensionality reduction (MDR) approach\(^6\) (software version 2.0 beta 8.4, www.epistasis.org).

**RESULTS**

The eye-colour distribution in the 1015 individuals (mean age 34.80, SD 17.01) turned out as follows: (a) ESR: 32 (6.2%) blue eyes, 35 (6.7%) green/grey, and 452 (87.1%) brown and (b) WSR: 46 (9.3%) blue eyes, 105 (21.2%) green/grey, and 345 (69.5%) brown. The eye-colour distribution in the 1015 individuals (mean age 34.80, SD 17.01) turned out as follows: (a) ESR: 32 (6.2%) blue eyes, 35 (6.7%) green/grey, and 452 (87.1%) brown and (b) WSR: 46 (9.3%) blue eyes, 105 (21.2%) green/grey, and 345 (69.5%) brown.

The genetic control (GC) inflation factor \( \lambda \) to check for stratifications in our populations was calculated being \( \lambda \) of 1.010115 and indicating the absence of strong stratification in our samples.

The GC inflation factor was calculated considering all individuals from each population as a unique group.

GWAS results for a first analysis\(^4\) are displayed in Table 1. The most significant association was obtained with two SNPs, known as being strongly correlated with both blue and brown eye colours, located
within HERC2 gene (rs1129038 and rs12913832; \( P \)-value = 4.99 \( \times 10^{-37} \)). Other significant associations (\( P \)-value < 5 \( \times 10^{-8} \)) were found with both CTNNA2 (OMIM: 114025) and OCA2 genes.

A second GWAS analysis was performed using the phenotype ‘brown versus not brown’ in a categorical way. In this case too, the most significant association was found with seven different SNPs located on HERC2 gene and one on OCA2 gene (Table 1).

According to SPA analysis, we found evidence of positive selection into the region of HERC2 and OCA2 genes (scores are over the top 2% in agreement with previous data). Interestingly, high SPA scores were also obtained for CTNNA2 gene. BAYESCAN results confirm these outcomes, namely, that SNPs in the CTNNA2 gene display evidence of positive selection.

Finally, looking at the most significant SNPs in HERC2 gene, we were able to construct a seven-marker haplo-block, which detected 10 different haplotypes in our populations. Six of them show a frequency > 5%: 5'-GGAAACAG-3' (27.2%), 5'-AAGGCCGA-3' (20.3%), 5'-GAAGCCGA-3' (20.1%), 5'-GAAGACAC-3' (17.7%), 5'-AAGGTTGA-3' (7.9%), and 5'-GGAAACAG-3' (5.5%). All of them had correlation with eye colour (0.036 > \( P > 2.05 \times 10^{-22} \)), haplotype 5'-AAGGCCGA-3' being the most associated one (\( P = 2.05 \times 10^{-22} \)).

The 10 haplotypes originated by recombination and/or transversion from two ancestral haplotypes: 5'-AAGGCCGA-3' (H1) and 5'-GGAAACAG-3' (H2).

<table>
<thead>
<tr>
<th>Gene</th>
<th>Name</th>
<th>Chr</th>
<th>Position</th>
<th>A1</th>
<th>A2</th>
<th>Coded_allele_freq</th>
<th>( P )-value</th>
<th>Pgc_value</th>
</tr>
</thead>
<tbody>
<tr>
<td>OCA2</td>
<td>rs7495174</td>
<td>15</td>
<td>28342238</td>
<td>G</td>
<td>A</td>
<td>0.3</td>
<td>4.51 ( \times 10^{-9} )</td>
<td>4.99 ( \times 10^{-9} )</td>
</tr>
<tr>
<td>HERC2</td>
<td>rs1129038</td>
<td>15</td>
<td>28356859</td>
<td>A</td>
<td>G</td>
<td>0.3</td>
<td>4.99 ( \times 10^{-37} )</td>
<td>7.96 ( \times 10^{-37} )</td>
</tr>
<tr>
<td></td>
<td>rs12959329</td>
<td>15</td>
<td>28359258</td>
<td>G</td>
<td>A</td>
<td>0.27549467</td>
<td>1.01 ( \times 10^{-9} )</td>
<td>1.13 ( \times 10^{-9} )</td>
</tr>
<tr>
<td></td>
<td>rs12913832</td>
<td>15</td>
<td>28365618</td>
<td>G</td>
<td>A</td>
<td>0.28082192</td>
<td>4.99 ( \times 10^{-37} )</td>
<td>7.96 ( \times 10^{-37} )</td>
</tr>
<tr>
<td></td>
<td>rs3935591</td>
<td>15</td>
<td>28374012</td>
<td>G</td>
<td>A</td>
<td>0.46183206</td>
<td>1.44 ( \times 10^{-13} )</td>
<td>1.69 ( \times 10^{-13} )</td>
</tr>
<tr>
<td></td>
<td>rs11636232</td>
<td>15</td>
<td>28386626</td>
<td>G</td>
<td>T</td>
<td>0.07810107</td>
<td>1.28 ( \times 10^{-10} )</td>
<td>1.45 ( \times 10^{-10} )</td>
</tr>
<tr>
<td></td>
<td>rs916977</td>
<td>15</td>
<td>28513364</td>
<td>G</td>
<td>A</td>
<td>0.48782344</td>
<td>3.29 ( \times 10^{-16} )</td>
<td>4.00 ( \times 10^{-16} )</td>
</tr>
<tr>
<td></td>
<td>rs1667394</td>
<td>15</td>
<td>28530182</td>
<td>A</td>
<td>G</td>
<td>0.49162861</td>
<td>4.92 ( \times 10^{-16} )</td>
<td>5.96 ( \times 10^{-16} )</td>
</tr>
<tr>
<td>CTNNA2</td>
<td>rs13431891</td>
<td>2</td>
<td>80605790</td>
<td>A</td>
<td>G</td>
<td>0.938271</td>
<td>4.06 ( \times 10^{-8} )</td>
<td>4.44 ( \times 10^{-8} )</td>
</tr>
</tbody>
</table>

