Spatially Explicit Models to Investigate Geographic Patterns in the Distribution of Forensic STRs: Application to the North-Eastern Mediterranean

Francesco Messina[1], Andrea Finocchio[1], Nejat Akar[2], Aphrodite Loutradis[3], Emmanuel I. Michalodimitrakis[4], Radim Brdicka[5], Carla Jodice[1], Andrea Novelletto[1]*

1 Department of Biology, University "Tor Vergata", Rome, Italy, 2 Pediatrics Department, TOBB-Economy and Technology University Hospital, Ankara, Turkey, 3 National Center for Thalassemias, Athens, Greece, 4 Department of Forensic Sciences, University of Crete School of Medicine, Heraklion, Crete, Greece, 5 Institute of Haematology and Blood Transfusion, Praha, Czech Republic

* novelletto@bio.uniroma2.it

Abstract

Human forensic STRs used for individual identification have been reported to have little power for inter-population analyses. Several methods have been developed which incorporate information on the spatial distribution of individuals to arrive at a description of the arrangement of diversity. We genotyped at 16 forensic STRs a large population sample obtained from many locations in Italy, Greece and Turkey, i.e. three countries crucial to the understanding of discontinuities at the European/Asian junction and the genetic legacy of ancient migrations, but seldom represented together in previous studies. Using spatial PCA on the full dataset, we detected patterns of population affinities in the area. Additionally, we devised objective criteria to reduce the overall complexity into reduced datasets. Independent spatially explicit methods applied to these latter datasets converged in showing that the extraction of information on long- to medium-range geographical trends and structuring from the overall diversity is possible. All analyses returned the picture of a background clinal variation, with regional discontinuities captured by each of the reduced datasets. Several aspects of our results are confirmed on external STR datasets and replicate those of genome-wide SNP typings. High levels of gene flow were inferred within the main continental areas by coalescent simulations. These results are promising from a microevolutionary perspective, in view of the fast pace at which forensic data are being accumulated for many locales. It is foreseeable that this will allow the exploitation of an invaluable genotypic resource, assembled for other (forensic) purposes, to clarify important aspects in the formation of local gene pools.

Introduction

The power of forensic STR loci for individual identification has led to the accumulation of huge amounts of data, whose interest goes beyond forensic issues and potentially relates to the gene geography of human populations and microevolutionary processes. Surveys of population samples with large arrays of STR loci [1, 2] have shown the ability of these markers to reveal many aspects of genetic structure in diverse human populations at the continental and sub-continental scales. Based on these findings, different authors interpreted the arrangement of the total diversity as mainly clinal or mainly clustered, at least at the spatial scale of the examined population samples [3, 4]. The same ability for the subset of autosomal STR loci commonly used for forensic purposes is not well established. Rowold and Herrera [5] used five forensic STRs and obtained phylogenetic relationships of worldwide populations consistent with other markers. Silva et al. [6] detected a progressive reduction of diversity with increasing distance from Eastern Africa, a signature of the serial founder effect which accompanied the spread of modern humans out of Africa and beyond. However, these latter authors found low fixation indices, and suggested that, as far as these markers have been specifically selected to maximize the within-population (or between-subjects) variance, they carry a tiny proportion of information useful for between-population inferences. Additionally, it was suggested that the ethnic composition of population panels used in surveys with forensic and non-forensic loci may have resulted in lower fixation indexes in the former ones [7]. Recently, it was also shown that the widely used index Fst is mathematically constrained for loci with high heterozygosity, while statistics derived from multivariate methods are indeed able to extract ancestry information, reviving the value of loci for individual identification also in population identifiability, at least at the continental scale [8].

In recent years, several methods have been developed which, in addition to genotypic data, incorporate information on the spatial distribution of individuals within a species or a population to arrive at a description of the arrangement of diversity across that space [9–14]. These methods address different questions as compared to those (e.g. [15, 16]) aimed at detecting hidden population structuring. In fact, to test hypotheses based on the spatial context, a method should be spatially explicit, i.e. it should directly take spatial information into account as a component of the adjusted model or of the optimized criterion, thereby focusing on the part of the variability which is spatially structured [17]. The key step in this class of procedures is to consider entities (individuals or populations), and to weight their proximity via plots which partition the space, such as tessellations, lattices and connection networks. This class of methods has been increasingly used to answer questions on gene flow, barriers, admixture and the detection of sporadic immigrants in human and non-human populations [10, 18, 19].

We report here on the typing of a large population sample obtained with a fine-grained sampling scheme involving locations in four main areas along a transect over Southern Europe and the Near East, i.e. Southern Italy, Continental Greece, the Aegean Islands and Turkey (S1 Fig and S1 Table). Populations from the same areas are seldom represented together in previous studies. Yet, this geographical transect, which embraces most of the North Eastern Mediterranean, is crucial for the understanding of population movements that took place over many millennia, contributing to the making of the Southern European gene pool. These include, at least: i) the entry of anatomically modern humans in Europe [20]; ii) the spread of the Neolithic culture, which brought peoples, crops and livestock to the West, most likely as a punctuated series of mainly maritime migration episodes [21–24]; iii) the expansion of the Hellenic world around the 8th century B.C.E. [25]; iv) historical conquests in Anatolia and the Southern Balkans [26, 27]. The relative impact of each of these on the overall diversity of the

current genome pool is still contentious, but it is to be expected that events occurring at different times and by different processes have left distinct spatial signatures.

We used 16 autosomal STR loci, included in a popular kit designed for individual identification. Thus, a feeble signal of structuring could be expected, especially in the transect here examined, which represents only a subset of the continental diversity. In this work we wanted to challenge the possibility of separating the main signal of the underlying arrangement of inter-population diversity from the background noise generated by the extreme inter-individual diversity of these markers. In this regard, we reasoned that the complex geography of the area constrained gene flow and population movements, not only across Seas, but also inland, due to the scattered arrangement of human settlements in the mountainous environment of Southern Italy and Greece. This prompted us to use many, relatively small samples, preserving their geographic attributes, and methods based in a geographic framework. A positive outcome in detecting spatial patterns would be promising, in view of the fast pace at which forensic data are being accumulated for many locales. This will be the basis for a description of the geographic distribution of diversity at an unprecedentedly fine scale.

Results

The final dataset comprised 1559 subjects, all typed at 16 STR loci. Overall, 288 alleles were recorded, of which 27 not present in the allelic ladder provided with the kit (overladder) and named here on the basis of their molecular weight (S2 Table). Overladder alleles were abundant at SE33, where they are known to arise also from variation outside the repeated stretch [28]; we did not investigate further the molecular structure of these variants and treated them as separate alleles. The number of alleles per locus varied between 9 (loci D10S1248, D16S539 and TH01) and 65 (SE33).

The screening of the entire dataset with RELPAIR [29] confirmed that all genotypes were different from each other, as expected from the low combined probability of identity (2.35E-21) of the kit [30].

We wanted to assay the potential presence of null alleles by estimating their frequency at each locus in all location samples (16 x 41 = 656). In this regard, in the absence of homozygotes for the null allele, the estimated frequency depends on reduced heterozygosity as compared to the expected one. We then performed in parallel an estimation of the expected and observed heterozygosities as summarized by the index Fis. As expected, the two measures turned out to co-vary (S2 Fig). The estimated frequencies of null alleles reached up to 0.13, often grossly exceeding the values experimentally determined by comparing assays with different primer pairs and reported at http://www.cstl.nist.gov/biotech/strbase/NullAlleles.htm (always <1/100 subjects; frequency <0.005). Also averages over populations and over loci (S3 Fig top) often exceeded this value, without particular outliers and departure from normality.

The distribution of Fis was quite symmetrical around 0, with only a slight excess of positive values (359/656 = 55%), rarely exceeding 0.3. To check whether the diminished heterozygosity was an experimental artefact or was a real population feature, we contrasted our Fis values averaged over loci with those obtained in comparable populations with 645 STRs (Italians and Palestinians from Table S2 in ref. [31]). Among Italians (same population background but different subjects in the two studies), Fis values of +0.0090 (95% C.I. = -0.0016–0.0164) and +0.0123 were obtained in the two studies, respectively. Among Palestinians (same subjects in the two studies), the Fis values were +0.0124 (95% C.I. = -0.0414–0.0494) and +0.0186, respectively. We then concluded that our genotyping did not suffer a significant allele loss (by either PCR failure or profile reading errors) and interpreted the reduced heterozygosity as an effect of endogamy in our closely geographically confined location samples. Therefore, we continued

our analysis by considering all loci truly codominant. In these conditions, the exact test for the Hardy-Weinberg equilibrium (HWE) did not show departures from the expectation (S4 Fig). The mean of population-specific Fis averaged over loci was 0.0119, ranging -0.058 to +0.045, with values above the average for 5 out of 6 Turkish locations, in agreement with consanguinity reports (www.consang.net), and 6 out of 7 for Continental Greece (S3 Fig bottom). Allele frequencies are reported in S3 Table.

Spatial analysis of the full dataset

The overall Fst among the 41 location samples was 0.0022, a value which reveals the inability of forensic STRs to summarize differentiation into Fst, due to the large number of alleles, many of which with low frequencies [32], inflated intra-population variances [6, 7], and the mathematical constraint on the index [8]. Nevertheless, the departure of our value from the null (P<1E-4), showed that the sampling scheme may have the power to detect some kind of structuring. With such low level of differentiation, a plot of the between-location pairwise Fst values (nmMDS) not only did not cluster any of the locations by geography, but positioned even the outgroups within a unique cloud. However, when subjects from locations in Italy, Greece, Crete, Aegean Islands and Turkey were grouped into 11 larger samples, a pattern coherent with geography was obtained. The nmMDS plot (stress = 0.120) summarizing this 13 x 13 (including the outgroups) pairwise Fst matrix (S5 Fig) displayed the two outgroups widely separated on axes 1 and 3. Turkish samples were oriented towards the South Eastern outgroup (Palestinians), with positive and negative values on dimensions 1 and 2, respectively. Continental Greek and Cretan samples were oriented towards the North Western outgroup (Czechs in Brno), but with strongly positive dimension 3 values. Italian samples occupied a central position in all dimensions. The same Fst matrix was significantly correlated (r = 0.42; P = 0.017) with the matrix of geographic distances. This result was encouraging in prompting the use of spatially explicit models to highlight geographical patterns more clearly, and at the resolution of our sampling scheme (41 locations).

