Research paper

Ancient ancestry informative markers for identifying fine-scale ancient population structure in Eurasians

Umberto Esposito[1], Ranajit Das[2], Mehdi Pirooznia[3], and Eran Elhaik[1*]

[1] Department of Animal and Plant Sciences, University of Sheffield, Sheffield, UK S10 2TN

[2] Manipal University, Manipal Centre for Natural Sciences (MCNS), Manipal, Karnataka, India

[3] Johns Hopkins University, Department of Psychiatry and Behavioral Sciences, Baltimore, MD, USA 21205

* Please address all correspondence to Eran Elhaik at e.elhaik@sheffield.ac.uk

Keywords: ancient DNA, ancient ancestry informative markers, population structure, PCA, admixture mapping

Abstract

The rapid accumulation of ancient human genomes from various places and time periods, mainly from the past 15,000 years, allows us to probe the past with an unparalleled accuracy and reconstruct trends in human biodiversity. Alongside providing novel insights into the population history, population structure permits correcting for population stratification, a practical concern in gene mapping in association studies. However, it remains unclear which markers best capture ancient population structure as not all markers are equally informative. Moreover, the high missingness rates in ancient, oftentimes haploid, DNA, may distort the population structure and prohibit genomic comparisons. In past studies, ancestry informative markers (AIMs) were harnessed to address such problems, yet whether AIMs finding methods are applicable to aDNA remains unclear. Here, we define ancient AIM (aAIMs) and develop a framework to evaluate established and novel AIMs-finding methods. We show that a novel principal component analysis (PCA)-based method outperforms all methods in capturing ancient population structure and identifying admixed individuals. Our results highlight important features of the genetic structure of ancient Eurasians and the choice of strategies to identify informative markers. This work can inform the design and interpretation of population and medical studies employing ancient DNA.

Author summary

Ancient DNA studies aim to identify geographical origin, migration routes, and disease susceptibility genes through the analysis of genetic markers such as single nucleotide polymorphisms (SNPs) in growing cohorts of ancient data. In addition to the existence of sub-structure in the studied population (i.e., differences in ancestry), ancient DNA suffers from high missingness rates and is oftentimes haploid, which may distort the inferred population structure and lead to spurious results. It is thereby imperative to address this possible bias by identifying the most accurate population structure. Due to the success of past studies in addressing similar problems using ancestry informative markers (AIMs), we defined ancient ancestry informative markers (aAIMs) that like AIMs can be used to interrogate ancient population structure. To find aAIMs, we designed a framework to evaluate established and novel AIMs-finding methods. We developed a database of 150,278 autosomal SNPs from 302 ancient genomes and 21 populations recovered from Europe, the Middle East, and North Eurasia dated to time periods from 14,000 to 1,500 years ago. We then applied two existing and three novel AIMs-finding methods and compared their performances against the complete dataset. We found that a novel principal component analysis (PCA)-based method captured the ancient population structure most accurately. Importantly, we introduce here a novel concept of aAIMs, a novel method that effectively identifies aAIMs, and a framework to compare the performances of AIMs. The outcome of our studies can improve the accuracy of genetic studies employing ancient DNA.

Introduction

Population stratification or geographic variation are a major concern in population, biomedical, and evolutionary studies. In genetic association studies, mismatching cases and controls introduces genetic heterogeneity that can lead to spurious associations and obscure the true association [1, 2]. In large groups the stratification bias may be less pronounced, but it is practically unavoidable in the case of rare diseases due to the difficulties in recruiting genetically homogeneous participants [3]. These problems are particularly challenging since the human population structure itself remain contentious. Nonetheless, is now clear that conquering population structure requires considering ancient DNA (aDNA) [4-6].

The advent of next-generation sequencing and the availability of large-scale genomic data and genotyping techniques have facilitated investigations of genomic variability that are central to understanding our evolutionary history and genomic origins. Over the last decade a plethora of ancient human genome sequencing projects have been accomplished, generating more than a thousand ancient genomes [7]. The revolution in aDNA sequencing has aided in investigations of ancient human migration, human adaptation, agricultural lifestyle, and disease co-evolution [7]. Notwithstanding its usefulness in delineating the evolutionary history of mankind, aDNA data can be problematic due to its haploidity and high missingness [6], which require having a large number of SNPs to infer population structure. However, SNPs are not equally informative and may distort the population structure. The plethora of mismatching

markers sequenced in different genomes has revived the "of AIMs matrix" problem and the difficulty of comparing genomes. These problems are not new and rather reminiscent of the early stages of human population genetics. Then, one of the most successful solutions was using ancestry informative markers (AIMs).

AIMs are SNPs which exhibit large variation in minor allele frequencies (MAF) among populations. Over the past two centuries DNA studies scour genomes for these genetic patterns and produced numerous AIM sets for various purposes including determining an individual's ancestry, detecting stratification in biomedical studies, inferring geographic structure, and localizing biogeographical origins [e.g., 8, 9-12]. AIM panels can delineate population structure in a cost effective manner by identifying population specific markers, which in turn help in detecting and correcting for variation in individual ancestry that can confound methods like admixture mapping, Mendelian Randomization trials, association studies, and forensics by increasing false positive results and/or reducing Power [e.g., 13, 14, 15]. In the case of genetic association studies, AIMs-based solution has been preferred over methods like genomic control (GC) correction, which is only applicable in genome-wide scale data [16]. However, it remains uncertain which AIMs to use since all AIMs panels have limitations [17] and their applicability to ancient genomes was never tested. The characteristics of ideal AIMs are remain contentious with some authors preferring common SNPs (minor allele frequency >1%) [16], SNPs with high F_{ST} [18], SNPs with high pairwise MAF between populations [17], or SNPs that satisfy several criteria. Consequently, AIMs may not overlap across studies that focus on particular populations and even those reported in global studies do not necessarily

overlap. Finally, studies typically show that AIMs can separate populations or broadly classify individuals into subcontinental populations, rather than capture the population structure of the complete SNP set or allow fine-population mapping. Given the uncertainties surrounding AIMs, their potential incompatibility to capture ancient structure and admixtures, and the challenges imposed by aDNA data, it is unclear whether, if at all, AIMs-finding methods or AIMs can be utilized to study ancient population structure.

In this study, we defined ancient ancestry informative markers (aAIMs) as SNPs that vary in their MAF across ancient populations (Figure 1) and attempted to identify and validate the first autosomal aAIMs. Since AIMs-finding tools were never tested on aDNA, it was necessary to compare their ability in finding aAIMs. For that, we interrogated a comprehensive dataset of 302 ancient genomes classified to 21 populations from Europe, the Middle East, and North Eurasia. This dataset was used to compare two existing AIMs finding algorithms: Infocalc [19] and Wright's F_{ST} [20, 21], three novel Admixture- and PCA-based algorithms, and two random SNP sets in identifying aAIMs that can capture the population structure and identify admixed individuals. Our study offers a methodological framework to evaluate AIMs, contrasts different strategies to find aAIMs, and reports the first set of aAIMs.

Results

Ancient genomic data

We developed a framework to identify and evaluate the efficacy of aAIM candidates in capturing ancient population structure and allow admixture mapping (Figure 2). We constructed a dataset of 150,278 autosomal SNPs from 302 ancient genomes and 21 populations recovered from Europe, the Middle East, and North Eurasia dated to time periods spanning the past 14,000 years till 1,500 years ago (Figure 3, Table S1). Due to the limited availability of ancient genomes, our dataset was inconsistent over time and space. For instance, there were 57 Central European genomes from the Late Neolithic to the Bronze Age, but populations, such as, Central Western Mesolithic Europeans, Bronze Age Jordanians, Chalcolithic, and Mesolithic Russians, comprised of three genomes each.

Missingness largely varied between the samples and markers. The sample-based missingness ranged from 0.05% (KK1) to 99.2% (I1951) with an average of 54%. The sample-based missingness varied among populations with Levantine Epipaleolithic Neolithic genomes having the highest missingness (n=19, μ=90±16%) and Mesolithic Swedish genomes having the lowest one (n=8, μ=29±27%). The variant-based missingness ranged from 30% to 98% with an average of 54%.

Principal component analysis (PCA) of the ancient genomes substantiated previous observations of a Europe–Middle East contrast along the vertical principal component

(PC1) and parallel clines (PC2) in both Europe and the Middle East (Figure 4). Genomes from the Epipaleolithic and Neolithic Levantine clustered at one extreme of the Near East-Europe cline with some overlapping with Neolithic Turkish and Central European genomes. Neolithic Iranians clustered between Central European genomes. While ancient Spanish, Armenian, Central EU, and British genomes occupied the intermediate position of Near Eastern and North Eurasian genomes, Russian and Swedish genomes clustered at the end of the Near East-Europe cline.

We next applied an unsupervised ADMIXTURE analysis to the dataset. Analyzing the results generated with various number of splits (*K*) (Figure S1), no choice of *K* minimized the cross-validation error (CVE) (Figure S2), likely because the high noise and missingness in the data prevented the CVE from stabilizing. We observed that at *K*=10, multiple genomes (e.g., Britain Iron Saxon, Mesolithic Neolithic Caucasus population, Bronze Age Jordanian and Epipaleolithic Levantine, Chalcolithic, Mesolithic and Early Mid Bronze Russian, Early Neolithic Spanish, Mesolithic and Mid Neolithic Swedish, and Neolithic Turkish) appeared homogeneous in relation to their population and assigned to a distinct allele frequency profile or admixture components (Figure 5). In these figure, putative ancient ancestral components, like the *Early Neolithic European* (brown) and *Russia Mid Late Bronze* (magenta), predominantly found among European genomes, may be identified. Except their predominance in Neolithic Turkish genomes, these components also exist in most Neolithic Central Europeans. Some 20-30% of Central European genomes have discernible fractions of *Europe Late Neolithic-Early Bronze* (navy-blue) and *Russia Mid-Late Bronze* (deep-pink) components, respectively.

Two components (cyan and dark purple) appeared sporadically in a few populations, likely due to noise.

Identifying aAIM candidates

To identify aAIM candidates, we employed Infocalc and F_{ST}, commonly used to detect AIMs. We also implemented three novel Admixture- and PCA-based methods (Admixture$_1$, Admixture$_2$, and PCA-derived [PD]). Finally, we selected two random SNP sets of 10,000 and 15,000 markers, which approximated the number of AIMs identified by the various methods (Rand$_{10k}$ and Rand$_{15k}$). Four criteria were adopted to evaluate how the candidate aAIMs capture the population structure depicted by the complete SNP set (CSS): first, by qualitatively comparing the dispersal of genomes obtained from a PCA to that of the CSS. Second, by comparing the Euclidean distances between the admixture proportions of each genome and those obtained from the CSS. To avoid inconsistencies between the SNP sets, we used admixture components obtained through a *supervised* ADMIXTURE (see *methods*). Third, by testing which aAIMs classify individuals to populations most accurately. The abilities to identify admixed individuals and evaluated for the top performing method.