The Silk Road, a well-known trading route of the ancient Chinese civilisation, has been an important way for cultural and commercial exchange between subjects living in East, South, and Western Asia with those settled in the Mediterranean and European regions for almost 3000 years. The collapse of the Mongol empire and the late fifteenth-century discovery of the sea route from Europe to Asia led to a progressive decline of the Silk Road’s trade in Central Asia. The geography of the Silk Road is a complex interaction between the

**DISCUSSION**

The Silk Road, a well-known trading route of the ancient Chinese civilisation, has been an important way for cultural and commercial exchange between subjects living in East, South, and Western Asia with those settled in the Mediterranean and European regions for almost 3000 years. The collapse of the Mongol empire and the late fifteenth-century discovery of the sea route from Europe to Asia led to a progressive decline of the Silk Road’s trade in Central Asia. The geography of the Silk Road is a complex interaction between the
different physical and climate areas featuring mountains, moorlands or grasslands, and river valleys and oases, often neighbouring uninhabitable deserts. Thus, populations could be dispersed (in the grasslands) or concentrated in the oases and river valleys. The majority of the population is of mixed Turkish descent. The Uighurs are the largest ethnic group, whereas Kyrgyz, Kazaks, Uzbeks, and Tartars are further strongly represented populations. From a linguistic point of view, different varieties of old Turkish are spoken.

The colour of the eye’s iris can vary dramatically between individuals, due to genetic differences. Present results confirm, also in these populations, the role of genes such as \( \text{HERC2} \) and \( \text{OCA2} \), already known to be involved in defining iris colour. Moreover, our findings demonstrate that not only \( \text{HERC2} \) and \( \text{OCA2} \) genes but also \( \text{CTNNA2} \) gene could be under selection pressure. As reported by Sturm and Duffy,\(^7\) eye colour is a feature that may fall under multiple selection pressures, including sexual, cultural, and environmental factors (ie, the level of sunlight). In this light, the presence of brown eyes, especially in populations living in the ESR, might be probably due to the combined action of both environmental and cultural factors.

One of the interesting outcomes of the present data is the demonstration on how genetic information can be used to predict eye colours. For example, homozygous individuals for a specific \( \text{HERC2} \) haplotype (H2) lead to a higher probability of carrying brown eyes in all populations, whereas carriers of the other haplotype (H1) have a significant probability to show a blue or green/grey iris colour. Furthermore, homozygous individuals for the same haplotype have a different probability to develop green/grey iris colour depending on the region in which they live (ie, a person belonging to the Caucasus region has a higher probability as compared with individuals living in Central Asia). An explanation for these findings is the possible presence of population-specific polymorphisms that might interact with \( \text{HERC2} \) and \( \text{OCA2} \) genes and thus contribute to the phenotypes. These polymorphisms could have a higher frequency in the Caucasus region because of a different history of gene flow between the various

---

**Figure 1** Results obtained from the partitioning analysis. (a) A regression tree with two nodes based on \( \text{HERC2} \) haplotypes and country (or region) of origin of each individual (ARM, Armenia; AZR, Azerbaijan; CRM, Crimea (peninsula of Ukraine); GEO, Georgia; KAZ, Kazakhstan; TJK, Tajikistan; UZB, Uzbekistan), most striking is the major node where we have the higher percentage of blue eyes (50.6%) in individuals carrying H1/H1. In the first node, the origin of population of the individuals carrying H1/H1 is not reported because the \( p \)-value is not significant. (b) Tree obtained from model-based recursive partitioning; East refers to populations from ESR, and West refers to those from WSR.

**Figure 2** MDR analysis. Representation of haplotype-SNP combinations among attributes considered in interaction model for brown versus not brown; the dark grey cells represent the genotype combinations associated with ‘high risk’ of having a specific eye colour, light grey cells are associated with ‘low risk’, and white cells mean lack of data. Each cell is a representation of the number of individuals with brown eyes (left bar) and not brown eyes (right bar).
populations. Apart from a significant relevance at population level, present findings might also be extremely useful in forensic medicine.

CONFLICT OF INTEREST
The authors declare no conflict of interest.

ACKNOWLEDGEMENTS
We thank Laura Esposito, Angela D’Eustacchio, Federico Fornasier, and Emmanouil Athanasakis for their technical support. We also thank Terra Madre organization and the Terra Madre communities who participated in the project. This work is part of the scientific activities carried out within the expedition Marcopolo2010–2012 and it has largely been benefited from funds coming from the sponsors of the scientific expedition.