We first applied the Spatial PCA (sPCA) method implemented in adegenet [9] by using the full dataset of 1559 subjects x 16 loci (288 alleles). We obtained 7 and 33 positive and negative eigenvalues, respectively. The first two positive (global) eigenvalues (6.93E-3 and 4.33E-3) were remarkably larger than the third (1.87E-3) and the following ones, suggesting a boundary between strong and weak structures of the data [17], though a test of the global spatial structure did not reject the null hypothesis (P = 0.10). At any rate, sPC1 and sPC2 determined an increase in the Moran I of 10 and 3 folds, respectively, as compared to an ordinary PCA, indicating that the method could capture greater similarities between locations connected than non-connected along the network. In sPC1 (Fig 1A) the North-Western outgroup and most of the Italian locations (16/18) were identified by negative scores, whereas Greek and Turkish locations plus the South-Eastern outgroup were identified by positive scores. Two locations in continental Italy (Cilento promontory and Locri), one in continental Greece (Agrinion), one in the Aegean Sea (Khios) and one in continental Turkey (Central Anatolia) contrasted with this pattern. In sPC2 (Fig 1B) all of the Turkish locations and the South-Eastern outgroup were identified by extreme negative scores, whereas the North-Western outgroup, all of the Continental Greek and Western Cretan locations were characterized by positive scores. Eastern Crete showed moderately negative values. Italian locations were characterized by small negative (11/18) and positive (7/18) values.

We next asked which alleles were the main contributors to these (global) patterns. We then examined the distribution of squared loadings on each of sPC's 1 and 2, using increasing levels of stringency (Table 1). All loci were represented in the allele lists, with some of them (e.g.
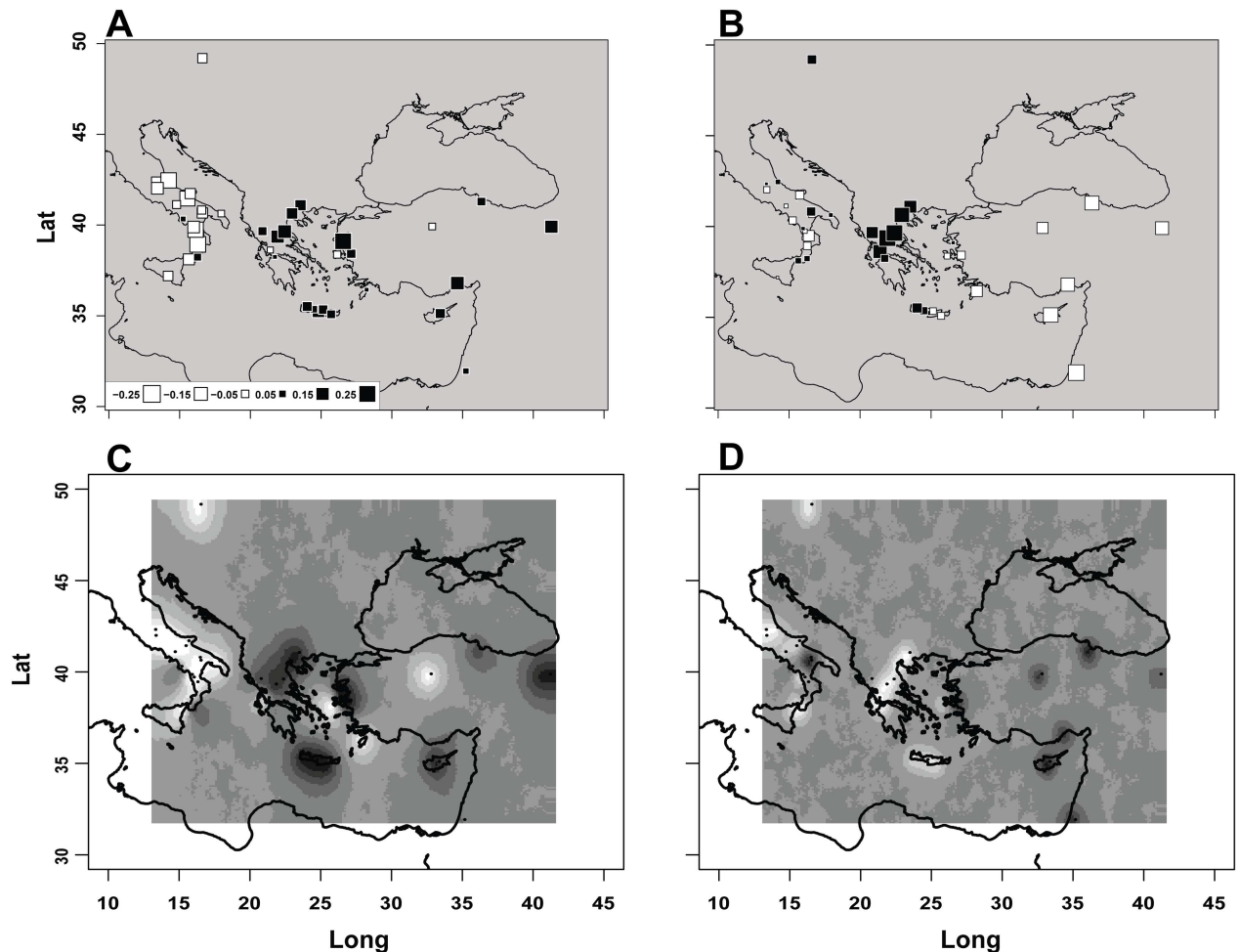
Fig 1. Maps of: A) scores for the 41 locations in sPC1 obtained on the full dataset with adegenet; B) scores in sPC2 obtained as in A; C) posterior assignment probabilities of the 41 locations to either of two clusters obtained on the reduced dataset derived from sPC1 with Geneland; D) posterior assignment probabilities of the 41 locations to either of two clusters obtained on the reduced dataset derived from sPC2 with Geneland. In A and B white and black squares represent negative and positive scores, respectively, with square size proportional to the absolute value (inset in panel A). In each of panels C and D shades of grey indicate probabilities of assignment to one of two mutually exclusive clusters from 0 (dark grey) to 1 (white). Color versions of panels C and D are reported in S7 Fig.

doi:10.1371/journal.pone.0167065.g001

FGA and D12S391) appearing only in the top 10% quantile. On the other hand, some loci (e.g. D10S1248 and D3S1358) contributed strongly to both sPC's, sometimes even with the same allele. Some loci contributed to the same sPC with more than one allele, a not unexpected finding in view of the intrinsic negative correlation between the frequencies of alternative alleles at the same locus. Finally, only in 9 out of 16 loci the most frequent allele emerged, confirming that a relevant spatial signal was often due to alleles that contribute only marginally to the overall variance [8]. These observations suggested that some loci, and some of their alleles, convey a clearer signal of spatial structuring, and one or more reduced datasets can be obtained, which could retain most of the geographic information but with an abated background noise. In order to identify the optimal sets of loci and alleles we combined three criteria: a) independence between the reduced datasets derived from sPC1 and 2, i.e. the same allele should not appear in both reduced datasets; b) minimization of internal negative correlation, i.e. multiple alleles from the same locus should be avoided in the same reduced dataset; c) improved ability

Table 1. Loci and alleles with the strongest impact on the SpatialPCA eigenvalues 1 and 2.

Top quantile in the distribution of squared loadings	sPC1					sPC2				
Locus (allele)[a]	10%	7.50%	5%	2.50%	1.25%	10%	7.50%	5%	2.50%	1.25%
	D10S1248(13)	D10S1248(13)	D10S1248(13)	D10S1248(13)	D10S1248(13)	**D10S1248(14)[b]**	**D10S1248(14)[b]**	**D10S1248(14)[b]**	**D16S539(12)**	**D16S539(12)**
	D10S1248(14)[b]	**D10S1248(14)[b]**	**D10S1248(14)[b]**	**D10S1248(14)[b]**	vWA(17)[b]	D16S539(10)	**D16S539(12)**	**D16S539(12)**	D2S1338(24)	D18S51(15)
	vWA(16)	vWA(16)	vWA(17)[b]	vWA(17)[b]	D3S1358(16)[b]	**D16S539(12)**	D2S1338(19)	D2S1338(19)	D18S51(15)	D18S51(19)
	vWA(17)[b]	vWA(17)[b]	**D16S539(12)**	**D2S1338(17)[b]**	D1S1656(13)	D16S539(13)	D2S1338(24)	D2S1338(24)	D18S51(19)	D3S1358(18)
	D16S539(12)	**D16S539(12)**	**D2S1338(17)[b]**	D19S433(13)		**D2S1338(17)[b]**	D21S11(32.2)	**D18S51(12)**	**D22S1045(15)[b]**	
	D2S1338(17)[b]	**D2S1338(17)[b]**	D8S1179(13)[b]	**D2S441(11)[b]**		D2S1338(18)	**D18S51(12)**	D18S51(15)	D22S1045(16)	
	D8S1179(13)[b]	D8S1179(13)[b]	**D18S51(12)**	D3S1358(16)[b]		D2S1338(19)	D18S51(15)	D18S51(19)	TH01(9)	
	D21S11(30)	**D18S51(12)**	D19S433(13)	D1S1656(13)		D2S1338(24)	D18S51(19)	**D22S1045(15)[b]**	D3S1358(18)	
	D18S51(12)	D18S51(14)[b]	**D19S433(15.2)**			D8S1179(14)	**D22S1045(11)**	D22S1045(16)		
	D18S51(14)[b]	D18S51(16)	TH01(80)			D21S11(32.2)	**D22S1045(15)[b]**	D19S433(15)		
	D18S51(16)	D19S433(13)	**D2S441(11)[b]**			**D18S51(12)**	D22S1045(16)	TH01(9)		
	D22S1045(11)	**D19S433(15.2)**	D2S441(14)			D18S51(15)	D22S1045(17)	D3S1358(18)		
	D22S1045(17.2)	**TH01(8)**	D3S1358(14)			D18S51(19)	D19S433(15)	D1S1656(16)		
	D19S433(13)	FGA(25)	D3S1358(16)[b]			**D22S1045(11)**	**D19S433(15.2)**	D1S1656(16.3)		
	D19S433(15.2)	**D2S441(11)[b]**	D1S1656(13)			**D22S1045(15)[b]**	TH01(9)	D12S391(21)		
	TH01(8)	D2S441(14)				D22S1045(16)	D3S1358(17)			
	FGA(22)	D3S1358(14)				D22S1045(17)	D3S1358(18)			
	FGA(25)	D3S1358(16)[b]				D19S433(15)	**D1S1656(15)[b]**			
	D2S441(11)[b]	D1S1656(12)				**TH01(8)**	D1S1656(16)			
	D2S441(14)	D1S1656(13)				TH01(9)	D1S1656(16.3)			
	D3S1358(14)	**D1S1656(15)[b]**				D3S1358(17)	D1S1656(17.3)			
	D3S1358(16)[b]	SE33(16)				D3S1358(18)	D12S391(21)			
	D1S1656(12)					D1S1656(11)				
	D1S1656(13)					**D1S1656(15)[b]**				
	D1S1656(15)[b]					D1S1656(16)				
	D12S391(22)					D1S1656(16.3)				
	SE33(16)					D1S1656(17.3)				
	SE33(19)					D12S391(21)				
	SE33(28.2)									