As with the CSS, genomes with over 90% missingness were removed, leaving each dataset with 223-263 genomes (Table S2). 310 SNPs without data were removed from the Rand$_{10k}$ dataset. The final number of aAIM candidates identified using each method is shown in Table S3. Overlapping aAIMs between the methods are remarkably small and range from 560 (Rand$_{10k}$ and Admixture$_1$) to 2,160 (Admixture$_1$ and Admixture$_2$).

206 Interestingly, Infocalc and F_{ST}, oftentimes used together share only ~10% of their aAIM
207 candidates. The PD method shares 13.7% of its aAIMs with F_{ST} and ~10% with Infocalc.

208 Comparing the sequence properties of the aAIM candidates, we found that for ancient
209 populations (Figure 6a) Infocalc's aAIMs mirrored the MAF of the CSS with most
210 variants having low MAF (45% of the aAIMs have MAF<0.1). The F_{ST} aAIM also had
211 high frequency of low-mid MAF values. By contrast, the PD and Admixture-based
212 methods exhibited higher frequencies of high MAF SNPs with $Admixture_2$ having the
213 highest proportion of high MAF aAIMs (75% of the aAIMS have MAF>0.4).
214 Interestingly, the MAF distributions exhibited similar distributions in modern populations
215 (Figure 6b), though with fewer alleles in lowest MAF bins for all the methods.
216 Unsurprisingly, most of the aAIM variants were non-functional (94.6-96.3%) and vary
217 little from the CSS's annotation (Table S4).
218
219 ## Comparative testing of aAIM candidates
220
221 We compared the performances of aAIMs candidates to each other and to the CSS in
222 capturing the population structure and classifying individuals to populations through
223 three analyses. First, we calculated the PCA for each SNP set and compared the
224 population dispersion along the primary two axis. Similarly to the CSS (Figure 4), all the
225 methods depicted the Europe–Middle East contrast (PC1) and parallel clines (PC2) in the
226 European genomes so that ancient Jordanian, Levantine, Turkic, and Spanish genomes
227 clustered at one extreme of the Near East-Europe cline, whereas the genomes from
228 Russia and Sweden clustered at the other end (Figure S3). However, much like the

229 random sets, Infocalc and F_{ST} did not separate Levantine and Turkish individuals from
230 each other. Infocalc also merged the Caucasus individuals with central Europeans. The
231 admixture-based methods and PD separated all the ancient populations, similarly to the
232 CSS and better, in the case of Scandinavians and Russians.
233
234 We next quantitatively assessed which dataset produced the closest admixture signature
235 to that of the CSS (Figures 5). For that, we calculated the admixture proportions in
236 relation to ten putatively ancient ancestral populations (Figures S4-5) and then computed
237 their Euclidean distances to their counterparts obtained by the CSS (Figure 7). The PD
238 aAIMs had significantly short Euclidean distances (μ=0.13, σ= 0.1, n=302) compared to
239 all other aAIMs (Welch t-test p-values: Infocalc 0.002, F_{ST} $8.5\text{x}10^{-13}$, Admixture_1 $2.2\text{x}10^{-}$
240 ^{16}, Admixture_1 $2\text{x}10^{-16}$, Rand_{10k} $5\text{x}10^{-6}$, and Rand_{15k} 0.001). Infocalc's aAIMs produced
241 the second shortest distances from the CSS (μ=0.17, σ=0.15), however they were not
242 statistically shorter than the distances obtained by the two random datasets (Welch t-test
243 p-values: Rand_{10k} 0.12 and Rand_{15k} 0.77 respectively), suggesting that Infocalc was
244 unable to capture the population structure. F_{ST}-derived AIMs (μ=0.2, σ=0.13) performed
245 worse than the Rand_{15k} aAIMs (Welch t-test p-value 0.004), and similarly to the Rand_{10k}
246 aAIMs (Welch t-test, p-value=0.13). The admixture-based datasets performed worst of all
247 aAIMs (μ_1=0.22, σ_1=0.15 and μ_2=0.24, σ_1=0.16) and significantly worse than the two
248 random datasets (Welch t-test: Admixture_1 [Rand_{10k} p-value=0.002] and [Rand_{15k} p-
249 value=$1.6\text{x}10^{-5}$]; Admixture_2 [Rand_{10k} p-value=$1.7\text{x}10^{-5}$] and [Rand_{15k} p-value=$2.5\text{x}10^{-8}$]).
250

We last assessed which aAIMs dataset allows classifying individuals to population groups most accurately. For that, an admixture-based population classifier was applied to the admixture proportions produced by all the datasets and their accuracy was compared to that of the CSS (76±25%) and the known population classification (Table S1). The mean classification accuracy per population ranged from 3% (F_{ST}) to 61% (PD) with the PD outperforming all other methods (Table 1). In other words, ~13k (8%) of the SNPs are sufficiently informative to classify individuals to populations with 80% of the accuracy of the CSS. In nine out of 21 population groups (22% of the individuals) PD-based classification was similar or more accurate than the CSS. All other methods performed similarly or worse than the random SNP sets (42±22% and 50±23%) with Infocalc (50±23%) outperforming the remaining methods. Of note are the poor performances of F_{ST} aAIMs, likely due to the high sensitivity of F_{ST} to aDNA data. As expected, high missingness was associated with incorrect predictions (Figure S6). For example, the low-coverage low-quality Britain Anglo-Saxon genomes proved challenging for all the methods (0-40%) but predicted correctly with the CSS (100%). Due to the high accuracy of the PD aAIMs compared to the alternative datasets, we continued to analyze its aAIMs,

Inference of admixed samples

Admixture mapping is a powerful method of gene mapping to map phenotypic variation or diseases that show differential risk by ancestry and takes advantage of higher densities of genetic variants and extensions to admixed populations [22]. Thereby a large number

of markers throughout the genome is necessary to allow inference of local chromosomal ancestry blocks. Figure 8 illustrates the genome-wide distribution of PD aAIMs. To test whether these aAIMs can identify admixture in hybrid individuals, ancient individuals were hybridized to form 120 mixed individuals, each associated with three datasets: CSS, PD aAIMs, and a random SNP set of the size of PD aAIMs (Table 2).

The genetic distances between the CSS and PD aAIMs were significantly smaller (μ=0.05, σ=0.04) than the distances between the CSS and the random SNP sets (μ=0.45, σ=0.15, Welch *t*-test *p*-values=2.2×10^{-8}) as well as between the OD and the random SNP sets (μ=0.43, σ=0.15, Welch *t*-test *p*-values=1.9×10^{-8}). We, thus, demonstrated that PD aAIMs can be used to infer admixed individuals and be used in future admixture mapping involving aDNA.

Discussion

The use of ancient genomes in research is at its infancy and expected to intensify as data are becoming available. It is reasonable to expect that many of the tools employed to study modern-day genomes will need to be adapted to the ancient DNA environment. Some of the most useful tools in addressing population, biomedical, and evolutionary questions were ancestry informative markers (AIMs), however it is unclear whether they

297 are applicable to ancient genomic data, which not only represent populations with
298 different population structure, but has some unique characteristics like high missingness
299 and haploid genomes [6].
300
301 In this study, we defined ancient ancestry informative markers (aAIMs) (Figure 1) and
302 sought to identify those using various methods. The number of aAIMs identified by each
303 method ranges from 9 to 15 thousands. These numbers of the same magnitude of large
304 AIMs studies [e.g., 23, 24] and reasonable provided the potential relatedness of the
305 ancient populations and the near absence of heterozygote markers in the data. To find
306 which of the aAIMs candidates produced by each method best represent the true
307 population structure, we used the complete SNP set as a benchmark for qualitatively and
308 quantitatively comparisons.
309
310 Identifying the ideal AIMs set that would be both small and include redundancies (in case
311 of sequencing failure), capture the population structure, and allow identifying admixed
312 individuals remains one of the challenges of population genetics. We showed that aAIMs
313 identified through a PCA-derived (PD) method outperformed all other methods in
314 agreement with previous studies that tested PCA-based methods [16]. Some
315 classifications made by the PD were more accurate than those made using the CSS,
316 which highlights the negative influence of ancestry *uninformative* markers. To the best of
317 our knowledge, such markers and their influence were never explored. Infocalc and F_{ST}
318 aAIMs, typically used in conjunction to identify AIMs [10] and have been reported to
319 perform well in admixed populations [25] have oftentimes underperformed random

SNPs. Not only was F_{ST} already shown to be particularly small within continental populations [26], but these methods may be particularly sensitive to ancient DNA data that is both haploid and has high missingness (Figure S6). We also found no relationships between MAF and aAIMs performances (Figure 5). Enrichment for high or low MAF SNPs did not guarantee success, although the PD harbored more common SNPs than most underperforming methods.

The applicability of the PD aAIMs for admixture mapping combined with tools that can homogenize cases and controls [e.g., 27] enable future association studies to be carried out on ancient DNA samples. Indeed, Cassidy et al. [28] provided evidence for the existence of Hemochromatosis alleles in ancient genomes and point at the association of hemochromatosis alleles in ancient Irish. Due to the nature of the ancient data and to enable admixture mapping studies we refrained from optimizing the number of aAIMs. Further investigations with additional data may identify formerly common markers associated with the disease that with time became rare and undetectable.

Our study has several limitations. We studied an uneven number of Eurasian populations from various times and locations, causing a skew towards markers that predict central European populations from the Late-Neolithic Bronze Age. A modest attempt to reduce this bias was made by including modern-day African and Asian populations, however a more comprehensive analyses should be made when more global genomes are available. Second, the aAIMs were calculated independently by each method with individual populations considered independent, although the PCA and ADMIXTURE plots indicate

that central European populations may not be independent. Finally, due to high missingness of the data, it is likely that our study missed informative markers that could improve the classification accuracy in newly sequenced populations. We thereby advise applying our method to more comprehensive aDNA datasets when such will be available.

In summary, AIMs are some of the most effective tools that spear-headed population genetics over the past two decades and ancillary to the challenge of understanding population structure. We defined ancient AIMs (aAIMs), proposed a framework to evaluate AIMs-finding methods, demonstrated the accuracy of a novel aAIMs-finding method, and reported the most successful set of aAIMs. Future analyses may benefit from using our method to uncover powerful aAIMs and using our aAIMs to refine ancient population structure models.