2 Rennie IG: Don’t it make my blue eyes brown: heterochromia and other abnormalities of the iris. Eye (Lond) 2012; 26: 29–50.
4.3.2. Article: Bitter perception and food liking
A Population-Based Approach to Study the Impact of PROP Perception on Food Liking in Populations along the Silk Road

Antonietta Robino1,2*, Massimo Mezzavilla1,2, Nicola Pirastu1,2, Maddalena Dognini3, Beverly J. Tepper4, Paolo Gasparini1,2

1 Institute for Maternal and Child Health – IRCCS “Burlo Garofolo”, Trieste, Italy, 2 University of Trieste, Trieste, Italy, 3 Condotta Slow Food, Trieste, Italy, 4 Department of Food Science, School of Environmental and Biological Sciences, Rutgers University, New Brunswick, New Jersey, United States of America

Abstract

Taste is one of the main factors determining food choices. Differences in PROP bitter taste perception have been implicated in individual differences in food preferences and selection. The present study examined associations between, PROP phenotypes, self-reported food liking and TAS2R38 polymorphisms, the major gene implicated in PROP bitterness, in six different populations of the Caucasus and Central Asia, located along the ancient Silk Road. Differences in the distribution of PROP phenotypes across populations were detected, with a higher frequency of super tasters in Tajikistan (31.3%) and Armenia (39.0%) and a higher frequency of non tasters in Georgia (50.9%). While no relationships were observed between PROP phenotypes and food liking using standard statistical tests, we used an approach based on comparison of distance matrices derived from these data. The first matrix compared the food liking ratings of each population to all others pairwise using the Kruskal-Wallis test (at p < 0.00063), and the second one compared the distribution of PROP phenotypes across all populations in a similar manner calculating the chi-square statistic as a distance measure. A strong correlation between the two matrices was found (Mantel test: r = 0.67, p-value = 0.03), suggesting that the pattern of food liking across populations was closely related to the distribution of PROP phenotypes. This same relationship was not observed when TAS2R38 genotypes were substituted for PROP phenotypes in this analysis. Our data suggest that a population-based approach utilizing distance matrices is a useful technique for detecting PROP-related differences in food liking and can be applied to other taste phenotypes.

Introduction

Bitter taste perception is a variable trait both within and between human populations, and large individual differences in responsiveness to bitterness have been well documented [1]. Bitter perception in humans is mediated by a family of 25 TAS2R taste receptors [2]. Among them, the most studied is the TAS2R38 gene, associated with the ability to taste PTC (phenylthiocarbamide) and PROP (6-n-propylthiouracil) [3]. Approximately 75% of the world’s population are considered “tasters”, and perceive these substances as moderately to intensely bitter. These compounds are weak or tasteless for the remaining 25% of the population, who are considered “non tasters” [4–5]. Bartoshuk et al revealed that taster individuals can be divided into two sub-groups: medium tasters, who perceived moderate intensity from PTC/PROP, and super-tasters, who perceived these compounds as extremely bitter. Thus, the population distribution of non tasters, medium tasters and super tasters is approximately 25%, 50% and 25% respectively [6].

Sequence variations in the TAS2R38 gene produce three amino acid substitutions: A49P, A262V and V291I that define two common haplotypes, namely PAV and AVI. The AVI haplotype (AVI/AVI homozygous individuals) specifies the non taster phenotype, while it was supposed that the PAV haplotype (PAV/PAV homozygous or PAV/AVI heterozygous individuals) specifies the taster phenotype [7]. Although supertasting is typically associated with heightened responses to the bitterness elicited by PROP, TAS2R38 variations cannot explain “general” supertasting more broadly defined as the ability to perceive oral sensations more strongly without regard to PROP status or TAS2R38 polymorphisms [8].

Rare haplotypes (AAI, AAV, PAI, and PVI) have also been observed at a frequency of 1–5% [9], but are mainly found in African populations [10].

PTC and PROP are synthetic compounds, not found in nature, but they are chemically similar to isothiocyanates commonly found in broccoli, cabbage and other bitter-tasting Brassica vegetables [11]. The presence of the thiourea group (N-C=S) within these compounds is responsible for their bitter taste. Although the TAS2R38 receptor is also capable of binding non-thiourea substances (e.g., limonin, ethylpyrazine), compounds with the N-C=S moiety are considered the primary ligands for this receptor.
Although variation in _TAS2R38_ accounts for 50%–80% of the phenotype [7–14], recent data suggest that other genetic and non-genetic factors may play a role. This evidence includes a recently identified SNP (rs2274335) in the gustin (CA6) gene that has been implicated in taste bud growth and maintenance [15]. The presence of the major allele at this location in gustin is associated with greater taste bud densities, which are commonly observed on the tongue of supertasters [16]. Other evidence suggests that differential release of the salivary protein Ps-1 and II-2 that belong to the family of basic Proline-rich Proteins (bPRPs), may play a permissive role in PROP taste bitterness. Moreover, supplementation of Ps-1 to individuals who lack this protein enhances PROP bitterness intensity [16–17].