[a]. Listed in the order of increasing molecular weights in the blue, green, black and red series of the electropherograms

[b]. Most frequent allele in the locus

Multiple alleles from the same locus are underlined. Alleles shared between the two sPC's are shown in boldface.

doi:10.1371/journal.pone.0167065.t001

Table 2. Fst analysis in the main geographic regions.

	sPC 1		sPC 2	
Region	**Fst**	**P**	**Fst**	**P**
All 41 locations	0.0036	<0.01	0.0031	<0.01
Italian Peninsula	0.0056	<1E-4	0.0029	0.08
Continental Greece	-0.0064	n.s.	-0.0044	n.s.
Crete	0.0012	n.s.	-0.0010	n.s.
Turkey	0.0021	n.s.	-0.0001	n.s.

to detect structuring by the reduced datasets as compared to the full dataset. From the lists reported in Table 1, we thus considered optimal the choice of the alleles producing the top 2.5% of squared loadings, and recoded all individual genotypes considering each of the 8 alleles against all the remaining alleles at the respective locus (hereafter "reduced datasets"). As a result, these reduced datasets were quite complementary. First, only two loci (D2S1338 and D3S1358) contributed to both reduced datasets (each with different alleles); second, in only 3 cases two alleles of the same locus were included; third, the reduced datasets derived from sPC's 1 and 2 produced an increase of Fst to 0.0036 (P<0.01) and 0.0031 (P<0.01), respectively (top row in Table 2), i.e. approximately 50% higher than the full dataset.

The frequency maps of alleles included in the reduced datasets are shown in S6 Fig. A common feature of these maps is that a general clinal variation appears, though this is evident mostly outside the polygon defined by the sampling locations, where the kriging prediction of frequencies suffers of larger errors [33]. The correlograms for the individual alleles displayed different patterns in the arrangement of similarity with distance. For sPC1 four correlograms were significant over the entire range of distance classes: D2S1338(17) and D1S1656(13) displayed a clinal pattern over the whole range (Bonferroni corrected P<0.005 and P<0.05, respectively), whereas D10S1248(13) and D10S1248(14) displayed a minimum of the Moran I in the 800–1200 km class (Bonferroni corrected P<0.05 and P<0.001, respectively). For sPC2 only D3S1358(18) displayed a clinal pattern over the whole range (Bonferroni corrected P < 0.05), whereas D16S539(12) also showed a minimum of the Moran I in the 800–1200 km class (Bonferroni corrected P < 0.005). Allele D18S51(19) showed a strong correlation at the shortest distances but a flat pattern in the remaining distance classes. In summary, a geographic structuring for these alleles, extracted with the sPCA from a pool 20 times larger, was evident. However, we noticed that the 800–1200 km distance class included most of the comparisons between locations in Greece vs those to the West and to the East of it. In fact, when looking at the maps more carefully, it is noticeable that almost all of them display strong foci of high/low frequencies in the areas more densely covered by our locations, and such a pattern is particularly recurrent for Continental Greek locations, especially in sPC2.

Spatial analyses of the reduced datasets

We then wanted to control that using only 8 out of 288 alleles from each sPC did not alter substantially the spatial patterns of similarity among locations. To this aim we added to our analyses two independent methods, separately for the reduced datasets derived from sPC's 1 and 2. The membership probability maps (Fig 1C and 1D; S7 Fig) replicated those obtained with sPCA on the full set of alleles in many respects. The reduced dataset derived from sPC1 (Fig 1C) produced probabilities of assignment to each of two mutually exclusive population clusters ranging from 0.99:0.01 (e.g. in the Italian Peninsula) to 0.01:0.99 (Mitilini). The same 16 Italians locations identified by negative scores in Fig 1A were assigned to a single cluster with
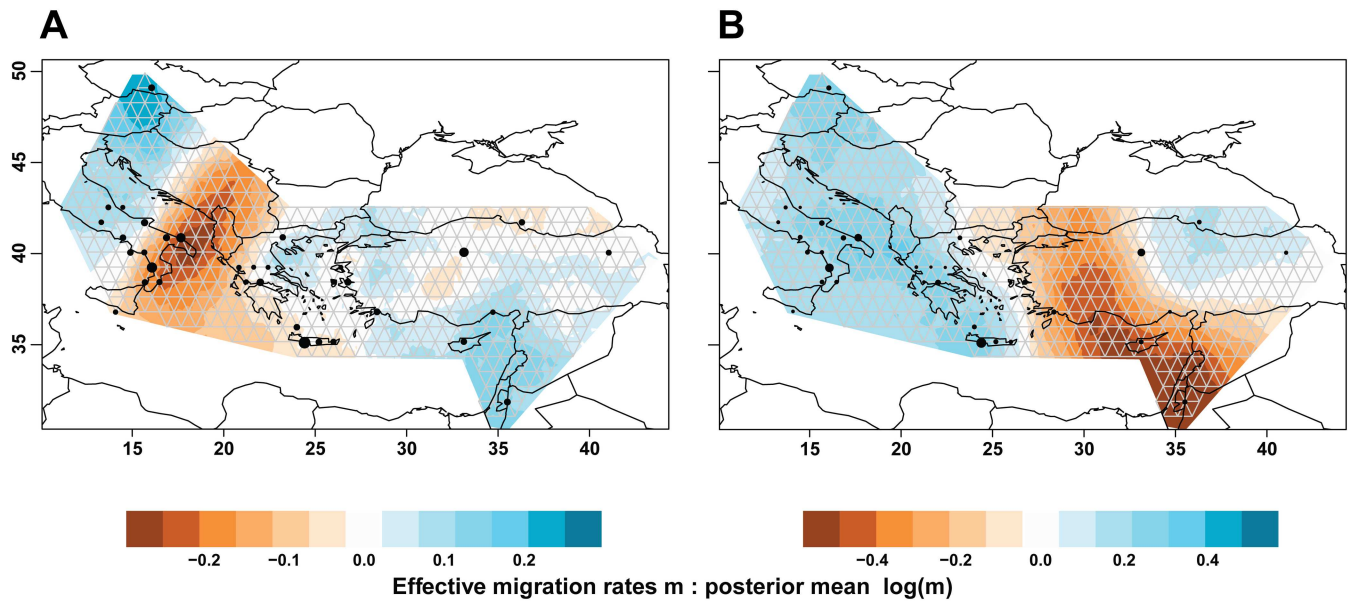
Fig 2. Representation of effective migration surfaces as obtained with EEMS on the reduced datasets derived from sPC 1 (A) and sPC2 (B). The coloured area covers only the user-defined polygon. The grid used by the program is shown in grey. Note that only 34 sampled demes appear (black dots, with size proportional to the n. of individuals), assigned to a grid vertex and not necessarily coinciding exactly with the original sampling location. Pooled locations were (numbered as in S1 Table): 6+7, 9+10, 13+14+15+16, 25+26, 30+32. Note the different colour scales between the two maps. In both maps brown belts correspond to low migration values, i.e. barriers to gene flow.

doi:10.1371/journal.pone.0167065.g002

probabilities of 0.7 or higher (white shade in Fig 1C and S7 Fig). The same Continental Greek, Cretan (plus Mitilini) and Turkish locations, which were identified by positive scores in Fig 1A, were assigned to the alternative cluster with probabilities of 0.7 or higher (dark grey and red shades in Fig 1C and S7 Fig, respectively). The reduced dataset derived from sPC2 (Fig 1D) produced probabilities ranging from 0.96:0.04 (e.g. in Continental Greece) to 0.03:0.97 (Black Sea coast). In this case all Continental Greek and Cretan locations, identified by positive scores in Fig 1B, were assigned to a single cluster with probabilities of 0.7 or higher (white shade in Fig 1D and S7 Fig), whereas Turkish locations plus the South-Eastern outgroup were assigned to the alternative cluster with probabilities of 0.7 or higher (dark grey and red shades in Fig 1D and S7 Fig, respectively). The affiliation of some locations shifted as compared with Fig 1B. In continental Italy 14 locations were now assigned to the same cluster as the Greek ones with probabilities > 0.6 and only two to the alternative cluster with probabilities >0.8, and all Cretan locations were coherently assigned to the same cluster as the Greek ones with probabilities >0.7.