Methods

Sample collection

Ancient DNA genomic data were obtained from 11 publications depicting 207 ancient genomes (Table S1). In the case of sequence data, sequence reads were aligned to the human reference assembly (UCSC hg19-http://genome.ucsc.edu/) using the Burrows Wheeler Aligner (BWA version 0.7.15) [29], allowing two mismatches in the 30-base seed. Alignments were then imported to binary (bam) format, sorted, and indexed using SAMtools (version 1.3.1) [30]. Picard (version 2.1.1) (http://picard.sourceforge.net/) was

366 then used for MarkDuplicates to remove reads with identical outer mapping coordinates
367 (which are likely PCA artifacts). The Genome Analysis Toolkit RealignerTargetCreator
368 module (GATK version 3.6) [31, 32] was used to generate SNP and small InDel calls for
369 the data within the targeted regions only. GATK InDelRealigner/BaseRecalibrator was
370 then used for local read realignment around known InDels and for base quality score
371 recalibration of predicted variant sites based on dbSNP 138 and 1000 Genomes known
372 sites, resulting in corrections for base reported quality. The recalibration was followed by
373 SNP/InDel calling with the GATK HaplotypeCaller. Variants were filtered for a
374 minimum confidence score of 30 and minimum mapping quality of 40. At the genotype
375 level, all genotypes that had a genotype depth less than 4 (GD < 4) or a genotype quality
376 score less than 10 (GQ < 10) were removed from the dataset by setting them to missing in
377 the VCF. GATK DepthofCoverage was then used to re-examine coverage following the
378 realignment. VCFtools (version 0.1.14) [33] were used to convert the VCF file to PLINK
379 format [34]. The final dataset comprised of 150,278 autosomal SNPs from 302 ancient
380 DNA (aDNA) genomes (Table S1; Additional file 1). Eight aDNA genomes (I0247,
381 I1584, I1955, ATP9, IR1, Kostenki14, MA1, and Ust_Ishim) without any country/region
382 designation were omitted in the closest population determination calculations. For
383 coherency, the genomes were divided into 21 populations, based on the sampling
384 country/region and their era.
385
386 Data analyses
387

388 The genetic structure canvas of ancient Eurasian genomes. The population structure of
389 the ancient genomes was described using principal component analysis (PCA)
390 implemented in PLINK v1.9 (https://www.cog-genomics.org/plink/1.9/). Individuals with
391 high SNP missingness were removed using --mind 0.9 flag alongside the --pca command
392 for all the aAIMs datasets. We also applied the model-based clustering methods
393 implemented in ADMIXTURE v1.3 [35]. All PCA and Admixture plots were generated
394 in R v3.2.3. Minor allele frequency (MAF) was calculated using PLINK (--maf
395 command) for ancient populations and for modern ones, MAF was calculated from the
396 1000 Genomes populations (ALL.2of4intersection.20100804.genotypes) [36]. Percentage
397 of rare and novel variants and other functional information were obtained through VEP
398 (McLaren et al. 2016).
399
400 Identifying aAIMs via five methods. aAIMs were considered markers that can infer the
401 ancestry of ancient DNA (aDNA) genomes in a similar accuracy to the complete SNP set
402 (CSS). We compared methods to detect candidate AIMs, three of which are novel:
403
404 **1.** Infocalc (Rosenberg et al. 2003), which determines the amount of information
405 multiallelic markers provide about an individual's ancestry by calculating the
406 informativeness (*I*) of each of each markers separately and ranks the SNPs by
407 their informativeness. Infocalc determines *I* based on the mathematical expression
408 described in Rosenberg et al. (2003). We compared the performances of four
409 choices of the top 5,000, 10,000, 15,000, and 20,000 most informative markers in

410 the Infocalc v1.1 output file (results not shown). The 15,000 dataset outperformed
411 all other datasets and was selected for further analyses.

412 **2.** *F_{ST}*. Wright's fixation indices (F_{ST}) [21] measures the degree of differentiation
413 among populations potentially arising due to genetic structure within populations.
414 Given a set of populations (Table S1), we employed PLINK [34] to estimate *F_{ST}*
415 separately for all the markers using –fst command alongside --within flag, that
416 defines population IDs of the genomes. Due to the high fragmentation of the data,
417 *F_{ST}* values could only be calculated for 46% of the dataset. We compared the
418 performances of four choices of the highest 5,000, 10,000, 15,000, and 20,000 F_{st}
419 values. The 15,000 dataset outperformed all other datasets and was selected for
420 further analyses.

421 **3.** Admixture$_1$. This method assumes that AIMs have high allelic frequencies in
422 certain subpopulations and drive the differentiation of admixture components.
423 Analyzing ADMIXTURE's output file (P file) for *K* of 10, we identified the
424 markers (rows) that had high allele frequency (>0.9) in only one admixture
425 component (columns). We identified 9,309 from the five columns with the highest
426 number of such markers.

427 **4.** Admixture$_2$. This method assumes that AIMs embody both high allelic
428 frequencies in certain subpopulations and high variance between these allelic
429 frequencies that differentiate of admixture components. Analyzing
430 ADMIXTURE's output file (P file) for *K* of 10, we identified 11,418 SNPs that
431 for each SNP (rows) had high variance (≥0.04) and high allele frequency range
432 (maxima - minima ≥ 0.65) between the admixture component (columns).

433 **5.** PC-based (PD) approach. This methods assumes that AIMs can replicate the
434 population structure of subpopulations represented by the variation in the first two
435 PCs. This is an interactive PC-based approach that identifies the smallest set of
436 markers necessary to capture the population structure of a group of individuals as
437 captured by the CSS. More specifically, for each population group (Table S1) in
438 which at least 100 SNPs were available, we calculated PCA and used PC1 and
439 PC2 to plot the individuals after all SNPs with high missingness (>0.05) were
440 excised. If the population group had insufficient SNPs we relaxed the missingness
441 threshold by additional 0.05, though 0.05 were sufficient for almost all groups.
442 We then scored the SNPs by their informativeness as in [37] and visually
443 compared the plot to that obtained from the CSS (Figure S7). If the plots were
444 dissimilar, we repeated the analysis using additional 100 top scored SNPs until
445 either the plots exhibited high similarity or a threshold of 2000 SNPs was reached.
446 We were unable to complete the analyses for 3 populations due to the small
447 number of individuals. The PD method is available on
448 https://github.com/eelhaik/PCA-derived-aAIMs. On average 861 SNPs were
449 found per population group. Overall, the dataset comprised of 13,027 SNPs.
450
451 To compare the prediction accuracy of the aAIMs subsets, two control datasets ($Rand_{10k}$
452 and $Rand_{15k}$) were generated by randomly sampling 10,000 and 15,000 SNPs from the
453 CSS, respectively. aAIMs identified by all methods are available as Additional file 2.
454

Classifying individuals to populations from genomic data. **Identifying ancient admixture components**. We selected a random hundred ancient genomes and removed six for insufficient data (>95% missingness). To those, we added 20 Han Chinese and 20 Yoruba modern genomes from the 1000 Genomes Project (Durbin et al. 2010). We then applied *supervised* ADMIXTURE with various *K*'s ranging from 8 to 13. While we were unable to find a single *K* where culturally related genomes exhibited homogeneous admixture patterns, the most robust population substructure was found for *K* of 10. Two more components were obtained by analyzing Spanish and German genomes that appeared indistinguishable along with five Yoruba genomes separately. We observed very little admixture with the Han and Yoruba. Overall, we identified 10 admixture components in ancient genomes, corresponding to allele frequencies of hypothetical populations. Similarly to Elhaik et al. [9], we simulated 15 samples for each hypothetical population, by generating 30 alleles whose average corresponds to the mean allele frequency of that population and assigning those genotypes to the simulated individuals. The putative ancestral ancient populations are available in Additional file 3. **Relabeling populations**. Initially, the labels from the corresponding papers were used to classify individuals to population. The consistency of these labels with data was evaluated by carrying out a *supervised* ADMIXTURE analysis on the genomic data combined with the 150 putative ancient ancestral individuals. Due to the high similarity of the admixture patterns between individuals of different groups living in similar periods or entire groups (e.g., Neolithic individuals from Hungary and those from Germany), we re-labeled some of the population to reduce the number of populations and create more genomically homogeneous populations,. For instance, Natufian and Neolithic samples from Jordan are

grouped into the label Levant Epipaleolithic Neolithic. Overall, we identified 23 populations, whose labels are all of the form "area_time period." In the case of the Caucasus labelling, all the samples from Iran (except Iran_HotuIIIb) were excavated in the Zagros Mountains, south of the Caucasus. Given their admixture similarity with Armenians and Georgians from the same periods and their proximity to the Caucasus, this area was labelled as Caucasus. Iran_HotuIIIb was found in a more eastern region, just below the southeastern edge of the Caspian Sea, and given its similarity to Georgians and other Iranians it was included in the group Caucasus Mesolithic Neolithic.

Genomically defining reference populations. For each population with $N_P > 4$, where N_P is the number of individuals in the population, individuals were grouped in clusters through an iterative process that uses a *k*-means clustering technique paired with multiple pairwise *F*-tests. Iterations ran over the number of k clusters [2, $N_P/2$]. At each iteration i, *k*-means was used to identify the k clusters, then the *F*-test was applied on each pair of clusters to test whether they are significantly ($P<0.05$) different. If the two clusters are different from all the pairs at iteration i, the process advances to $i+1$ until at least one pair violates the condition, in which case $k_{op}=i-1$ is the optimal number of clusters or reference populations within that population. **Assigning individuals to populations.** We developed an admixture-based classifier, which is not sensitive to exclusion of random groups of individuals nor inclusion of large numbers of individuals from admixed groups and was trained on a third of the data. Using *supervised* ADMIXTURE, we calculated the admixture proportions of the individuals in relation to the putative ancient ancestral populations. Population assignment was then made based on the minimal Euclidean distance between the admixture components of each genome and those of the reference

501 populations. The assignment accuracy was calculated based on the known classification
502 (Table S1).
503
504 Assessing admixture mapping. **Creating hybrid individuals**. We selected 15 individuals
505 from five populations that showed homogeneity in their admixture components (Figure 5)
506 and randomly sampled 120 pairs. Since selecting random alleles from each parent was
507 problematic due to the high missingness of the data, we randomly selected half the
508 genotypes of each parent to form 120 "offspring" or hybrid genomes. Each hybrid had
509 three SNP sets: the CSS, PD aAIMs, and a random SNP set of the size of PD aAIMs with
510 SNPs selected randomly for each hybrid. **Assessing admixture accuracy**. We defined
511 genetic distances (*d*) as the Euclidean distance between two set of admixture proportions.
512 We applied a *supervised* admixture to the three SNP sets of each hybrid and calculated
513 their distances *d* from each another.
514
515 Graphics. Maps were drawn using the 'rworldmap' package implemented in R v3.2.3.
516
517 Availability of data and materials. The dataset supporting the conclusions of this article
518 are included within the article and its additional files.
519
520 **Acknowledgment**
521 This study was partially supported by the MRC Confidence in Concept Scheme award
522 2014-University of Sheffield to E.E. (Ref: MC_PC_14115). We thank Grace Holland

who was partially supported by the UK EPSRC Doctoral Training Partnership Grant EP/N509735/1 as a Vacation Bursary Training Project.