Greater perception of PROP is generally [18–19–20–21–22–23], but not always [24–25–26] associated with dislike and avoidance of _Brassica_ vegetables. There are also numerous reports that supertasters dislike bitter foods that do not contain the thiourea group, as well as other foods that produce strong oral sensations such as sweets, added fats, spicy foods and alcoholic beverages [19–20–21–22–27–28–29]. In light of these observations, PROP-tasting has gained attention as a general marker for oral sensations and food preferences [18]. This view remains controversial, however, since some studies report no relationship between PROP tasting and general food preferences [25–30] and other markers for oral sensations have emerged [31].

Despite these uncertainties, PROP-tasting has shown relationships with food intake and body weight variation that may have import long-term health implications. Specifically, studies showed that non taster women maintained higher body weights [32–33–34] and consumed more calories and high-fat foods [35–36] than supertaster women when offered a variety of foods in a buffet feeding regimen. Preliminary evidence from a mixed-gender, weight-loss intervention showed that almost half (47%) of the obese participants were non tasters as compared to the prevalence rate of 28–32% in the general population [37]. Finally, differences in the risk of colorectal cancer which is mediated, in part, by diet has been reported across _TAS2R38_ polymorphic groups [38–39].

Most population-based studies relating PROP-tasting to food and/or beverage selection have been conducted in Caucasian subjects residing in North America, Australia and Western Europe [19–23–28–40–41]. Notable exceptions are the study by Baranowski et al. [42] who studied White, Hispanic and Black children in the U.S.A. and the study by Mennella et al. [43] on African-American and non-Hispanic children and their mothers. Noteworthy works were also conducted in Asian populations [44–45–46]. There is a wealth of historic data on the diversity of PROP-tasting in different ethnic groups around the globe which shows that the frequency of non-tasting ranges from 3% to >40% worldwide [4].

However, data are sparse on the distribution of _TAS2R38_ polymorphisms in these same populations. Some recent investigations addressing this question include a genetic analysis of Indian-born Asian Indians living in the U.S.A. [47], and a study of genotype-phenotype relationships in a sample derived from West Central and East Africa [10]. However, neither study investigated PROP perception. _TAS2R38_ polymorphisms and food liking in different rural communities from the Caucasus region (Georgia, Armenia and Azerbaijan), Central Asia (Uzbekistan and Kazakhstan) and Tajikistan. Data were obtained as part of the scientific expedition Marcopolo 2010 (www.marcopolo2010.it), whose main goals were to analyse individual differences in the human senses (e.g. taste, smell, hearing, vision) across the Silk Road, a major pathway for cultural, commercial, and genetic exchange between individuals from China and Mediterranean countries for almost 3,000 years.

### Materials and Methods

#### Study population

A total of 496 subjects participated in the study (206 males and 290 females), coming from 20 different communities of six countries in the Caucasus and Central Asia: Georgia, Armenia, Azerbaijan, Uzbekistan, Kazakhstan and Tajikistan (Figure 1). Sample sizes from each of the communities are described in Table S1. All communities belong to the Terra Madre organization (www.terramadre.org). Information, such as age, sex, lifestyle, eating habits, professional activity, smoking and alcohol consumption were collected.

All subjects provided written informed consent before participation. Approval for the research protocol was obtained from the ethical committee of IRCCS-Burlo Garofolo Hospital.

#### Genotype analyses

Saliva samples were collected from all participants using the Oragene DNA collection kit and DNA was extracted (DNA Genotek, Ontario, Canada). Three polymorphisms in the _TAS2R38_ receptor gene (rs1726866, rs10246939 and rs713598) define the genotype. The first two were genotyped with the Omni Express 700K Illumina Chip. The third one was analysed using TaqMan probe-based assays (Applied Biosystems, Foster City, CA, USA).

#### PROP tasting

PROP taste intensity was determined in all subjects using a filter paper method previously described [48]. Each subject was given two paper disks, the first one was impregnated with 1.0 mol/l NaCl (VWR Scientific, Bridgeport, NJ), and the second disk was impregnated with 50 mmol/1 6-n-2-propylthiouracil (cat. no. P3755; Sigma-Aldrich, St Louis, MO). The subject was asked to rinse the mouth with bottled water, place the paper disk on the tip of the tongue and rate the intensity of the taste using the labelled magnitude scale (LMS). The subjects were also required to rinse with water between tasting each disk and to wait a minimum 30 s before tasting the PROP disk.

The LMS is a quasi-logarithmic 100-mm scale anchored with the labels ‘barely taste it’, ‘weak’, ‘strong’, ‘very strong’ or ‘strongest imaginable’ oral sensation [49].

For this study the LMS was translated in the local language of each community. In addition, translators verbally defined the label descriptors of the scale to each participant and also instructed him/her to make a mark anywhere on the scale, not only near the descriptors.

Using LMS numerical cut-off scores of <15 and >67, the subjects were classified as super tasters and non tasters, respectively. Medium tasters fell between those two limits (>16 and 67). NaCl ratings were used as a reference standard for classifying subject who gave a borderline rating to PROP. The use of NaCl as a reference standard is based on the observation that super tasters give higher ratings to PROP than NaCl, medium tasters give similar ratings to both, and non tasters give higher ratings to NaCl than to PROP [50]. These procedures were developed and validated in previous studies [48] and have been used in numerous areas.
investigations in English-speaking and non-English speaking populations followed in our previous studies [34–51–52].