Also, the effective migration surfaces analysis confirmed these broad patterns. The reduced dataset derived from sPC1 (Fig 2A) produced lower effective migration rates across the Southern Adriatic and Ionian Seas, with increased rates to the West of this belt, a pattern indicative of isolation by distance across the Aegean Sea, and again enhanced migration in the South-Eastern sector of our sampling range. In this map, the putative connection of two Italian locations (Cilento promontory and Locri) with Continental Greek locations, apparent in Fig 1A and 1C and S7 Fig, did not show up. In fact the authors of the method remark that, while a similarity of this kind could be captured by a narrow corridor, inserting such a corridor would make the overall model fit worse [10]. The reduced dataset derived from sPC2 (Fig 2B) produced effective migration rates strongly lowered along a belt stretching from the South-Eastern outgroup to Western Anatolia, shifted to the East as compared to the discontinuity in the

scores and assignment probabilities of Fig 1B and 1D and S7 Fig. To the West of this belt an area including all the remaining demes displayed high migration rates (blue in Fig 2B).

Dependency on the sampling scheme

As such, the results outlined above seem to reveal long-range structuring. However, our geographic range is unusual, being stretched over 30 degrees of longitude but compressed into 10 degrees of latitude, and with obvious constraints as to the spatial organization of sampled points. We thus sought to reinforce our findings by answering two questions: 1) How do our methods respond to a rarefaction of the sampled points? and 2) Are similar patterns detectable with external datasets, from broadly the same area examined here?

Briefly, the following lessons emerged from this work (S1 text). First, the rarefaction of locations implies that radically different pairs of points get connected in the sPCA network. In a situation with large frequency variations also at short distances (as determined e.g. by drift, founder effects and sampling variance), Moran I contributions to sPC's in complementary though spatially overlapping subsets may shift from positive to negative, and vice-versa. Thus, different alleles may show up as the main contributors to the global pattern in each case. Under these circumstances, though the main patterns were conserved, the more sparse locations suffered a larger uncertainty in their assignment to clusters, with increasing discrepancy between sPCA and the Bayesian method (see next paragraph). Second, our methods implicitly addressed global patterns of similarity. Frequencies obtained from external databases for those alleles recognized as the strongest contributors to our sPC's, and retained in the reduced datasets, indeed displayed the same clines extended outside the polygon of our locations. Third, the number of outgroups used here (2/41) did not overwhelmingly determine the clustering pattern, as far as they lay within the same broad-range cline(s).

Local differentiation

All the results described so far converged in showing that the extraction of information on long- to medium-range geographical trends from the overall diversity is possible. Further, a small subset of <6% of all alleles seems to retain most of the geographic information, and is able to enhance the low differentiation between locations. Analyses of both the full and the reduced datasets returned the picture of a background clinal variation onto which prominent local highs and lows generate a patchy pattern (visualized as frequency foci in contour maps). The presence of such local effects was seemingly captured also by the sPCA, which returned an excess of low magnitude negative eigenvalues (33/40), i.e. in which locations relatively close to each other (and connected in the network of S8 Fig) produced negative Moran I's. This raised questions on whether these minor sPC's indeed were indicators of a real heterogeneity [34], even between adjacent locations within the same country. We then quantified more carefully the differentiation within regional areas, by calculating the fixation index Fst (using the reduced datasets derived from sPC's1 and 2, separately) for the locations in the areas listed in Table 2. Locations in the Italian Peninsula produced a significant Fst value with the sPC1 reduced dataset, 50% higher than that obtained on 41 locations. None of the other areas provided evidence for a similar differentiation. In pairwise comparisons, three of the Italian locations stood out i.e. Locri, Cilento promontory and Lungro. When these three locations were removed, the Fst dropped to 0.0018 (n.s.). On the basis of sPC1 scores and assignment probabilities (Fig 1A and 1C), Locri and Cilento promontory appeared affiliated with locations in Greece and the Aegean, which also produced positive sPC1 scores, whereas Lungro produced the most extreme negative score. In summary, with the exception of these three locations,

within each of the four areas (Table 2) low differentiation was observed. This testifies of strong connectivity and prolonged gene flow between the populations that we tried to represent with our sampling, at distances ranging from some tenths (such as within Crete and Calabria) to one thousand (such as between Turkish locations) kilometres. In order to set upper and lower bounds to migration rates compatible with our observations, we used continuous-time coalescent simulations of genomic diversity [35], tailored on the Italian location sample sizes (Model 1 in S9 Fig), but valid also for the other areas. When the separation of demes (14) was modelled to have occurred from 224 to 276 generations (6500–8000 years) ago, migration rates (m) lower than 0.0025 produced high Fst's, incompatible with our observed value of 0.0018 (P<0.008 or less), whereas m values of 0.005 or greater were acceptable, with 0.01 as the best fitting (S10 Fig).

In this context, which factor(s) may justify the Italian peculiarity? One possible explanation is the harsh mountainous landscape of inner Southern Italy. However, we notice that the increased Fst is attributable to three locations only, distinguished from the other ones by peculiar histories. In fact, Locri and Cilento promontory coincide with two of the most important cities founded by Greek colonists in the 7-6th centuries B.C.E. (Locri Epizefiri and Velia, respectively) during the establishment of "Magna Grecia". As opposed to other colonies of Magna Grecia, continuity of human settlement until today in the same locations or in the immediate surroundings is documented. On the other hand, Lungro was the place for the main settlement of migrants from Albania in the 15th century C.E. We then performed coalescent simulations using Model 2 (S9 Fig), in which 3 new incoming demes join the previous 14 at 96 and 18 generations in the past. Also in this case, complete isolation could be excluded. Interestingly, however, the m value producing the best fit with the observed Fst value of 0.0056 was one order of magnitude lower than that estimated for the 14 demes (0.001 vs 0.01), and a uniform value of 0.01 for all the 17 demes turned out to be barely compatible (P = 0.07) with it (S10 Fig). This result suggested that some degree of reduced gene flow may be responsible for the excess differentiation of the three outlier locations, in line with demographic analyses in one of them [36].

Discussion

We used recently developed spatially explicit models [10, 13, 17] to analyse forensic STR data focusing on the composition of gene pools rather than on the geographical assignment of individual genotypes. We addressed populations currently living in a geographic range of great relevance for the gene geography of Europe as a whole. Each of the four main areas here considered (Southern Italy, Continental Greece, the Aegean Islands, Turkey and sub-regions within them) can be regarded as a stepping stone for any westward migratory movement from the Levant and the associated dispersal of culture(s) [22]. Genetically, this is supported by phylogeographic surveys of this same geographic region with uniparental genetic systems [37–45], which led to the identification of the traces of an East-to-West gene flow, accompanied by molecular radiation of each of several lineages. Also, genetic affinities in the occurrence of male-borne lineages were observed between Aegean Islands and specific mainland populations, traceable to colonizations widely distant in time [43, 46]. Studies directly addressed at the colonization routes, led Paschou et al. [47] and Fernández et al. [48] to favour Crete and Cyprus (reached by seafaring) as early steps for Neolithic farmers, as compared to inland routes leading directly to the Southern Balkans. In this broad context, a special case can be made of Southern Italy, which was impacted by two main immigration events from the East, i.e. the arrival of the first agriculturalists in the early Neolithic [49] and the settlement of Greek colonists as a series of newly founded town-states starting from the

8th century B.C.E. The colonists in the two cases may have had different genetic ancestries and may have followed different routes. Also, they were allowed to grow for largely different amounts of time and under radically different competition/interaction regimes with already settled human groups, leading to a large uncertainty on their relative genetic legacy to the extant gene pool.

As to genotyping, our results revealed a background level of inbreeding in the majority of locations (roughly two thirds), which is not attributable to the use of 16 loci only, but is consistently replicated with larger numbers of loci in many world populations (see Fig S2 in [31]). Our estimated values of the inbreeding coefficient agree with the limited population size of the sampled locations and the traditional marital habits [36]. Also, they exceed the values (averaged over subjects) of the direct estimations of homozygosity obtained in two population samples overlapping with those examined here, i.e. Palestinians [31] and Italians [50]. This indicates that, when considering local samples, a slight excess of homozygosity is a common occurrence, even in individuals more remotely related than currently detectable with closely spaced genetic markers.

Methodological considerations and caveats

Pioneering studies with the Y chromosome [51] showed that collating forensic data can result in a high power of detecting spatial patterns even at the sub-national level. The spatial analyses here performed showed that the extraction of information on geographical trends from the overall diversity is to some extent possible also for autosomal forensic STRs, provided that explicit models are employed. Our serial usage of different methods was inspired by the obvious need of reducing the overall complexity of the dataset, yet retaining the signal of "global" (sensu [17]) spatial trends and of genetic structuring. Thus, we are not in the position to rank the performance of each of them. Only when used jointly, they provide a concordant view of the population structure.

The results reported here call for a replication of the analytical procedures on additional forensic STR datasets, in order to test their performance against known features of the gene geography of Europe and other areas [52]. We showed (see S1 text) that the spacing between sampling locations and their positioning in the surveyed area may affect the list of loci and alleles with strongest spatial signals. However, we also showed that the clines for alleles displaying the strongest contributions in this work extended also beyond the area covered by our locations, and in the same direction. It is thus to be expected that enlarging or shifting the surveyed geographic range will lead to lists of loci and alleles only partially overlapping with those reported here, and will require a new tuning of the parameters used in the analyses (density of the connection network, threshold for loadings, etc.). Furthermore, these considerations call for a testing of the same methods on simpler landscapes. These could allow a better trade-off between a pre-designed sampling scheme (see Chapter 3 in [33]) and the representation of interesting or peculiar locales.