References

1. Marchini J, Cardon LR, Phillips MS, Donnelly P: **The effects of human population structure on large genetic association studies.** *Nat Genet* 2004, **36:**512-517.
2. Scannell JW, Blanckley A, Boldon H, Warrington B: **Diagnosing the decline in pharmaceutical R&D efficiency.** *Nat Rev Drug Discov* 2012, **11:**191-200.
3. Yusuf S, Wittes J: **Interpreting geographic variations in results of randomized, controlled trials.** *N Engl J Med* 2016, **375:**2263-2271.
4. Rasmussen S, Allentoft ME, Nielsen K, Orlando L, Sikora M, Sjogren KG, Pedersen AG, Schubert M, Van Dam A, Kapel CM, et al: **Early divergent strains of Yersinia pestis in Eurasia 5,000 years ago.** *Cell* 2015, **163:**571-582.
5. D'Costa VM, King CE, Kalan L, Morar M, Sung WWL, Schwarz C, Froese D, Zazula G, Calmels F, Debruyne R, et al: **Antibiotic resistance is ancient.** *Nature* 2011, **477:**457.
6. Morozova I, Flegontov P, Mikheyev AS, Bruskin S, Asgharian H, Ponomarenko P, Klyuchnikov V, ArunKumar G, Prokhortchouk E, Gankin Y, et al: **Toward high-resolution population genomics using archaeological samples.** *DNA Res* 2016, **23:**295-310.
7. Marciniak S, Perry GH: **Harnessing ancient genomes to study the history of human adaptation.** *Nat Rev Genet* 2017, **advance online publication**.
8. Elhaik E, Greenspan E, Staats S, Krahn T, Tyler-Smith C, Xue Y, Tofanelli S, Francalacci P, Cucca F, Pagani L, et al: **The GenoChip: a new tool for genetic anthropology.** *Genome Biol Evol* 2013, **5:**1021-1031.
9. Elhaik E, Tatarinova T, Chebotarev D, Piras IS, Maria Calò C, De Montis A, Atzori M, Marini M, Tofanelli S, Francalacci P, et al: **Geographic population structure analysis of worldwide human populations infers their biogeographical origins.** *Nat Commun* 2014, **5:**1-12.
10. Phillips C, Parson W, Lundsberg B, Santos C, Freire-Aradas A, Torres M, Eduardoff M, Borsting C, Johansen P, Fondevila M, et al: **Building a forensic ancestry panel from the ground up: The EUROFORGEN Global AIM-SNP set.** *Forensic Sci Int Genet* 2014, **11:**13-25.
11. Mao X, Bigham AW, Mei R, Gutierrez G, Weiss KM, Brutsaert TD, Leon-Velarde F, Moore LG, Vargas E, McKeigue PM, et al: **A genomewide admixture mapping panel for Hispanic/Latino populations.** *Am J Hum Genet* 2007, **80:**1171-1178.

12. Elhaik E, Yusuf L, Anderson AIJ, Pirooznia M, Arnellos D, Vilshansky G, Ercal G, Lu Y, Webster T, Baird ML, Esposito U: **The Diversity of REcent and Ancient huMan (DREAM): a new microarray for genetic anthropology and genealogy, forensics, and personalized medicine.** *Genome Biol Evol* 2017, **9:**3225-3237.
13. Qin H, Zhu X: **Power comparison of admixture mapping and direct association analysis in genome-wide association studies.** *Genet Epidemiol* 2012, **36:**235-243.
14. Seldin MF, Price AL: **Application of ancestry informative markers to association studies in European Americans.** *PLoS Genet* 2008, **4:**e5.
15. Halder I, Shriver M, Thomas M, Fernandez JR, Frudakis T: **A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications.** *Hum Mutat* 2008, **29:**648-658.
16. Huckins LM, Boraska V, Franklin CS, Floyd JAB, Southam L, Gcan, Wtccc, Sullivan PF, Bulik CM, Collier DA, et al: **Using ancestry-informative markers to identify fine structure across 15 populations of European origin.** *Eur J Hum Genet* 2014, **22:**1190.
17. Kidd KK, Speed WC, Pakstis AJ, Furtado MR, Fang R, Madbouly A, Maiers M, Middha M, Friedlaender FR, Kidd JR: **Progress toward an efficient panel of SNPs for ancestry inference.** *Forensic Sci Int Genet* 2014, **10:**23-32.
18. Xu S, Huang W, Qian J, Jin L: **Analysis of genomic admixture in Uyghur and its implication in mapping strategy.** *Am J Hum Genet* 2008, **82:**883-894.
19. Rosenberg NA, Li LM, Ward R, Pritchard JK: **Informativeness of genetic markers for inference of ancestry.** *Am J Hum Genet* 2003, **73:**1402-1422.
20. Kidd JR, Friedlaender FR, Speed WC, Pakstis AJ, De La Vega FM, Kidd KK: **Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples.** *Investig Genet* 2011, **2:**1.
21. Wright S: *Evolution and the genetics of populations. A treatise in three volumes.* Chicago, IL: University of Chicago Press; 1968.
22. Shriner D: **Overview of admixture mapping.** *Curr Protoc Hum Genet* 2013, **Chapter 1:**Unit 1 23.
23. Tian C, Hinds DA, Shigeta R, Adler SG, Lee A, Pahl MV, Silva G, Belmont JW, Hanson RL, Knowler WC, et al: **A genomewide single-nucleotide-polymorphism panel for Mexican American admixture mapping.** *Am J Hum Genet* 2007, **80:**1014-1023.
24. Paschou P, Lewis J, Javed A, Drineas P: **Ancestry informative markers for fine-scale individual assignment to worldwide populations.** *J Med Genet* 2010, **47:**835-847.
25. Ding L, Wiener H, Abebe T, Altaye M, Go RC, Kercsmar C, Grabowski G, Martin LJ, Hershey GK, Chakorborty R, Baye TM: **Comparison of measures of marker informativeness for ancestry and admixture mapping.** *BMC Genomics* 2011, **12:**622.
26. Elhaik E: **Empirical distributions of F_{ST} from large-scale Human polymorphism data.** *PLoS One* 2012, **7:**e49837.

27. Elhaik E, Ryan D: **Pair Matcher (PaM): fast model-based optimisation of treatment/case-control matches using demographic and genetic data.** *bioRxiv* 2017.
28. Cassidy LM, Martiniano R, Murphy EM, Teasdale MD, Mallory J, Hartwell B, Bradley DG: **Neolithic and Bronze Age migration to Ireland and establishment of the insular Atlantic genome.** *Proc Natl Acad Sci U S A* 2016, **113:**368-373.
29. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25:**1754-1760.
30. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.
31. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: **The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome Res* 2010, **20:**1297-1303.
32. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al: **A framework for variation discovery and genotyping using next-generation DNA sequencing data.** *Nat Genet* 2011, **43:**491-498.
33. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al: **The variant call format and VCFtools.** *Bioinformatics* 2011, **27:**2156-2158.
34. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 2007, **81:**559-575.
35. Alexander DH, Novembre J, Lange K: **Fast model-based estimation of ancestry in unrelated individuals.** *Genome Res* 2009, **19:**1655-1664.
36. Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Gibbs RA, Hurles ME, McVean GA: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467:**1061-1073.
37. Paschou P, Ziv E, Burchard EG, Choudhry S, Rodriguez-Cintron W, Mahoney MW, Drineas P: **PCA-correlated SNPs for structure identification in worldwide human populations.** *PLoS Genet* 2007, **3:**1672-1686.
38. Marcus JH, Novembre J: **Visualizing the geography of genetic variants.** *Bioinformatics* 2017, **33:**594-595.

Figure Legends

Figure 1

Geographic distribution of the highly differentiated rs7896530 in ancient (A) and modern-day (B) populations. The geographic distributions of the T (black) and G (yellow) alleles in ancient and modern-day populations were obtained from our dataset (Table S1) and the Geography of Genetic Variants Browser [38], respectively.

657 Figure 2

658 **A scheme to identify and evaluate aAIMs.**

659

660

661 Figure 3

662 **Geographical locations of the ancient genomes**. The geographical coordinates of the
663 ancient genomes. The shapes plotted in the map designate the country of origin of the
664 genomes and their colors designate the era. The total number of ancient genomes from a
665 specific era is shown.

666

667

668 Figure 4

669 **Scatter plot of all ancient populations along the first two principal components**.

670 Symbols corresponding to individuals and their color and shape correspond to the

671 location map and the era table, respectively.

672

673

674 Figure 5

675 **Ancient population structure inferred by ADMIXTURE analysis.** Each individual is

676 represented by a vertical (100%) stacked column of genetic components proportions

677 shown in color for *K*=10.

678

679

Figure 6

Minor allele frequency distributions for aAIMs identified with various methods. MAF frequencies were calculated for ancient (A) and modern-day (B) populations. To avoid confusion, the distributions represent the frequency of the minor allele in each datasets, which was the same one in 91.5% of the genotypes.

687 Figure 7

688 **Violin plots comparing the Euclidean distances between the admixture proportions**

689 **of the ancient genomes obtained from the CSS and those obtained from the aAIM**

690 **sets**.

691

692

693 Figure 8

694 **Genome wide distribution of SNPs in the CSS (dots) and PD (red bars) datasets**.

695

696

Supporting Information Legends

Elhaik et al 2018 – Supp – Figures S1-S7 and Tables S1-S4

Additional file 1.zip – Genotype data of the aDNA samples

Additional file 2.zip – aAIMs candidates used in all analyses

Additional file 3.zip – Genotype file of the putative ancient ancestral populations

Tables

Table 1

Accuracy in classifying individuals to populations using the aAIM candidates. Mean and standard deviation for each SNP set are provided in the last row.