Food Liking Questionnaire

Participants completed a 79-item food liking questionnaire that was based on an instrument used in a previous study [51] and supplemented with foods specific to the diets of the communities we studied. The original questionnaire focused on bitter-tasting, strong-flavored and high-fat foods that had been associated with PROP status in our previous studies [21–51]. The selection of the supplemental foods was based on a survey conducted by collaborators from the Terra Madre organization who carried out a preliminary survey on the local foods consumed by these populations [53]. The questionnaire assessed general food likes and dislikes (e.g. garlic, milk, banana, orange juice). It was administered in the local language of each community by translators who were familiar with the local culture.

Subjects rated their liking of each item on a 5-point scale ranging from “like extremely” (score 5) to “dislike extremely” (score 1). The option “never tasted” was also included.

Statistical analysis

The Chi-square test was used to examine the association between TAS2R38 genotypes and PROP status for the whole cohort. Chi-square tests were also performed to determine whether the relationship between TAS2R38 genotypes and PROP status differed among the populations tested. Correspondence Analysis [54] was also applied to the two-way contingency table of PROP status and participants’ country of residence to obtain a graphical representation of the relationship between the two variables.

Considering the potential relationship between PROP perception and food liking reported in the existing literature [18–19–20–21–28–29], analysis of covariance (ANCOVA) was performed to determine the influence of PROP taster status and TAS2R38 genotypes on liking of each food. This analysis was applied to the entire cohort and to each population separately. Sex and age were used as the covariates. Due to the large number of comparisons, statistical significance was set at p = 0.00063, following Bonferroni correction (p = 0.05 / 79 foods).

In addition, the foods were grouped using the same general categories as in Ulrich et al. [21] and the ANCOVA calculations described previously were repeated. The food groups included fruits (strawberries, lemons, orange juice), vegetables (artichokes, spinach, turnip, cooked carrots, asparagus, fava beans, cabbage), alcohol (red wine, white wine, vodka, brandy, beer), condiments (olives, sardines, onion, garlic, kilka, adgika, chilli pepper), sweets (ice cream, cake, sweet ricotta, biscuits, biscuits with cream, jam, honey, milk chocolate). The mean number of foods within each food group was calculated for each subject and was used for the analyses.

We also sought to determine if variations in food likes and dislikes across populations were related to the distribution of PROP phenotypes or TAS2R38 genotypes. To accomplish this task, a series of data matrices were constructed. First, the Kruskal-Wallis test was performed (at p < 0.00063) comparing the food liking of each population to all others, pairwise. The number of foods that showed statistically significant differences between population pairs were tallied and entered into a distance matrix.
Higher values indicated dissimilar patterns (large distances) in food liking between populations, and lower values indicated similar patterns (small distances) between them. For example, if the pairwise difference between two populations was high, these two populations had many differences in food liking. On the contrary, if the pair-wise difference was small, the two populations shared similar food liking responses.

In order to describe the phenotypic dissimilarities in bitter perception between populations, we created another distance matrix. Here, we calculated the chi-square statistic (as a distance measure) between phenotypic groups (non taster, medium taster and super taster) for each population, pairwise. Here, higher values represent a large difference (i.e., distance) in PROP bitterness between population pairs, and lower values represent a small difference in bitterness between population pairs. The data inputs and procedures for this analysis are similar to those of multiple correspondence analysis (MCA) [55] where data are categorical rather than continuous.

In order to assess possible bias due to the differences in sample size between populations, we performed a bootstrap analysis [56]. We constructed a series of distance matrices by repeatedly (1000 times) sampling 47 individual (the n of the smallest population) from each population. We compared each distance matrix built after bootstrapping with the original one (built using the full dataset) and found a high correlation between them (r>0.9), showing that differences in sample size did not affect our results.

Then, we calculated the Fst (Fixation Index) [57] to estimate genetic differences between populations for the SNPs which define TAS2R38 haplotypes. We also constructed a matrix of Fst values using the whole genome (~356,000 SNPs) to obtain a global estimate of genetic diversity in our sample. Pairwise Fst was performed using the R package Adegenet v1.3-4 [58]. Finally, the Mantel test [59] was used to determine the (dis)similarities between distance matrices. The Mantel r statistic is a standardized Pearson correlation coefficient calculated following random rearrangement of the data matrices across multiple permutations. 1000 iterations were used for a critical cut-off value of p<0.05.

Results and Discussion

PROP phenotypes and haplotypes

All 496 individuals were tested for PROP taste intensity. The distribution of PROP status in each population was analysed and is shown in Table 1. In the overall sample 37.0% of individuals were non tasters, 40.0% were medium tasters and 23.0% were super tasters. Interestingly, the distribution of phenotypes varied among the populations (X-squared = 42.1077, p-value = 7.1e-06). In particular, the prevalence of non tasters was higher in Georgia (50.9%) as compared to other populations, while the proportion of super tasters was higher in Armenia (39.0%) and Tajikistan (31.3%) relative to the other populations.

Correspondence Analysis revealed the relationships among the populations living in different countries with respect to PROP phenotype. In agreement with the univariate analyses, Georgia was highly associated with the non taster phenotype while Armenia was closely associated with the super taster phenotype. Furthermore, medium tasters were highly represented in the cluster of populations consisting of Azerbaijan, Uzbekistan and Kazakhstan. Tajikistan was distinct from the other groups (having relatively equal frequencies of the three taster phenotypes), although it was more closely associated with the super taster phenotype in accordance with the high prevalence of super tasters in this population (Figure 2).