Finally, the geographic range here surveyed was suitable to reveal trends only in the East-West axis, with little or no power to detect similarities/discontinuities on the South-North axis. Moreover, our dataset is enriched in Greek subjects as compared to other studies in the area. As discussed elsewhere [53], these two factors determine that eigenvalues 1 and 2 of our sPCA cannot be expected to necessarily have a direct correspondence with eigenvalues 1 and 2 in ordinary PCA's of studies encompassing a broader range of European populations. In these conditions, the orthogonality [54] between the first two sPC's of the present study, which we tried to maintain also in the reduced datasets derived from them, may have e.g. enhanced the representation of structuring returned by sPC1 over sPC2.

Clinal and clustering patterns in the context of the gene-geography of the North-Eastern Mediterranean

With the above caveats in mind, the most prominent feature emerged in this study is the frequent clustering together of locations in the three main landmasses which protrude into the Central-Northern Mediterranean, i.e. the Italian peninsula, Greece and Turkey. Together with coalescent simulations, this indicated that high connectivity maintained a stronger homogeneity within the geographic peninsulas and islands (through land contacts) than between them, despite documented contacts by maritime routes over millennia.

Our sPC2, and the reduced dataset derived from it, set the Turkish locations apart from the ones to the West, with clearcut sPCA scores, assignment probabilities and a migration barrier. This feature and the spatial orientation of this sPC establish a straightforward similarity with other analyses of genome-wide SNP data in which Italian and Turkish subjects fell into separate clinal series covering the Middle East/Caucasus and Sardinia/Continental Europe, respectively [55, 56]. The centroids of Italian, Greek and a small number of Turkish subjects, turned out to be equidistant in three studies based on the POPRES resource [57–59]. In none of these studies Italian and Greek subjects were so well geographically referenced as we have done here, and it is thus not surprising that in the same studies the clouds of individual points largely overlapped. Our re-analysis (see S1 text) of subjects collected within the same geographical range considered here and typed at thousands of SNPs [47, 53] confirmed a clearcut discontinuity corresponding to the Aegean Sea. The same results argue for a hidden heterogeneity within Turkey which deserve future investigation and might be responsible for peculiar allele frequencies and assignment of our Central Anatolia location with sPC1.

A complementary and not necessarily alternative interpretation of the global spatial patterns highlighted here is a distinctiveness of Continental Greek locations, and especially the Eastern ones, which extended to Crete. Two studies [26, 27] stressed the relevance of expansions and admixture events in the 4-10th century C.E., which shaped the ancestry of Southern Balkans up to the Aegean coasts of Greece.

An Adriatic/Ionian discontinuity (Figs 1A, 1C and 2A) was captured by sPC1 and the reduced dataset derived from it. By analysing the pattern of sharing of genomic blocks identical by descent in the POPRES resource [27], an unusually little common ancestry was found between the Italian peninsula and other locations, seemingly deriving from longer ago than 2,500 years. Our re-analysis (see S1 text) of SNP data [47, 53], revealed an ill-defined line of distinctiveness between Sicily/Crete and Southern Continental Europe. In the presence of a single Southern Continental Italian subject in these data, it is not possible to establish how far this line extends northward. In this context, it is possible that our locations in Continental Italy spanned a barrier across the peninsula observed in other studies [60, 61], and more northerly locations (including the outgroup) acted as "attractors" in clustering, in view of their closeness to the other Italian ones (S8 Fig).

Yet another aspect is the retention of information on short-scale variation in the reduced datasets. Contrasting patterns in the distribution of Y-chromosomal and mtDNA markers emerged within Southern Italy [39, 62, 63]. Our coalescent simulations indicated that Fst values of the magnitude observed here could be better explained by high levels of gene flow prolonged over thousands of years. The estimated migration rates exceeded those obtained with a similar method for males and females in Asian patrilocal societies [64], and approach those experimentally measured in contemporary European populations [65, 66].

Some instances of the observed short-scale discontinuities of sPCA scores and assignment probabilities are also compatible with the historical accounts on the settlement of some of the locations in Southern Italy, which is in turn reflected also in the occurrence of Greek surnames

(Fig 5.7.4 in [52]). sPC scores and assignment probabilities for two Italian locations bear similarity to those of Eastern Greece and around the Aegean Sea, but the causal relation between the settlement of Magna Grecia and this observation remains to be determined. For example, traces of the Hellenic colonization were detected in some Sicilian locations only [67], and not in Southern Continental Italy [68]. Nevertheless, the method outlined here may be helpful to orient heuristic searches of specific markers in these locations, that could confirm/dismiss hypotheses on genetic contributions from the Early Neolithic Levant, the Hellenic world or the Balkans.

Implications

This work reaffirms that forensic STRs do contain some degree of ancestry-related information [5, 6, 69]. We endorse the concerns thoroughly discussed [8], which emerge from this perhaps unwanted property of such markers. Mathematically, the degree of population identifiability may have a non-null impact on calculations of the posterior probability of individual assignment. In practice, however, at the scale here considered the magnitude of such impact does not undermine the validity of forensic STRs in preserving ancestry anonymity.

Conclusion

Spatially explicit models empower the analysis of geographic structuring of extant human diversity at forensic STR loci up to a sub-continental and, possibly, to an even finer scale. It is foreseeable that this will allow the exploitation of an invaluable genotypic resource, assembled for other (forensic) purposes, to clarify important aspects in the formation of local gene pools. Intersecting the results obtained with different sets of loci for individual identification will come at the cost of reducing the number of shared loci but with the advantage of enormously increasing the sample size.

Materials and Methods

The samples

All samples were from collections of the authors, assembled in the 1980's, 1990's and 2000's to perform population studies in the region [39–41, 70–72]. The original blood sampling was performed by colleagues and operators at a number of collaborating Institutions and included the recording of the subject's place of residence. The subject was also asked to report the origin of his/her parents. Recent immigrants were excluded. A total of 40 villages/towns (hereafter "locations") were sampled (S1 Table and S1 Fig). In most cases a sample representing a location consisted of subjects with that residence, but in a minority of cases information on residence was collected at a finer scale, and residents in the neighbourhood were assigned to the nearest location.

As far as the proposed research did not involve any issue relevant for the donor's health, only a subset of the WMA Declaration of Helsinki and COE Oviedo Convention prescriptions were applicable and obeyed. For these reasons written consent was requested in most cases but, in some series collected before 1995, oral consent was considered sufficient and simply recorded in the corresponding log sheets (filed at the collecting Institutions). In all cases the consent included also storage and future use of the sample. Anonymized blood or DNA samples were then received at the corresponding author's laboratory. Extraction was performed immediately and DNA quantitated prior to storage at -20˚C. Concentration was not further re-checked before use in the present study. The study was prospectively examined and approved on November 21, 2014 by the intramural ethical committee (Comitato Etico

Indipendente—document number 0025422/2014), who expressly considered the list of collaborators, anonymity of samples and the compliance with consent regulations of previous publications which included the same samples.

The 40 locations comprised a group of residents in Moravia (Czech Republic) [71] as a reference for the Central European population (North-Western outgroup). In order to have an appropriate counterpart, we also included in the study (41st location) the Palestinians of the CEPH HGDP panel [73], as a South-Eastern outgroup.

STR profiles and quality controls

We obtained the genetic profiles at 16 autosomal STRs on 1984 individual assays using the AmpF*l*STR® NGM SElect™ PCR Amplification Kit (Life Technologies inc.) under the conditions recommended by the manufacturer, in 96-wells microtiter plates. All PCR products were separated with the same ABI PRISM® 3100-*Avant*™Genetic Analyzer, polymer and capillary types, and run conditions constant across the plate set. All plates included a negative (water) control, 8 replicates of the reference allelic ladder provided with the kit, as well the positive control provided by the manufacturer (Control DNA *007*) and our internal control. Electropherograms were generated with the GeneMapper® ID-X software, with allele nomenclature given in number of repeats at each locus. The following parameters were used: baseline = 51; analytical threshold = 50 RFU; stochastic threshold = 150 RFU. Profiles were inspected by two independent operators, considering Peak Height Ratios (PHR) <15% (together with appropriate molecular weight) for stutter peaks, and >50% for calling a heterozygous allele. This latter value is conservative and reduces the risk of dropping out true alleles. Independent spreadsheets were produced and compared. All discrepant results underwent a first round of reviewing. Profiles with missing amplification (signal below the analytical threshold) at one or more loci were discarded.

More than 300 subjects, spread over separate plates, were typed two or more times with 100% repeatability. These included 7 carriers of overladder alleles (S3 Table), in whom the variant peak was confirmed. In three subjects, a peak in between the D2S1045 and D19S433 bins was assigned to D19S433 based on peak height similarity within each locus, resulting in doubly heterozygous genotypes. As far as we analysed the CEPH Palestinians, we could directly compare our results with those already obtained in the same individuals [2]. Identical genotypes were obtained for the 4 loci shared between the two studies (D10S1248, D16S539, D19S433 and D22S1045), after translating allele sizes from bp to repeat numbers.

We also used the program RELPAIR [29] to detect hidden relatedness, with allele frequencies obtained in the whole series. Thresholds for the likelihood ratio took into account the number of pairwise comparisons within each location sample. This step led to the exclusion of 19 subjects, i.e. one member of each relative pair (Parent-Offspring = 12; Full Sibs = 7). Among full sibs, the relatedness of two Palestinians (HGDP00694 and HGDP00695) was detected, as previously reported [74].

Data analysis

Standard diversity indices. The potential presence of null alleles was checked with the program FreeNA [75]. Allele frequencies and fixation indexes were obtained with Arlequin v. 3.5.2.2 [76], both considering (for Fis) and not considering (for Fst) the individual level. The exact test for the HWE was performed with the same program and 1 million steps in the Markov chain. The matrix of pairwise Fst values for 11 location pools was represented by non-metric Multi Dimensional Scaling (nmMDS) with PAST v. 3.06 [77] and plotted with R. Pooling was as follows (numbers as in S1 Table and S1 Fig): Central Italy (2–5); Apulia (6, 7, 9–11);

Calabria and Sicily (8, 12–19); Western Greece (20–22); Eastern Greece (23–26); Western Crete (31, 32); Eastern Crete (30, 33, 34); North Eastern Mediterranean (35, 39–40); Aegean Islands (27–29); Black Sea (37); Central Anatolia (36–38).