Population	*n*	CSS	PD	F_{ST}	Infocalc	$Admixture_1$	$Admixture_2$	$Rand_{10k}$	$Rand_{15k}$
Britain Iron Saxon	10	10 (100)	4 (40)	0 (0)	0 (0)	0 (0)	0 (0)	1 (10)	3 (30)
Caucasus Chalcolithic Bronze	22	21 (95)	8 (36)	0 (0)	12 (55)	6 (27)	4 (18)	13 (59)	9 (41)
Caucasus Mesolithic Neolithic	9	6 (67)	7 (78)	0 (0)	6 (67)	1 (11)	7 (78)	4 (44)	4 (44)
Central EU Early Neolithic	26	17 (65)	14 (54)	4 (15)	18 (69)	4 (15)	5 (19)	14 (54)	18 (69)
Central EU Late Neolithic Bronze	57	16 (28)	17 (30)	19 (33)	19 (33)	13 (23)	21 (37)	25 (44)	21 (37)
Central EU Mid Neolithic Chalcolithic	6	2 (33)	3 (50)	0 (0)	3 (50)	3 (50)	3 (50)	2 (33)	2 (33)
Central Northern EU Late Neolithic Bronze	20	18 (90)	9 (45)	0 (0)	6 (30)	0 (0)	5 (25)	4 (20)	6 (30)
Central Western EU Mesolithic	3	3 (100)	2 (67)	0 (0)	3 (100)	0 (0)	0 (0)	1 (33)	3 (100)
Italy Mid Neolithic Chalcolithic	4	4 (100)	3 (75)	0 (0)	1 (25)	1 (25)	0 (0)	1 (25)	1 (25)
Jordan Bronze	3	3 (100)	2 (67)	0 (0)	0 (0)	2 (67)	3 (100)	1 (33)	2 (67)
Levant Epipaleolithic Neolithic	19	7 (37)	6 (32)	0 (0)	9 (47)	8 (42)	7 (37)	4 (21)	7 (37)
Russia Chalcolithic	3	2 (67)	3 (100)	0 (0)	1 (33)	0 (0)	2 (67)	1 (33)	1 (33)
Russia Early Mid Bronze	19	19 (100)	15 (79)	0 (0)	10 (53)	0 (0)	18 (95)	10 (53)	14 (74)
Russia Late Chalcolithic	9	6 (67)	6 (67)	0 (0)	5 (56)	0 (0)	1 (11)	3 (33)	3 (33)
Russia Mesolithic	3	2 (67)	2 (67)	0 (0)	2 (67)	0 (0)	1 (33)	2 (67)	2 (67)
Russia Mid Late Bronze	22	15 (68)	16 (73)	0 (0)	7 (32)	0 (0)	0 (0)	4 (18)	6 (27)
Spain Early Neolithic	6	4 (67)	5 (83)	0 (0)	6 (100)	4 (67)	4 (67)	4 (67)	5 (83)
Spain Mid Neolithic Chalcolithic	18	7 (39)	6 (33)	0 (0)	7 (39)	5 (28)	3 (17)	5 (28)	5 (28)
Sweden Mesolithic	8	8 (100)	8 (100)	0 (0)	7 (88)	4 (50)	1 (13)	6 (75)	7 (88)
Sweden Mid Neolithic	4	4 (100)	1 (25)	1 (25)	2 (50)	1 (25)	0 (0)	4 (100)	2 (50)
Turkey Neolithic	24	23 (96)	18 (75)	0 (0)	12 (50)	3 (13)	6 (25)	8 (33)	11 (46)
		76±25	61±23	3±9	50±27	21±23	33±32	42±22	50±23

Table 2

Accuracy of inferring hybrid individuals using the PD aAIMs. The parental populations and the number of hybrids generated from them are shown. Each hybrid was represented by three datasets: CSS, PD aAIMs, and a random SNP set. The average genetic distances (*d*) between the admixture components of these datasets per population are shown.

Parental population A	Parental population B	# Hybrids	$\overline{d(\mathrm{CSS}, \mathrm{PD})}$	$\overline{d(\mathrm{CSS}, random\ set)}$	$\overline{d(\mathrm{PD}, random\ set)}$
Britain Iron Saxon	Britain Iron Saxon	6	0.026	0.212	0.208
Britain Iron Saxon	Russia Late Chalcolithic	9	0.009	0.610	0.601
Britain Iron Saxon	Sweden Mesolithic	9	0.051	0.344	0.337
Britain Iron Saxon	Turkey Neolithic	9	0.003	0.428	0.431
Britain Iron Saxon	Spain Early Neolithic	9	0.108	0.221	0.241
Russia Late Chalcolithic	Russia Late Chalcolithic	6	0.009	0.443	0.448
Russia Late Chalcolithic	Sweden Mesolithic	9	0.062	0.578	0.561
Russia Late Chalcolithic	Turkey Neolithic	9	0.063	0.661	0.633
Russia Late Chalcolithic	Spain Early Neolithic	9	0.101	0.520	0.491
Sweden Mesolithic	Sweden Mesolithic	6	0.000	0.384	0.384
Sweden Mesolithic	Turkey Neolithic	9	0.055	0.567	0.522
Spain Early Neolithic	Sweden Mesolithic	9	0.108	0.402	0.377
Turkey Neolithic	Turkey Neolithic	6	0.001	0.627	0.626
Spain Early Neolithic	Turkey Neolithic	9	0.092	0.483	0.493
Spain Early Neolithic	Spain Early Neolithic	6	0.041	0.197	0.172

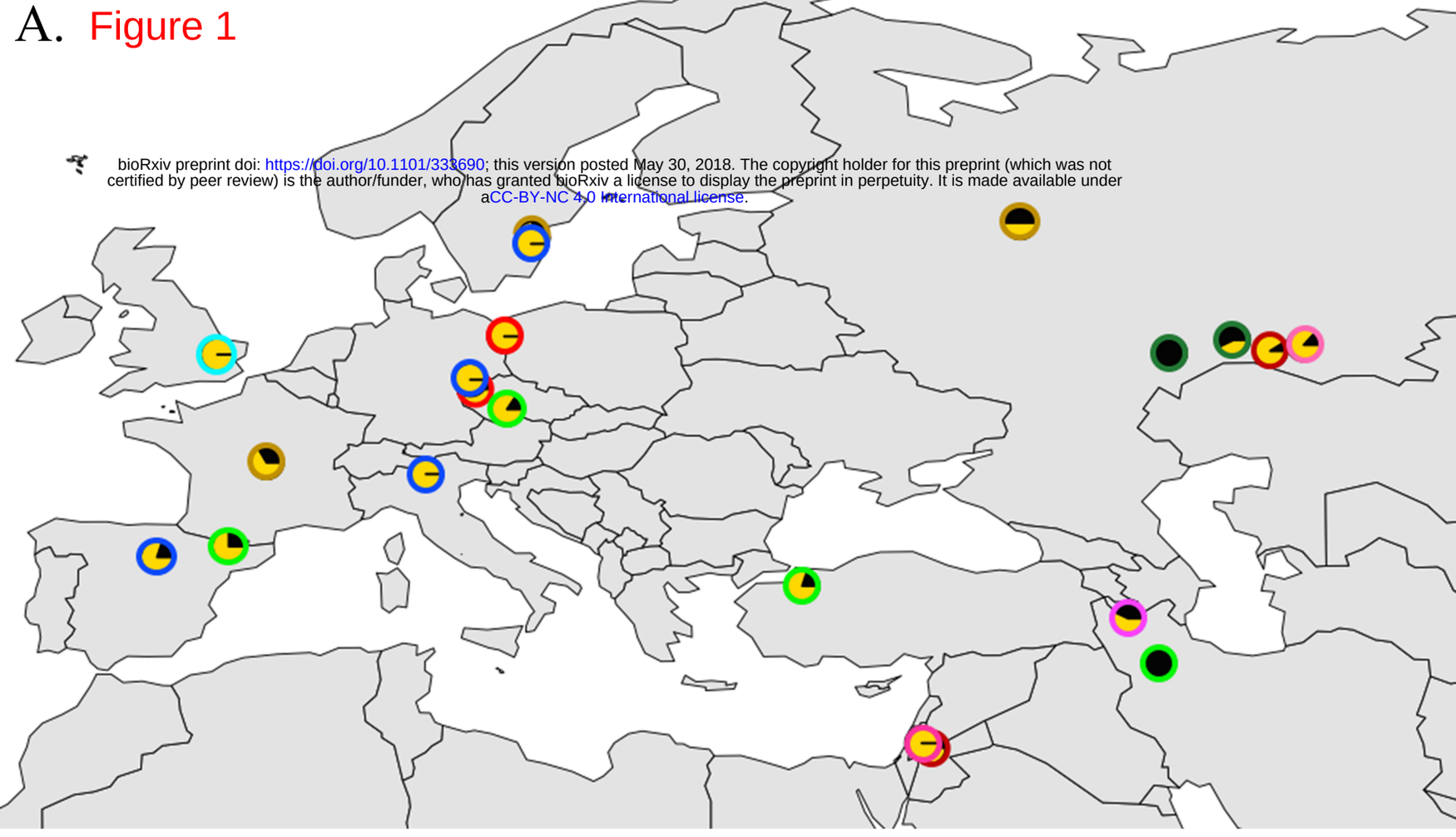
724 **Competing interests**

725 EE is a consultant to DNA Diagnostic Centre. The funders had no role in study design,

726 data collection and analysis, decision to publish, or preparation of the manuscript.

727

A. Figure 1



Mesolithic
Early-Neolithic/Neolithic
Mid-Neolithic/Mid-Neolithic Chalcolithic
Late-Neolithic Bronze
Chalcolithic/Late Chalcolithic
Early Bronze/Chalcolithic Bronze
Early-Mid Bronze
Mid-Late Bronze
Epipaleolithic
Anglo-Saxon

B.