These findings do not agree with a simple geographical explanation for the pattern of PROP phenotypes across populations. In particular, the phenotype differences between the populations of Armenia and Georgia were totally unexpected, because these two countries are closely located and have a long standing tradition of cultural and political exchange dating back to the Middle Ages, when the two countries were allied against the Muslim empire [60]. Differences in age, gender and smoking can influence PROP phenotypes [6–34–61–62]. However, these factors did not explain the variability across the populations studied here since our analyses adjusted for these factors. These data support recent findings suggesting that other genetic loci or non-genetic factors contribute to PROP tasting [15–17] and efforts to identify and fully characterize these factors should be ongoing goal.

In contrast to the phenotypic differences observed among populations, we found no differences in TAS2R38 haplotypes across populations (X-squared = 8.1822, p-value = 0.611) (Table 2). The AVI/AVI, AVI/PAV and PAV/PAV diplotypes accounted for 24.9%, 48.0% and 27.1%, respectively, of the overall sample, in agreement with the allelic frequencies typically reported in Caucasian populations [7]. As expected, there was a strong association between TAS2R38 diplotypes and PROP phenotypes (X-squared = 151.4019, p-value <2.2e-16). In the entire sample 82.9% of AVI/AVI homozygous individuals were non tasters, compared to 11.4% who were medium tasters and 5.7% who were super tasters. As expected, PAV/PAV homozygous and PAV/AVI heterozygous subjects were mainly medium or super tasters. We observed a similar correspondence between genotypes and phenotypes in each population. These data are reported in Table S2.

PROP phenotype and food liking

The relationship between PROP phenotype and liking for each food on the food liking questionnaire was examined for the entire cohort, and separately for each population, and no associations were found. A list of all foods with mean and standard deviation in the overall sample and in each population is reported in Table S3.

In addition, no relationship was revealed between PROP status and food preference groups. Data are shown in Table S4.
These same analyses were repeated for TAS2R38 haplotypes, and the outcome was the same; no associations were found.

Multi-dimensional analyses of food liking

A distance matrix describing the differences in food liking across the populations was constructed, and is graphically presented as a dendrogram in Figure 3. The figure shows three different groups: the first one composed only of Georgia, the second one composed of Uzbekistan, Kazakhstan and Azerbaijan and the third composed of Armenia and Tajikistan. It is clear that countries do not group according to geography, especially in the case of Armenia and Tajikistan.

We then determined if the PROP bitterness phenotypes could explain the observed clustering. Thus, we compared the two distance matrices (the PROP phenotype on one hand and the food liking on the other) and found a strong positive correlation between them (Mantel test: \( r = 0.67, p\text{-value} = 0.003 \)). The results of the Mantel test between each pair of distance matrices are summarized in Table 3.

In addition, distance matrices and their graphical representation are supplied in supplementary materials (Table S5 and Table S6, Figure S1).

We also tested if the TAS2R38 gene was associated with these groupings, and found no evidence of correlation (correlation = 0.02, p-value = 0.3) between the distance matrix of food liking and the matrix of genetic distance based on TAS2R38. In addition no correlation was found using the distance matrix based on the whole genome using ~356,000 SNPs.

These results have two important implications. First, they show that differences in food liking among populations strongly correlate with PROP taster status but not with TAS2R38 genotypes. This finding supports the view that polymorphisms in TAS2R38 primarily define the ability to taste PROP, but also recognizes that this gene is pleiotropic and influences multiple phenotypic

<table>
<thead>
<tr>
<th>Table 2. Distribution of TAS2R38 haplotype by sex and population.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TAS2R38 haplotype</strong></td>
</tr>
<tr>
<td>All (n = 496)</td>
</tr>
<tr>
<td>Sex</td>
</tr>
<tr>
<td>Males (n = 206)</td>
</tr>
<tr>
<td>Females (n = 290)</td>
</tr>
<tr>
<td>Population</td>
</tr>
<tr>
<td>Georgia (n = 116)</td>
</tr>
<tr>
<td>Azerbaijan (n = 47)</td>
</tr>
<tr>
<td>Uzbekistan (n = 91)</td>
</tr>
<tr>
<td>Kazakhstan(n = 57)</td>
</tr>
<tr>
<td>Tajikistan (n = 80)</td>
</tr>
<tr>
<td>Armenia (n = 105)</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pone.0091716.t002
traits such as the perception of non-thiourea, bitter and non-bitter tastes, other oral sensations, food liking, and downstream effects such as dietary behaviour and weight status [15–18–34]. We acknowledge, however, that PROP status may be one of several markers for chemosensory perceptions [31], and multiple markers may be involved. Moreover, post-receptor differences in peripheral signalling to the central nervous system (CNS) may also play a defining role in chemosensory responses [8–63]. A better understanding of all these mechanisms may be required to fully capture the depth and breath of human chemosensory experiences, and their influence on food selection.