The Mantel test as implemented in PASSaGE 2 [78] was used to test the correlation between the Fst and geographic distance (great circle) matrices.

Spatially explicit models. Spatial principal component analysis was performed with the R package adegenet [9]. This method uses the frequencies of alleles at all loci and performs a principal component analysis which takes into account simultaneously the frequency variance across locations and the similarity between locations connected in a spatial network as contrasted to those not connected (summarized by the Moran I index). It thus returns positive (and negative) scores according to whether closely spaced locations display higher (lower) similarities, thus distinguishing between global vs local frequency variation patterns. Geographic coordinates were those reported in S1 Table and a nearest neighbour (n = 12) connection network was used (S8 Fig). The distributions of squared loadings for the 288 total alleles on sPC's 1 and 2 were used to identify the strongest contributors to each sPC, represented in the top 1.25%, 2.50%, 5.0%, 7.5% and 10% quantiles, corresponding to 4, 8, 15, 22 and 29 alleles, respectively. Surface maps of the frequencies of the selected alleles (8 in each case, see Results) were constructed with the Kriging algorithm as implemented in the R package "fields", using a 200 x 200 rectangular grid spanning from the minimum to the maximum longitude and latitude values. The advantage of this method is that, at each sampling location, it returns the observed value (in this case the allele frequency) and thus the density of equal-frequency surfaces reflects the steepness of the expected cline(s) along lines connecting sampled locations. Coast contours and markers for sampling locations were overlaid with R functions. Correlograms of allele frequencies were obtained with PASSaGE 2 [78], using 6 distance classes (Upper bounds: 200, 400, 800, 1200, 1600 and 3200 km), chosen to obtain an appropriate number of observations within each class, and to render the shorter distances within than between each of the three main sampling countries (Italy, Greece and Turkey). Significance was tested with 10,000 permutations.

In order to retain most of the spatially informative aspects of the data while abating the background noise as compared to sPCA, we generated genotype datasets (reduced datasets) derived from sPC's 1 and 2. In each of them we recoded the individual genotypes considering the 8 alleles with the highest squared loadings on sPC's 1 and 2, respectively, against all the remaining alleles at the same locus [79]. Two independent spatially-explicit methods were applied to both datasets to describe affinities between location samples and patterns on anisotropic migration (barriers to/enhancements of migration).

The ability of the reduced datasets in detecting affinities between locations was assayed with the clustering algorithm implemented in the program Geneland [13]. This method employs a Bayesian approach that takes into account the spatial coordinates (in this case identical for all individuals of the same sampling location) of each genotyped individual to assign him/her to a population cluster among a number of clusters that can be bounded by the user. Both reduced datasets were then analysed under the uncorrelated model with 200,000 iterations with a thinning of 100 and an initial 20% burn-in, asking for the probability of assignment to either of two clusters (to directly compare with the qualitative result of sPCA). Such probabilities, obtained for each square of a 200 x 200 grid, were plotted from the output files with R functions as above.

In order to visualize estimated migration rates we used the program EEMS [10]. This program uses a regular triangular grid covering a polygon which embraces the entire geographic range of sampling. Each individual (in our case all individuals of the same sampling location) is assigned to the nearest vertex of the grid and the migration parameter m is estimated by

Bayesian inference for every edge of the grid. The processed output consists in maps in which colours of the estimated effective migration surface correspond to local deviations from isolation by distance: in particular effective migration is low in geographic regions where genetic similarity decays quickly. We considered a polygon extending from the North-Western to the South-Eastern outgroup and spanning over most of the Italian peninsula, the western Balkan peninsula and almost the entirety of Turkey (11 to 44 degrees of longitude and 50 to 30 of latitude). We performed a number of test runs to optimize the performance of the MCMC chain and to obtain a grid which could spatially resolve most of our locations. Eventually, we used a grid of 466 triangles (demes) which resulted in the reduction of the 41 location samples to 34 sampled demes, some locations being assigned to the same vertex and pooled. In order to use relaxed priors, the variances of the proposal distributions were increased 8x as compared to the defaults. Two independent runs of 2,000,000 steps, 50% burn-in and a thinning of 5000 were used for each reduced dataset. Postprocessing was as recommended by the authors.

Note that the first spatially explicit method allows for several types of connection networks. We paid attention in using a nearest neighbour network (n = 12, S8 Fig) to differentiate with the next two methods, which use a Delaunay triangulation (Geneland) and a triangular lattice (EEMS), respectively. In particular, the nearest neighbour and Delaunay schemes produce remarkably different connections, especially between locations at the outer boundaries of the area.

Coalescent simulations. In order to test the observed fixation indices against neutral evolutionary scenarios, coalescent simulations were obtained with Fastsimcoal2 [35]. We used two demographic models modified from that of Capocasa et al. [80] and graphically outlined in S9 Fig. In Model 1, a number of populations (initially set at 14 to simulate Italian locations) split from a large pool (100,000 current gene copies) which has been growing at a rate of 0.020/generation, i.e. a value valid also for pre-agricultural societies [81–83]. Generation time was assumed to be 29 years [84]. Splitting times (in generations before present) were sampled from a UNIFORM(224:276) distribution, to account for the long time of the spread of the Neolithic cultural package in the Italian peninsula [49], and deme sizes (in gene copies) were sampled from a UNIFORM(800:2400) distribution, to account for the autosomal effective size and larger census size as compared to Capocasa et al. [80]. These demes were set to grow at a rate of 0.017/generation, continuing to exchange gene copies with the main pool at rates of 1.0E-3 and 1.0E-4 for sending and receiving, respectively. The resulting distribution of the fixation index Fst was obtained for migration rates (m) between the 14 demes of 0, 0.0005, 0.001, 0.0025, 0.005, 0.01 and 0.02 (corresponding approximately to 0–30 migrant copies (M)/generation). A set of microsatellite loci with constrained number of alleles were modelled, to replicate the recoded loci described above. For each condition 2,000 simulations, with samples of size equal to the real samples, were run and analyzed with Arlequin v. 3.5.2.2. In Model 2, three additional demes were added to the scenario with m = 0.01, joining the set of 14 at 96 (2) and 18 (1) generations ago, respectively. These latter demes were simulated to exchange gene copies with the previous 14 at rates of 0, 0.001, 0.0025, 0.005 and 0.01.

Supporting Information

S1 Fig. Map of Southern Europe/North Eastern Mediterranean Sea, with the positioning of sampling locations numbered as in S1 Table.
(PDF)

S2 Fig. Scatterplot of Fis vs the estimated frequency of null alleles at 16 loci x 41 location samples.
(PDF)

S3 Fig. Barplots of the estimated frequencies of null alleles (top) and of the Fis values (bottom) averaged over loci for the 41 locations.
(PDF)

S4 Fig. Quantile-quantile plot of the probability values in 656 tests for HWE (16 loci x 41 location samples). The solid line indicates identity between calculated and expected values. The dotted line represents the significance level (nominal $\alpha = 0.05$) after Bonferroni correction.
(PDF)

S5 Fig. Scatterplot of the scores on the first three axes obtained by nmMDS based on the matrix of pairwise Fst values after grouping the 41 location samples into 13 geographical pools. Color shades from bright red to black refer to position on dimension 2.
(PDF)

S6 Fig. Frequency contour maps of 8 alleles producing the highest squared loadings on sPC 1 (page 1, red) and sPC 2 (page 3, green). The same alleles are listed in Table 1 in the 2.5% columns. Note the different colour scale used in each map. Values outside the polygon connecting the most external points are extrapolated. For each of the two map sets, the correlograms are shown in the same order (pages 2, 4). Black dots indicate significant class-specific values, i.e. individual values for which the null hypothesis (Moran I = 0) is rejected. The global significance of the correlograms consist in checking that at least one of the coefficients retains significance after considering multiple tests for distance classes with the Bonferroni correction. Ticks on the x axis are spaced to indicate the upper bounds of distance classes.
(PDF)

S7 Fig. Color maps (corresponding to Fig 1C and 1D) of assignment probabilities of the 41 locations to either of 2 population clusters obtained with Geneland on the reduced allele sets derived from sPC1 (A) and sPC2 (B).
(PDF)

S8 Fig. Map of Southern Europe/North Eastern Mediterranean Sea, with the nearest neighbour (n = 12) connection network used in adegenet.
(PDF)

S9 Fig. Scheme of the demographic models used for coalescent simulations. n = number of demes (only a subset shown for the sake of clarity); Ne = effective size (in gene copies); r = growth rate per generation. Black arrows represent instances of gene flow with the indicated fixed rate across simulations. Dashed arrows indicate instances of gene flow whose rate was varied across simulations.
(PDF)

S10 Fig. Distributions of simulated Fst values under the demographic scenarios of S9 Fig. A) Model 1: migration rates (m) among the 14 demes are shown; the curves for m = 0.02, 0.01 and 0.005 are shown in the inset for clarity. B) Model 2: the numbers indicate migration rates (m) between 3 recent demes and the 14 demes of Model 2. Note the different scale of the x axis as compared to panel A. In both panels the vertical lines indicate the Fst value obtained from real data.
(PDF)

S1 Table. List of sampling locations, geographic coordinates and sample sizes. For Italy and Greece the administrative region or the Island name is also reported.
(XLSX)

S2 Table. List of alleles not matching the AmpFLSTR® NGM SElect™ allelic ladder (over-ladder). Loci are in the order of increasing MW for the Blue, Green, Black and Red series.
(XLSX)

S3 Table. Spreadsheet with relative allele frequencies at 16 STR loci in the 41 location samples.
(XLSX)

S1 Text. We describe here three additional analyses, aimed at evaluating the robustness of our findings. The first one examines the effect of a 50% reduction of the sampled locations. The second and third one are based on external datasets and take into consideration forensic STRs for other samples within the same geographical frame, as well as publicly available genome-wide SNP data.
(PDF)

Acknowledgments

Author Contributions

Conceptualization: AN FM.