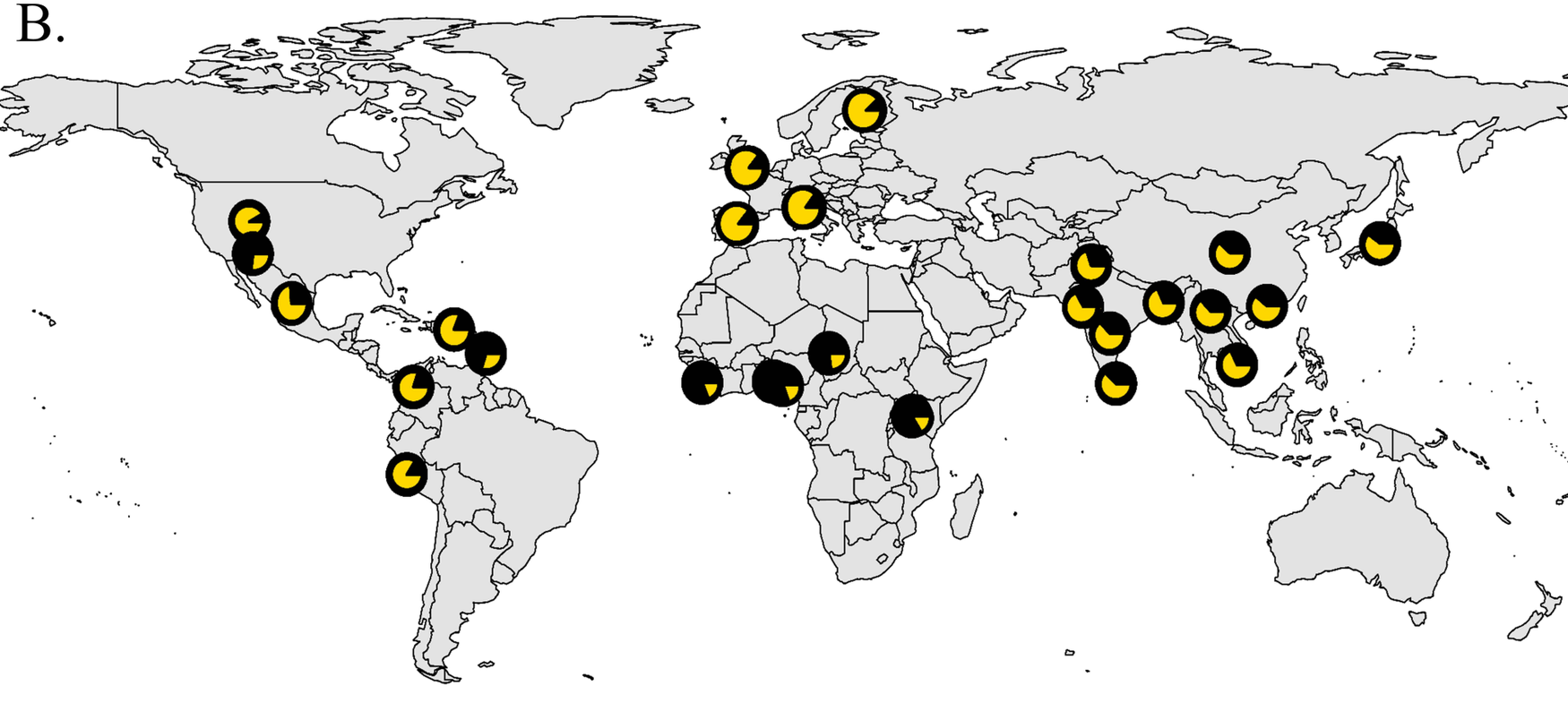


Figure 2

Ancient DNA dataset

Apply aAIMs methods to the aDNA dataset

Ancient genomic data

Initial aAIM candidate sets

Infocalc | Admixture$_1$ | PD | Rand$_{10k}$

F_{ST} | Admixture$_2$ | Rand$_{15}$

Filter samples/SNPs for missingness

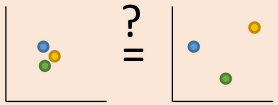
Final aAIM candidate sets

Calculate overlap between the aAIM sets
Calculate aAIMs overlap
Calculate aAIMs MAF distributions
Functional annotation

Identifying aAIM candidates

Calculate PCA for each aAIM set

1) Qualitative assessment of PCA plots

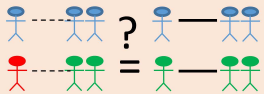


Calculate admixture for each aAIM set

2) Comparison of admixture plots

Classify individuals to populations

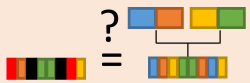
3) Comparison of individual classification



Further analyses with the best aAIMs

Hybrids dataset

4) Test hybrid inference



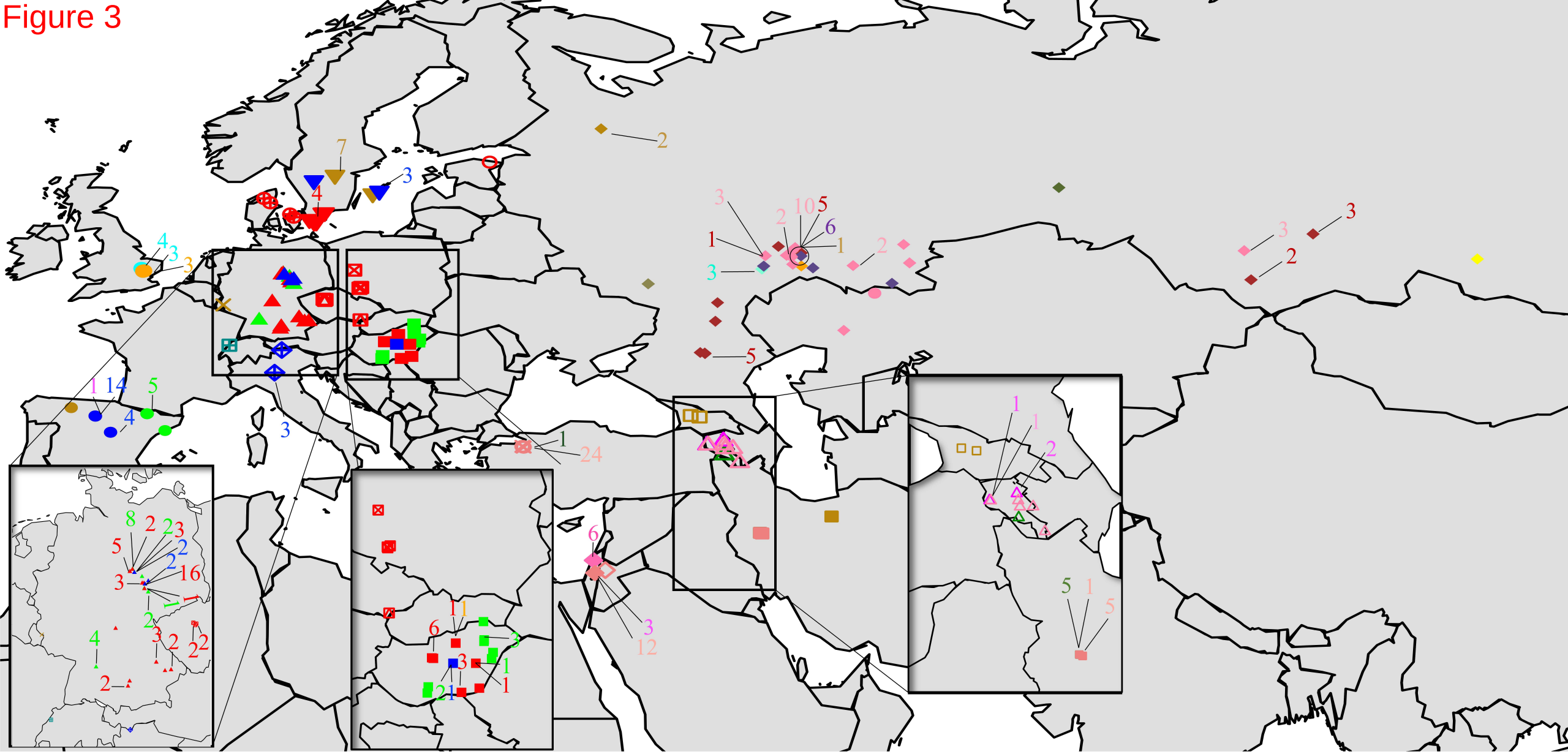
Calculate genome wide distribution of aAIMs

Criteria to evaluate the candidate aAIMs:

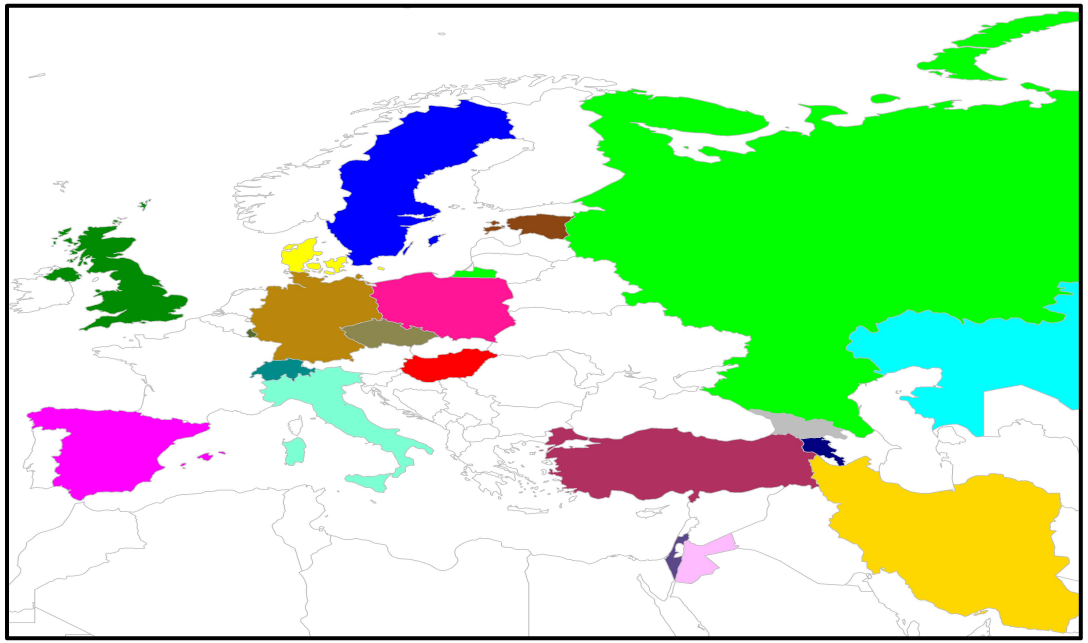
Can aAIMs capture the population structure?

Can aAIMs be used for admixture mapping?