Second, we did not observe any direct relationships between geography and the distribution of \( \text{TAS2R38} \) haplotypes or between geography and food liking in the populations we studied. Our findings differ from those of Pemberton et al. [47] who studied \( \text{TAS2R38} \) haplotypes in Asian Indians born in 15 geographic regions across India. They found that haplotype frequencies varied along a latitudinal cline with more tasters in the northern groups and more non tasters in the southern groups. Although Pemberton et al. [47] did not study food liking, it is intriguing that hot spices, like chilli pepper are more frequently consumed in southern India [64] in the same areas where non tasters predominate. Given the critical role of geography and climate in shaping the genetic features of world populations [65], we can only speculate that the geographical and ecological barriers to genetic and cultural exchanges in the groups residing in India along a north-south gradient were more formidable than those operating along the Silk Road which has been an east-west corridor for such exchanges for thousands of years.

However, asymmetrical gene flow and the availability of different crops could also be responsible for variability in genetic features across populations [66]. In addition this study was also limited by the sample size in each population.

These aspects can represent the limitations of the present study, therefore future studies involving a deeper analysis of other genes and environmental variables could further elucidate population differences in taste responses and food liking.

In conclusion, we used a population-based approach in which we exploited taste phenotypic differences among populations to reveal differences in food liking patterns across populations that could not be detected using standard methods. This approach, based on comparisons between distance matrices, can be applied to different population groups around the globe to obtain a comprehensive view of the role of PROP tasting in food preferences as well as to explore the role of novel taste-related traits in food choice.

![Cluster Dendrogram](image)

**Figure 3. Dendrogram based on differences in food preferences between populations.** Three groups are clear from the dendrogram: one composed by Georgia, the second one by Uzbekistan Kazakhstan and Azerbaijan and the third one by Armenia and Tajikistan. doi:10.1371/journal.pone.0091716.g003

<table>
<thead>
<tr>
<th>Table 3. Mantel test results between distance matrices analyzed.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geography</td>
</tr>
<tr>
<td>Genomic Fst</td>
</tr>
<tr>
<td>TAS2R38 Fst</td>
</tr>
<tr>
<td>PROP Status</td>
</tr>
<tr>
<td>Food liking</td>
</tr>
</tbody>
</table>

In bold are reported significant correlations. doi:10.1371/journal.pone.0091716.t003
Supporting Information

Figure S1  Multidimensional scaling of PROP status (a) and food preferences (b) matrices. (TIFF)

Table S1  Examined populations and their sample sizes. (DOC)

Table S2  The percentage of AVI/AVI, PAV/AVI and PAV/PAV subjects in the groups of NT (non taster), MT (medium taster) and ST (super taster) for each population. (DOC)

Table S3  Mean and standard deviation of each food included in the food liking questionnaire in the overall sample and in each population. (DOC)

Table S4  Mean and standard deviation of food liking groups in each population. (DOC)

References


PROP Perception and Food Preferences in Silk Road

Future developments

The Marco Polo expedition gives us the possibility to obtain several phenotype data from different populations. Environmental variables and rare alleles could have great impact on phenotype variability, and also several phenotypes could appear only in specific condition. As shown a fully description of population history and genetic structure has important implications for the design and understanding of disease mapping studies and reconstructing human evolutionary history. Despite it is recognized that the undetected population structure can cause spurious associations in disease mapping studies, we should avoid to merely solving the problem trying a simplistic way. Population structure has a deeper meaning, and we should understand and describe the role of migrations, history, genetic drift and several other aspects when we define a population. Only combining all these information with and a detailed population analysis we will be provided with useful tools to solve and use the genetic structure information to our advantage.

The Marco Polo scientific expedition samples are a wonderful laboratory for discovering new variants implicated in phenotype variation, finding interaction between genes and environments and describing the relationship between traits.

Future steps will include the development of a methodology that will combine the great array of population genetics data with trait variation in a viable way that could be used in association and epidemiological studies in these populations.


BLAKESLEE, A., 1918 Unlike reaction of different individuals to fragrance in verbena flowers. Science 48: 298-299.


MCRAE, J. F., J. D. MAINLAND, S. R. JAEGE, K. A. ADIPIETRO, H. MATSUNAMI et al., 2012 Genetic variation in the odorant receptor OR2J3 is associated with the ability to detect the “grassy” smelling odor, cis-3-hexen-1-ol. Chemical senses.
MORI, K., and H. SAKANO, 2011 How is the olfactory map formed and interpreted in the mammalian brain? Annual review of neuroscience 34: 467-499.
SELECT * FROM item
WHERE item_id = 9437;

<table>
<thead>
<tr>
<th>Item ID</th>
<th>Title</th>
<th>Authors</th>
<th>Journal</th>
<th>Year</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Archaeology and language: the puzzle of Indo-European origins.</td>
<td>C. Renfrew</td>
<td>CUP Archive</td>
<td>1990</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Olfaction and environment: Tsimane’o f Bolivian rainforest have lower threshold of odor detection than industrialized German people.</td>
<td>A. Sorokowska, P. Sorokowski, T. Hummel, T. Huanca</td>
<td>PLoS one</td>
<td>2013</td>
<td>8: e69203.</td>
</tr>
</tbody>
</table>