Data curation: FM.

Formal analysis: FM.

Funding acquisition: AN CJ.

Investigation: FM AF.

Methodology: AN FM CJ.

Project administration: AN.

Resources: AN AL EIM NA RB.

Supervision: AN.

Validation: FM AF.

Visualization: AN FM.

Writing – original draft: AN.

Writing – review & editing: AN FM CJ.

References

1. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, et al. Genetic structure of human populations. Science. 2002; 298(5602):2381–5. doi: 10.1126/science.1078311 PMID: 12493913

2. Pemberton TJ, DeGiorgio M, Rosenberg NA. Population structure in a comprehensive genomic data set on human microsatellite variation. G3: Genes|Genomes|Genetics. 2013; 3(5):891–907. doi: 10.1534/g3.113.005728 PMID: 23550135

3. Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW. Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet. 2005; 1(6): e70. doi: 10.1371/journal.pgen.0010070 PMID: 16355252

4. Serre D, Pääbo S. Evidence for gradients of human genetic diversity within and among continents. Genome Res. 2004; 14(9):1679–85. doi: 10.1101/gr.2529604 PMID: 15342553

5. Rowold DJ, Herrera RJ. Inferring recent human phylogenies using forensic STR technology. Forensic Sci Intl. 2003; 133(3):260–5.

6. Silva NM, Pereira L, Poloni ES, Currat M. Human neutral genetic variation and forensic STR data. PLoS ONE. 2012; 7(11):e49666. doi: 10.1371/journal.pone.0049666 PMID: 23185401

7. Steele CD, Court DS, Balding DJ. Worldwide estimates relative to five continental-scale populations. Ann Hum Genet. 2014; 78(6):468–77. doi: 10.1111/ahg.12081 PMID: 26460400

8. Algee-Hewitt BFB, Edge MD, Kim J, Li JZ, Rosenberg NA. Individual identifiability predicts population identifiability in forensic microsatellite markers. Cur Biol. 2016; 26(7):935–42.

9. Jombart T. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics. 2008; 24(11):1403–5. doi: 10.1093/bioinformatics/btn129 PMID: 18397895

10. Petkova D, Novembre J, Stephens M. Visualizing spatial population structure with estimated effective migration surfaces. Nat Genet. 2015; 48(1):94–100. doi: 10.1038/ng.3464 PMID: 26642242

11. Yang W-Y, Novembre J, Eskin E, Halperin E. A model-based approach for analysis of spatial structure in genetic data. Nat Genet. 2012; 44(6):725–31. doi: 10.1038/ng.2285 PMID: 22610118

12. Bradburd GS, Ralph PL, Coop GM. A spatial framework for understanding population structure and admixture. PLoS Genet. 2016; 12(1):e1005703. doi: 10.1371/journal.pgen.1005703 PMID: 26771578

13. Guillot G, Mortier F, Estoup A. Geneland: A computer package for landscape genetics. Mol Ecol Notes. 2005; 5:712–5.

14. Manel S, Berthoud F, Bellemain E, Gaudeul M, Luikart G, Swenson JE, et al. A new individual-based spatial approach for identifying genetic discontinuities in natural populations. Mol Ecol. 2007; 16 (10):2031–43. doi: 10.1111/j.1365-294X.2007.03293.x PMID: 17498230

15. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000; 155(2):945–59. PMID: 10835412

16. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009; 19(9):1655–64. doi: 10.1101/gr.094052.109 PMID: 19648217

17. Jombart T, Devillard S, Dufour AB, Pontier D. Revealing cryptic spatial patterns in genetic variability by a new multivariate method. Heredity. 2008; 101(1):92–103. doi: 10.1038/hdy.2008.34 PMID: 18446182

18. Dupanloup I, Schneider S, Excoffier L. A simulated annealing approach to define the genetic structure of populations. Mol Ecol. 2002; 11(12):2571–81. PMID: 12453240

19. Guillot G, Estoup A, Mortier F, Cosson JF. A spatial statistical model for landscape genetics. Genetics. 2005; 170(3):1261–80. doi: 10.1534/genetics.104.033803 PMID: 15520263

20. Mellars P. Palaeoanthropology: the earliest modern humans in Europe. Nature. 2011; 479(7374):483–5. doi: 10.1038/479483a PMID: 22113689

21. Pinhasi R, Thomas MG, Hofreiter M, Currat M, Burger J. The genetic history of Europeans. Trends Genet. 2012; 28(10):496–505. doi: 10.1016/j.tig.2012.06.006 PMID: 22889475

22. Rowley-Conwy P. Westward Ho! Curr Anthrop. 2011; 52(S4):S431–S51.

23. Brandt G, Szécsényi-Nagy A, Roth C, Alt KW, Haak W. Human paleogenetics of Europe—The known knowns and the known unknowns. J Hum Evol. 2015; 79:73–92. doi: 10.1016/j.jhevol.2014.06.017 PMID: 25467114

24. Lacan M, Keyser C, Ricaut Fo-X, Brucato N, Duranthon F, Guilaine J, et al. Ancient DNA reveals male diffusion through the Neolithic Mediterranean route. Proc Natl Acad Sci USA. 2011; 108(24):9788–91. doi: 10.1073/pnas.1100723108 PMID: 21628562

25. King RJ, DiCristofaro J, Kouvatsi A, Triantaphyllidis C, Scheidel W, Myres N, et al. The coming of the Greeks to Provence and Corsica: Y-chromosome models of archaic Greek colonization of the western Mediterranean. BMC Evol Biol. 2011; 11(1):69.

26. Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, et al. A genetic atlas of human admixture history. Science. 2014; 343(6172):747–51. doi: 10.1126/science.1243518 PMID: 24531965

27. Ralph P, Coop G. The geography of recent genetic ancestry across Europe. PLoS Biol. 2013; 11(5):e1001555. doi: 10.1371/journal.pbio.1001555 PMID: 23667324

28. Davis C, Ge J, King J, Malik N, Weirich V, Eisenberg AJ, et al. Variants observed for STR locus SE33: A concordance study. Forensic Sci Intl: Genetics. 2012; 6(4):494–7.

29. Epstein MP, Duren WL, Boehnke M. Improved inference of relationships for pairs of individuals. Am J Hum Genet. 2000; 67(5):1219–31. doi: 10.1016/S0002-9297(07)62952-8 PMID: 11032786

30. Green RL, Lagacé RE, Oldroyd NJ, Hennessy LK, Mulero JJ. Developmental validation of the AmpF*l*STR® NGM SElect™ PCR Amplification Kit: A next-generation STR multiplex with the SE33 locus. Forensic Sci Intl: Genetics. 2013; 7(1):41–51.

31. Pemberton TJ, Rosenberg NA. Population-genetic influences on genomic estimates of the inbreeding coefficient: a global perspective. Hum Hered. 2014; 77(1–4):37–48. doi: 10.1159/000362878 PMID: 25060268

32. Jakobsson M, Edge MD, Rosenberg NA. The relationship between Fst and the frequency of the most frequent allele. Genetics. 2013; 193(2):515–28. doi: 10.1534/genetics.112.144758 PMID: 23172852

33. Haining R. Spatial data analysis. Cambridge, UK: Cambridge University Press; 2003.

34. Biswas S, Scheinfeldt LB, Akey JM. Genome-wide insights into the patterns and determinants of fine-scale population structure in humans. Am J Hum Genet. 2009; 84(5):641–50. doi: 10.1016/j.ajhg.2009.04.015 PMID: 19442770

35. Excoffier L, Foll M. fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. Bioinformatics 2011; 27(9):1332–4. doi: 10.1093/bioinformatics/btr124 PMID: 21398675

36. Tagarelli G, Fiorini S, Piro A, Luiselli D, Tagarelli A, Pettener D. Ethnicity and biodemographic structure in the Arbëreshe of the province of Cosenza, Southern Italy, in the XIX century. Coll Antropol. 2007; 31(1):331–8. PMID: 17598420

37. Myres NM, Rootsi S, Lin AA, Järve M, King RJ, Kutuev I, et al. A major Y-chromosome haplogroup R1b Holocene era founder effect in Central and Western Europe. Eur J Hum Genet. 2011; 19(1):95–101. doi: 10.1038/ejhg.2010.146 PMID: 20736979

38. Semino O, Magri C, Benuzzi G, Lin AA, Al-Zahery N, Battaglia V, et al. Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area. Am J Hum Genet. 2004; 74(5):1023–34. doi: 10.1086/386295 PMID: 15069642

39. Di Giacomo F, Luca F, Anagnou N, Ciavarella G, Corbo RM, Cresta M, et al. Clinal patterns of human Y chromosomal diversity in continental Italy and Greece are dominated by drift and founder effects. Mol Phyl Evol. 2003; 28(3):387–95.