Figure 3



Upper Paleolithic (1)
Mesolithic (16)
Early-Neolithic (32)
Mid-Neolithic (32)
Late-Neolithic Bronze (77)
Neolithic (43)
Late Copper (9)
Chalcolithic (11)
Early Bronze (7)
Early-Mid Bronze (19)
Mid-Late Bronze (31)
Iron/Late Iron (5)
Russia Eneoliothic (3)
Russia Kostenki (1)
Russia MA1 (1)
Russia Ust Ishim (1)
Israel Epipaleolithic (6)
Britain Anglo-Saxon (7)



Upper Paleolithic (30,000 - 12,000 BP)	Epipaleolithic (12,000 BP - 8,300 BC)	Mesolithic (10,000 - 5,000 BC)
Early Neolithic (8,300 - 5,500 BC)	Neolithic (8,300 - 5,500 BC)	Mid Neolithic Chalcolithic (5,500 - 4,500 BC
Chalcolithic (4,500 - 4,000 BC)	Late Copper (4,000 -3,300 BC)	Late Neolithic Bronze (3,300 - 3,000 BC)
Early Bronze (3,000 - 2,700 BC)	Early Mid Bronze (2,000 - 1,750 BC)	Mid Late Bronze (1,400 - 1,200 BC)
Iron (1,200 - 586 BC)	Anglo-Saxons (450 – 1,066 AD)	Other–Chalc. Bronze & Iron

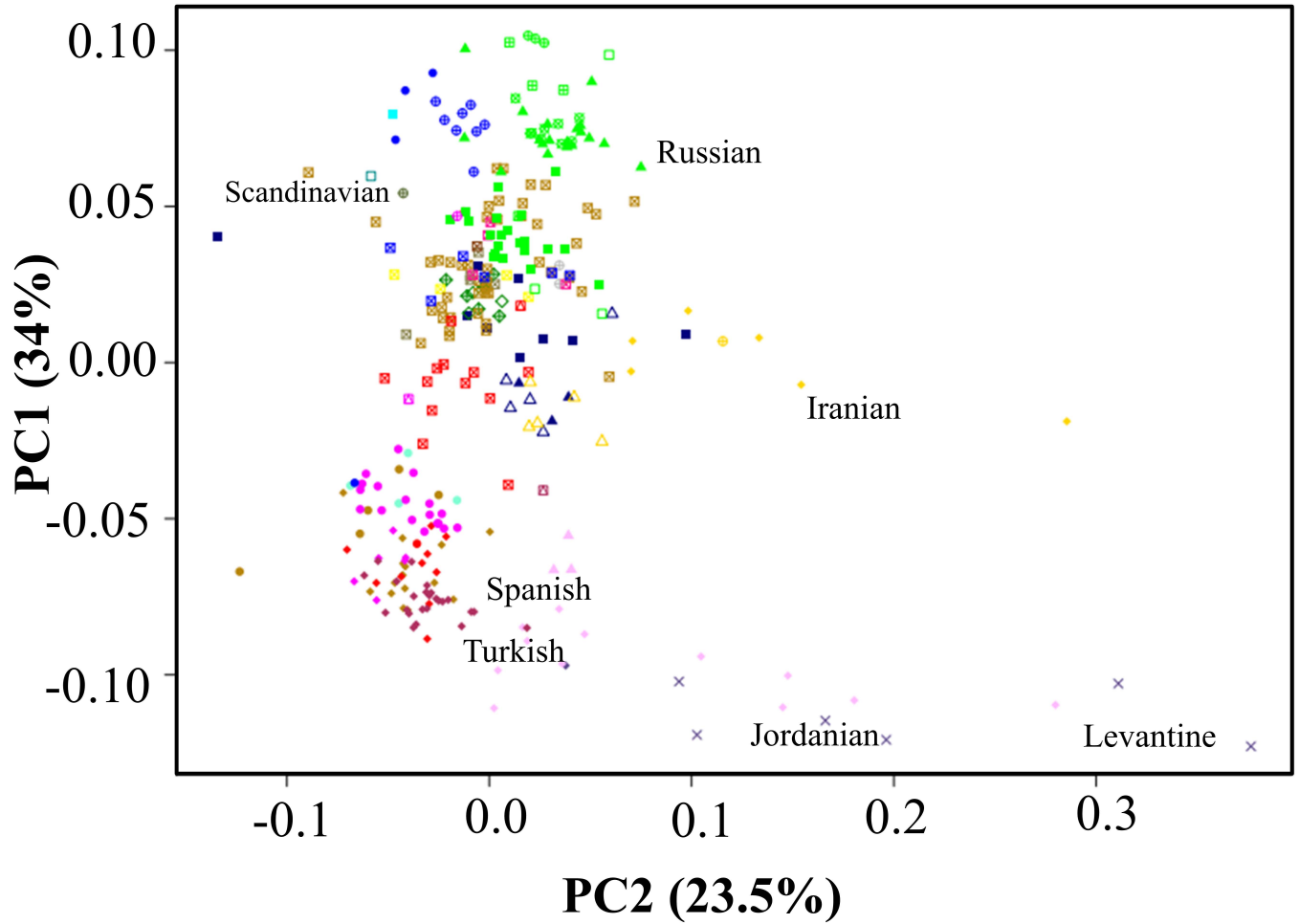
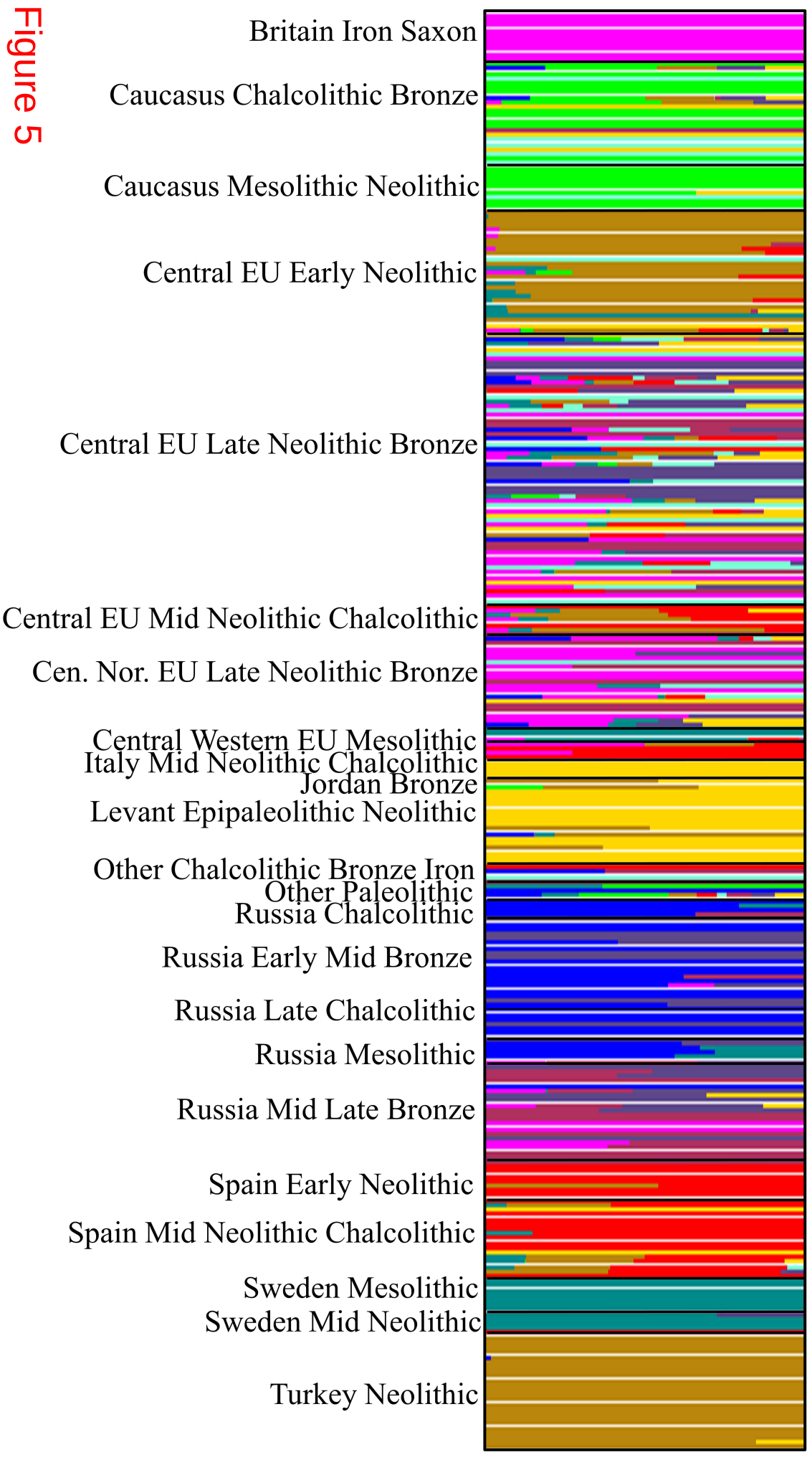


Figure 4

Figure 5


Britain Iron Saxon
Caucasus Chalcolithic Bronze
Caucasus Mesolithic Neolithic
Central EU Early Neolithic
Central EU Late Neolithic Bronze
Central EU Mid Neolithic Chalcolithic
Cen. Nor. EU Late Neolithic Bronze
Central Western EU Mesolithic
Italy Mid Neolithic Chalcolithic
Jordan Bronze
Levant Epipaleolithic Neolithic
Other Chalcolithic Bronze Iron
Other Paleolithic
Russia Chalcolithic
Russia Early Mid Bronze
Russia Late Chalcolithic
Russia Mesolithic
Russia Mid Late Bronze
Spain Early Neolithic
Spain Mid Neolithic Chalcolithic
Sweden Mesolithic
Sweden Mid Neolithic
Turkey Neolithic