CURRICULUM VITAE

BIOGRAPHICAL SKETCH

NAME Mezzavilla Massimo
BIRTH: 28/11/1986
CITIZENSHIP: Italian

POSITION/TITLE PhD Student in Molecular Genetics at University of Trieste

<table>
<thead>
<tr>
<th>INSTITUTION AND LOCATION</th>
<th>DEGREE</th>
<th>YEAR CONFERRED</th>
<th>FIELD OF STUDY</th>
</tr>
</thead>
<tbody>
<tr>
<td>University of Padua, Italy</td>
<td>Bachelor Degree</td>
<td>2008</td>
<td>Molecular Biology</td>
</tr>
<tr>
<td>University of Padua, Italy</td>
<td>Master Degree</td>
<td>2010</td>
<td>Molecular Biology</td>
</tr>
</tbody>
</table>

e-mail: massimo.mezzavilla@burlo.trieste.it
mm25@sanger.ac.uk
mzzvilla@gmail.com

Research and Professional Experience

January 2012 - today: PhD Student (Molecular Genetics) at the Department of Medical Chirurgical & Health Sciences of the University of Trieste (Italy). Principal fields of interest: Population Genetics of Isolate Populations, Population Genetics of Central Asia populations Epidemiology, genetics of smell perception, genetics of heart rate, genetic of taste.

March 2014 - September 2014: Visiting Student at Wellcome Trust Sanger Institute, Chris Tyler Smith group.

Summary of research in recent years

So far, my research has been focused on the field of Population Genetics. In particular, one of my main interests has been towards the analysis of genetic data from Isolate populations from all Europe (in collaboration with Chris Tyler-Smith group from Wellcome Trust Sanger Institute). The final goal of this research has been the identification of general genetic characteristics of the isolated populations, and answer several questions: what can we learn about their demographic history from the low-depth WGS?

Are there systematic differences between the genetic burden and predicted consequences of the genetic make-up of non-isolated and isolated populations?

What are the implications, direct and indirect, for studies of complex disease?

My second field is the study of Central Asia populations, this geographical area revealed one of the most interesting histories, disentangling the genetic history and understanding how this history shaped the phenotypic variation in these populations is the main goal of this research (in collaboration with Dott.ssa Vincenza Colonna from CNR, Naples). I am now studying the genetic history of Armenia population (in collaboration with Marc Haber, Wellcome Trust Sanger Institute), and also the genetic structure and effect of positive selection in Kalash population (in collaboration with Qasim Ayub, Wellcome Trust Sanger Institute).

A third field on interest is the study of olfactory phenotypes in different populations, the main scope is to find how genetic polymorphisms in odorant receptor alter smell function, and analyze difference between populations in olfactory perception.

Other fields in which I am working are: the study of heart rate in isolates Italian villages, the genetic of taste, and finally the study of demography of human population after the Out Of Africa (in collaboration with Guido Barbujani, University of Ferrara).

Currently I am also leading a project regarding the genetic characterization of the genetic isolate of Carlantino, with a focus on understanding the demographic events, which shaped the genetic variants that affect biomedical traits.
LIST OF PUBLICATIONS


* These authors contributed equally to this work

A4-Sheila Ulivi, Massimo Mezzavilla, Paolo Gasparini: Genetics of eye colours in different rural populations on the Silk Road. European journal of human genetics: EJHG 03/2013;
A5-I A M Marino, A Benazzo, C Agostini, M Mezzavilla, S M Hoban, T Patarnello, L Zane, G Bertorelle: *Evidence for past and present hybridization in three Antarctic icefish species provides new perspectives on an evolutionary radiation*. Molecular Ecology 08/2013;


A9-Massimo Mezzavilla, Annamaria Iorio, Marco Bobbo, Angela D'Eustacchio, Marco Merlo, Paolo Gasparini, Sheila Ulivi, Gianfranco Sinagra: *Insight into genetic determinants of resting heart rate*. Gene 03/2014;

genetic drift at missense and trait-associated variants. Accepted on Nature Communications.

A11-Massimo Mezzavilla, Diego Vozzi, Nicola Pirastu, Giorgia Girotto, Pio d’Adamo, Paolo Gasparini, Vincenza Colonna. Genetic landscape of populations along the Silk Road: admixture and migration patterns. BMC Genetics (5 December 2014)

A12-Massimo Mezzavilla, Maria Geppert, Chris Tyler-Smith, Lutz Roewer, Yali Xue. Insights into the origin of rare haplogroup C3* Y chromosomes in South America from high-density autosomal SNP genotyping. Forensic International: Genetics doi:10.1016/j.fsigen.2014.11.005

Accepted/in press

A13-Valentina Cenedese; Massimo Mezzavilla; Anna Morgan; Renato Marino; Cosimo Pietro Ettorre; Maurizio Margaglione; Paolo Gasparini; Anna Menini. Assessment of the olfactory function in Italian patients with type 3 von Willebrand disease caused by a homozygous 253 kb deletion involving VWF and TMEM16B/ANO2. Accepted to PLoS ONE.

A14-Massimo Mezzavilla, Diego Vozzi, Ramin Badii, Moza Khalifa Alkowari Khalid Abdulhadi, Giorgia Girotto, Paolo Gasparini. Increased rate of deleterious variants in long runs of homozygosity of an inbred population from Qatar. Accepted to Human Heredity

A15-Massimo Mezzavilla, Silvia Ghirotto. NeON: an R package to estimate human effective population size and divergence time from patterns of linkage disequilibrium between SNPs. Accepted to Journal of Computer Science & Systems Biology