40. Di Giacomo F, Luca F, Popa LO, Akar N, Anagnou N, Banyko J, et al. Y chromosomal haplogroup J as a signature of the post-neolithic colonization of Europe. Hum Genet. 2004; 115(5):357–71. doi: 10.1007/s00439-004-1168-9 PMID: 15322918

41. Malaspina P, Tsopanomichalou M, Duman T, Stefan M, Silvestri A, Rinaldi B, et al. A multistep process for the dispersal of a Y chromosomal lineage in the Mediterranean area. Ann Hum Genet. 2001; 65(Pt 4):339–49. doi: doi:10.1017/S0003480001008727 PMID: 11592923

42. Balaresque P, Bowden GR, Adams SM, Leung H-Y, King TE, Rosser ZH, et al. A predominantly Neolithic origin for European paternal lineages. PLoS Biol. 2010; 8(1):e1000285. doi: 10.1371/journal.pbio.1000285 PMID: 20087410

43. King RJ, Ozcan SS, Carter T, Kalfoglu E, Atasoy S, Triantaphyllidis C, et al. Differential Y-chromosome Anatolian influences on the Greek and Cretan Neolithic. Ann Hum Genet. 2008; 72(Pt 2):205–14. doi: 10.1111/j.1469-1809.2007.00414.x PMID: 18269686

44. Cinnioglu C, King R, Kivisild T, Kalfoglu E, Atasoy S, Cavalleri GL, et al. Excavating Y-chromosome haplotype strata in Anatolia. Hum Genet. 2004; 114(2):127–48. doi: 10.1007/s00439-003-1031-4 PMID: 14586639

45. Zalloua PA, Platt DE, El Sibai M, Khalife J, Makhoul N, Haber M, et al. Identifying genetic traces of historical expansions: Phoenician footprints in the Mediterranean. Am J Hum Genet. 2008; 83(5):633–42. doi: 10.1016/j.ajhg.2008.10.012 PMID: 18976729

46. Martinez L, Underhill PA, Zhivotovsky LA, Gayden T, Moschonas NK, Chow CE, et al. Paleolithic Y-haplogroup heritage predominates in a Cretan highland plateau. Eur J Hum Genet. 2007; 15(4):485–93. doi: 10.1038/sj.ejhg.5201769 PMID: 17264870

47. Paschou P, Drineas P, Yannaki E, Razou A, Kanaki K, Tsetsos F, et al. Maritime route of colonization of Europe. Proc Natl Acad Sci USA. 2014; 111(25):9211–6. doi: 10.1073/pnas.1320811111 PMID: 24927591

48. Fernàndez E, Pérez-Pérez A, Gamba C, Prats E, Cuesta P, Anfruns J, et al. Ancient DNA analysis of 8000 B.C. Near Eastern farmers supports an Early Neolithic pioneer maritime colonization of mainland Europe through Cyprus and the Aegean Islands. PLoS Genet. 2014; 10(6):e1004401. doi: 10.1371/journal.pgen.1004401 PMID: 24901650

49. Malone C. The Italian Neolithic: a synthesis of research. J World Prehist. 2003; 17(3):235–312.

50. Gazal S, Sahbatou M, Babron M-C, Génin E, Leutenegger A-L. High level of inbreeding in final phase of 1000 Genomes Project. Sci Rep. 2015; 5:17453. doi: 10.1038/srep17453 PMID: 26625947

51. Roewer L, Croucher PJ, Willuweit S, Lu TT, Kayser M, Lessig R, et al. Signature of recent historical events in the European Y-chromosomal STR haplotype distribution. Hum Genet. 2005; 116(4):279–91. doi: 10.1007/s00439-004-1201-z PMID: 15660227

52. Cavalli-Sforza LL, Menozzi P, Piazza A. The history and geography of human genes. Princeton, N.J.: Princeton University Press; 1994.

53. Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature. 2014; 513(7518):409–13. doi: 10.1038/nature13673 PMID: 25230663

54. Novembre J, Stephens M. Interpreting principal component analyses of spatial population genetic variation. Nat Genet. 2008; 40(5):646–9. doi: 10.1038/ng.139 PMID: 18425127

55. Yunusbayev B, Metspalu M, Järve M, Kutuev I, Rootsi S, Metspalu E, et al. The Caucasus as an asymmetric semipermeable barrier to ancient human migrations. Mol Biol Evol. 2011; 29(1):359–65. doi: 10.1093/molbev/msr221 PMID: 21917723

56. Behar DM, Yunusbayev B, Metspalu M, Metspalu E, Rosset S, Parik J, et al. The genome-wide structure of the Jewish people. Nature. 2010; 466(7303):238–42. doi: 10.1038/nature09103 PMID: 20531471

57. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al. Genes mirror geography within Europe. Nature. 2008; 456(7218):98–101. doi: 10.1038/nature07331 PMID: 18758442

58. Wang C, Zöllner S, Rosenberg NA. A quantitative comparison of the similarity between genes and geography in worldwide human populations. PLoS Genet. 2012; 8(8):e1002886. doi: 10.1371/journal.pgen.1002886 PMID: 22927824

59. Wang C, Szpiech ZA, Degnan JH, Jakobsson M, Pemberton TJ, Hardy JA, et al. Comparing spatial maps of human population-genetic variation using Procrustes analysis. Stat Appl Genet Mol Biol. 2010; 9:Article 13.

60. Nelis M, Esko T, Mägi R, Zimprich F, Zimprich A, Toncheva D, et al. Genetic structure of Europeans: A view from the North-East. PLoS ONE. 2009; 4(5):e5472. doi: 10.1371/journal.pone.0005472 PMID: 19424496

61. Di Gaetano C, Voglino F, Guarrera S, Fiorito G, Rosa F, Di Blasio AM, et al. An overview of the genetic structure within the Italian population from genome-wide data. PLoS ONE. 2012; 7(9):e43759. doi: 10.1371/journal.pone.0043759 PMID: 22984441

62. Capelli C, Brisighelli F, Scarnicci F, Arredi B, Caglia A, Vetrugno G, et al. Y chromosome genetic variation in the Italian peninsula is clinal and supports an admixture model for the Mesolithic-Neolithic encounter. Mol Phyl Evol. 2007; 44(1):228–39.

63. Sarno S, Boattini A, Carta M, Ferri G, Alù M, Yao DY, et al. An ancient Mediterranean melting pot: Investigating the uniparental genetic structure and population history of Sicily and Southern Italy. PLoS ONE. 2014; 9(4):e96074. doi: 10.1371/journal.pone.0096074 PMID: 24788788

64. Hamilton G, Stoneking M, Excoffier L. Molecular analysis reveals tighter social regulation of immigration in patrilocal populations than in matrilocal populations. Proc Natl Acad Sci USA. 2005; 102(21):7476–80. doi: 10.1073/pnas.0409253102 PMID: 15894624

65. Fix AG. Migration and colonization in human microevolution. Weiss KM, editor. Cambridge, U.K.: Cambridge University Press; 1999.

66. Cavalli-Sforza LL, Bodmer WF. The genetics of human populations. Kennedy D, Park RB, editors. San Francisco: W.H. Freeman & co.; 1971.

67. Di Gaetano C, Cerutti N, Crobu F, Robino C, Inturri S, Gino S, et al. Differential Greek and northern African migrations to Sicily are supported by genetic evidence from the Y chromosome. Eur J Hum Genet. 2008; 17(1):1–9.

68. Tofanelli S, Brisighelli F, Anagnostou P, Busby GBJ, Ferri G, Thomas MG, et al. The Greeks in the West: genetic signatures of the Hellenic colonisation in southern Italy and Sicily. Eur J Hum Genet. 2016; 24(3):429–36.

69. Rubi-Castellanos R, Martínez-Cortés G, Francisco Muñoz-Valle J, González-Martín A, Cerda-Flores RM, Anaya-Palafox M, et al. Pre-Hispanic Mesoamerican demography approximates the present-day ancestry of Mestizos throughout the territory of Mexico. Am J Phys Anthropol. 2009; 139(3):284–94. doi: 10.1002/ajpa.20980 PMID: 19140185

70. Batini C, Hallast P, Zadik D, Delser PM, Benazzo A, Ghirotto S, et al. Large-scale recent expansion of European patrilineages shown by population resequencing. Nat Commun. 2015; 6:7152. doi: 10.1038/ncomms8152 PMID: 25988751

71. Luca F, Di Giacomo F, Benincasa T, Popa LO, Banyko J, Kracmarova A, et al. Y-chromosomal variation in the Czech Republic. Am J Phys Anthropol. 2007; 132(1):132–9. doi: 10.1002/ajpa.20500 PMID: 17078035

72. Malaspina P, Cruciani F, Santolamazza P, Torroni A, Pangrazio A, Akar N, et al. Patterns of male-specific inter-population divergence in Europe, West Asia and North Africa. Ann Hum Genet. 2000; 64(Pt 5):395–412. PMID: 11281278

73. Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, et al. A human genome diversity cell line panel. Science. 2002; 296(5566):261–2. PMID: 11954565

74. Rosenberg NA. Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann Hum Genet. 2006; 70 (6):841–7.

75. Chapuis M-P, Estoup A. Microsatellite null alleles and estimation of population differentiation. Mol Biol Evol. 2007; 24(3):621–31. doi: 10.1093/molbev/msl191 PMID: 17150975

76. Excoffier L, Lischer HE. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour. 2010; 10(3):564–7. Epub 2011/05/14. doi: 10.1111/j.1755-0998.2010.02847.x PMID: 21565059

77. Hammer Ǿ, Harper DAT, Ryan PD. PAST:Paleontological statistics software package for education and data analysis. Paleontologia Electronica. 2001; 4:9.

78. Rosenberg MS. PASSAGE: Pattern analysis, spatial statistics and geographic exegesis. v. 1.3. Dept. of Biology, Arizona State University; 2001.

79. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006; 2(12): e190. doi: 10.1371/journal.pgen.0020190 PMID: 17194218

80. Capocasa M, Battaggia C, Anagnostou P, Montinaro F, Boschi I, Ferri G, et al. Detecting genetic isolation in human populations: A study of European language minorities. PLoS ONE. 2013; 8(2):e56371. doi: 10.1371/journal.pone.0056371 PMID: 23418562

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

82. Boone JL. Subsistence strategies and early human population history: an evolutionary ecological perspective. World Archaeol. 2002; 34(1):6–25. doi: 10.1080/00438240220134232 PMID: 16475305

83. Hamilton MJ, Burger O, DeLong JP, Walker RS, Moses ME, Brown JH. Population stability, cooperation, and the invasibility of the human species. Proc Natl Acad Sci USA. 2009; 106(30):12255–60. doi: 10.1073/pnas.0905708106 PMID: 19592508

84. Fenner JN. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am J Phys Anthropol. 2005; 128(2):415–23. doi: 10.1002/ajpa.20188 PMID: 15795887