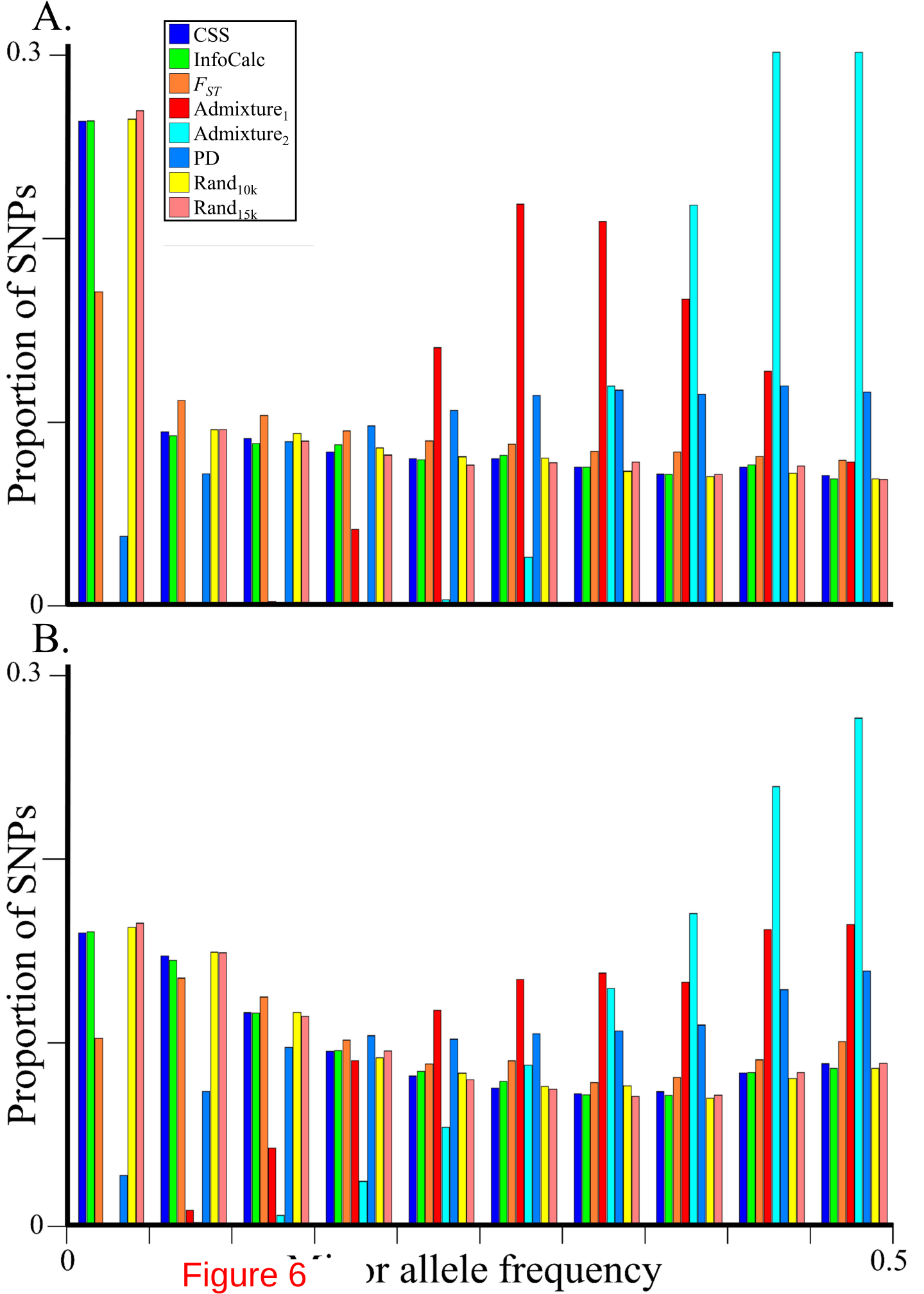
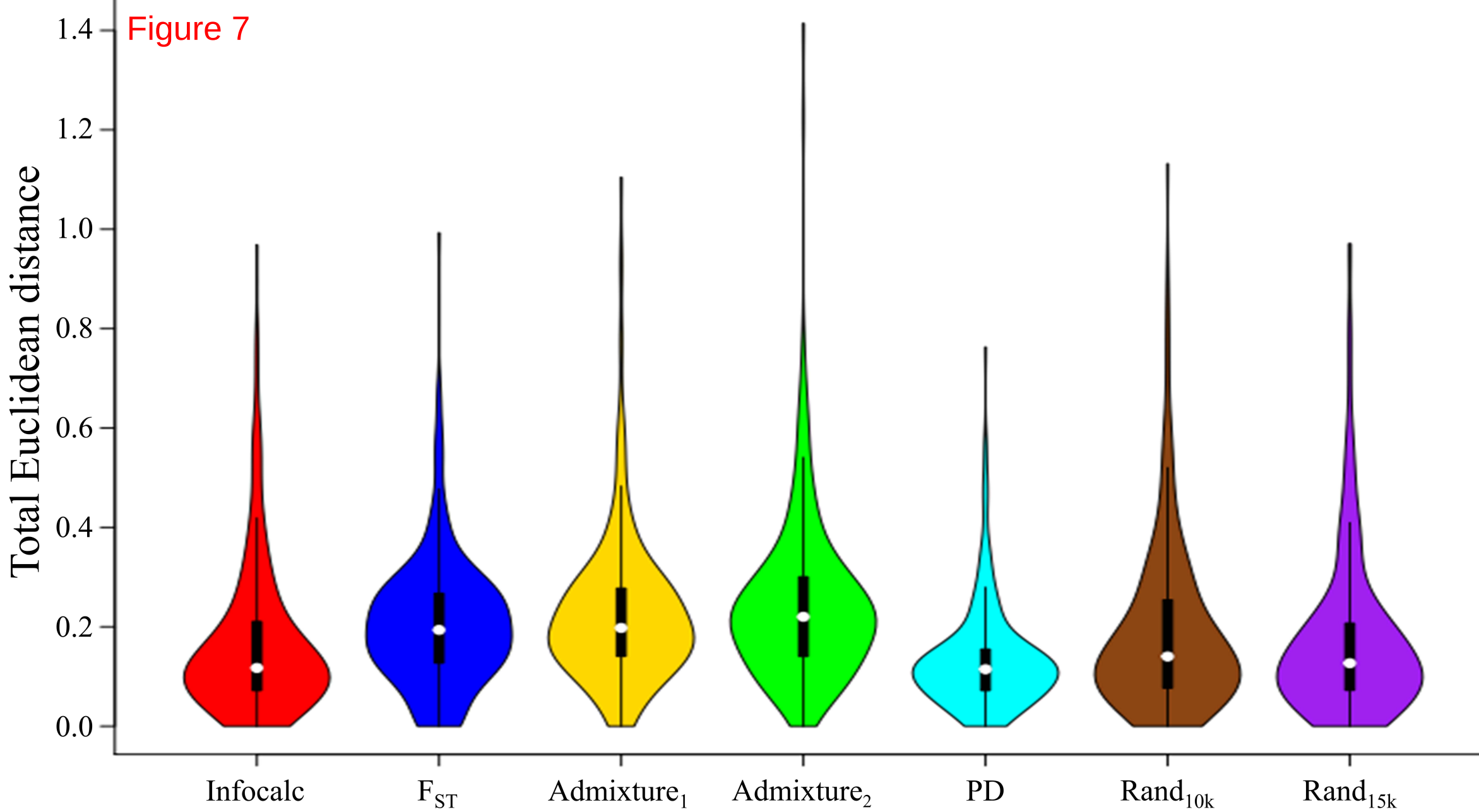
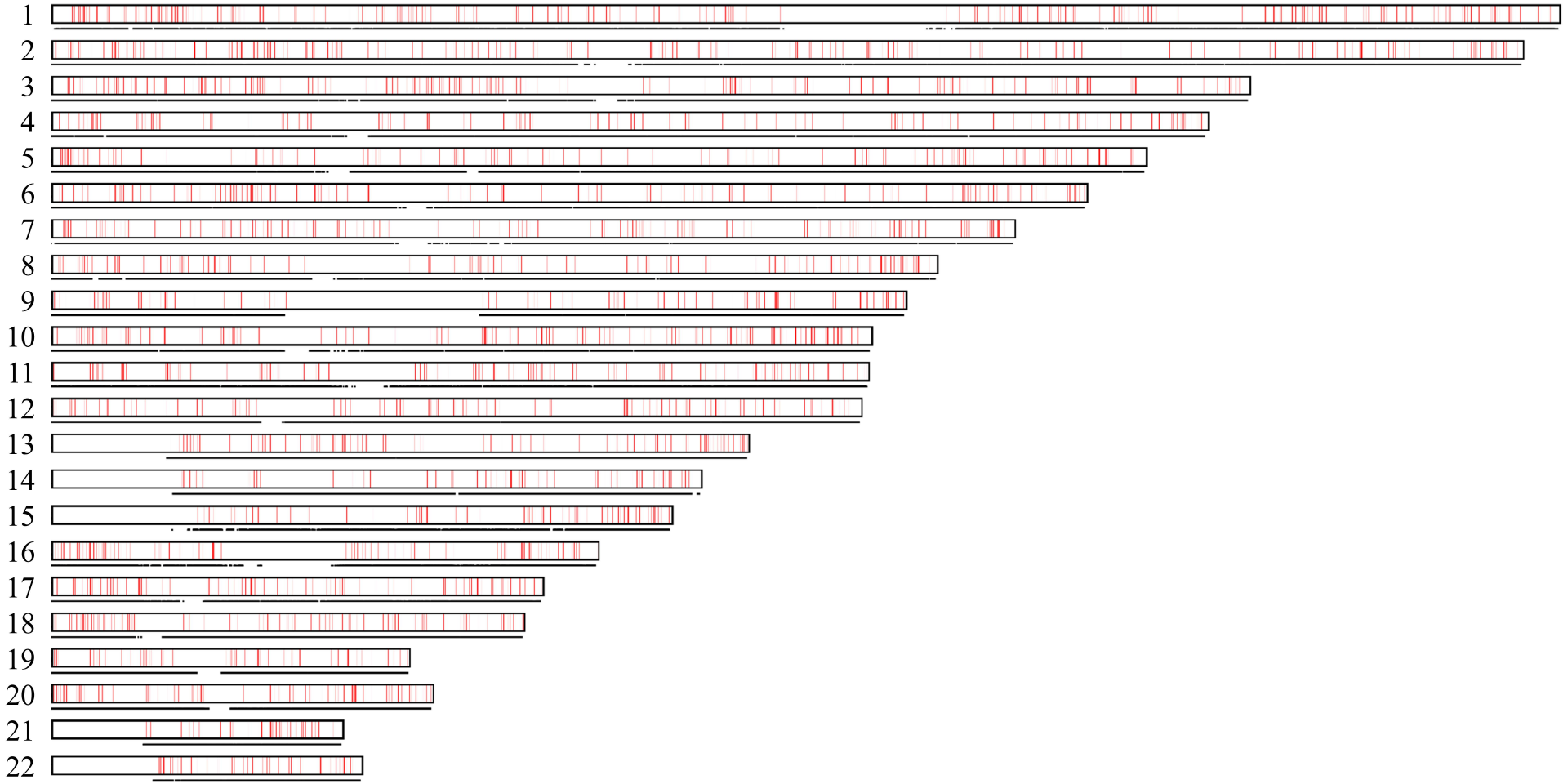

A.
B.
CSS
InfoCalc
F_{ST}
Admixture$_1$
Admixture$_2$
PD
Rand$_{10k}$
Rand$_{15k}$
0.3
0
Proportion of SNPs
0
0.5
r allele frequency

Figure 6

Figure 7


Total Euclidean distance
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Infocalc
F_{ST}
$Admixture_1$
$Admixture_2$
PD
$Rand_{10k}$
$Rand_{15k}$


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22