View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by White Rose Research Online

This is a repository copy of *Ancient ancestry informative markers for identifying fine-scale ancient population structure in Eurasians*.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/144736/

Version: Published Version

Article:
Esposito, U., Das, R., Syed, S. et al. (2 more authors) (2018) Ancient ancestry informative markers for identifying fine-scale ancient population structure in Eurasians. Genes, 9 (12). 625. ISSN 2073-4425

https://doi.org/10.3390/genes9120625

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

White Rose university consortium
Universities of Leeds, Sheffield & York

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

genes

MDPI

Article

Ancient Ancestry Informative Markers for Identifying Fine-Scale Ancient Population Structure in Eurasians

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

[1] Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK; u.esposito@sheffield.ac.uk (U.E.); syakir.syed@gmail.com (S.S.)
[2] Manipal University, Manipal Centre for Natural Sciences (MCNS), Manipal, Karnataka 576104, India; ranajit.das@manipal.edu
[3] Bioinformatics and Computational Biology, National Heart Lung and Blood Institute, National Institutes of Health, Bethesda, MD 20892, USA; mehdi.pirooznia@nih.gov
* Correspondence: e.elhaik@sheffield.ac.uk

Received: 6 November 2018; Accepted: 10 December 2018; Published: 12 December 2018

Abstract: The rapid accumulation of ancient human genomes from various areas and time periods potentially enables the expansion of studies of biodiversity, biogeography, forensics, population history, and epidemiology into past populations. However, most ancient DNA (aDNA) data were generated through microarrays designed for modern-day populations, which are known to misrepresent the population structure. Past studies addressed these problems by using ancestry informative markers (AIMs). It is, however, unclear whether AIMs derived from contemporary human genomes can capture ancient population structures, and whether AIM-finding methods are applicable to aDNA. Further the high missingness rates in ancient—and oftentimes haploid—DNA can also distort the population structure. Here, we define ancient AIMs (aAIMs) and develop a framework to evaluate established and novel AIM-finding methods in identifying the most informative markers. We show that aAIMs identified by a novel principal component analysis (PCA)-based method outperform all of the competing methods in classifying ancient individuals into populations and identifying admixed individuals. In some cases, predictions made using the aAIMs were more accurate than those made with a complete marker set. We discuss the features of the ancient Eurasian population structure and strategies to identify aAIMs. This work informs the design of single nucleotide polymorphism (SNP) microarrays and the interpretation of aDNA results, which enables a population-wide testing of primordialist theories.

Keywords: ancient DNA; ancient ancestry informative markers; population structure; principal component analysis; admixture mapping, primordialism

1. Introduction

1.1. Towards High-Resolution Population Models Using Ancient Samples

Over the past decade, genomic techniques have been reshaping our fundamental understanding of human prehistory and origins [1]. Ancient DNA (aDNA) human genomes have assisted in investigations of population structure, human migration, human adaptation, agricultural lifestyle, and disease co-evolution [2]. Ancient genome studies have already accelerated progress in the search for genetic variations underlying the inheritance of adaptations and forensics traits. Recently, Cassidy et al. [3] tested the allelic association of cystic fibrosis and hemochromatosis in ancient Irish samples, expanding genetic epidemiology onto ancient genomes. Such studies can potentially identify new risk factors for rare diseases.

1.2. Next Generation Sequencing Technologies to Study Ancient DNA

Whole genome sequencing and single nucleotide polymorphism (SNP) microarrays are the two leading approaches to aDNA sequencing. Although the former is preferable as it provides more data, by late 2017, only a quarter of the 1100 sequenced ancient genomes were whole genomes. The vast majority of these genomes (762) were captured by SNP microarrays [2], mainly the Human Origins [4] and Illumina 610-Quad arrays [5,6]—neither of which were designed for ancient humans—making it particularly challenging to identify and control for ancient population structure. Future microarrays, dedicated to aDNA, will thereby need a reliable set of polymorphic markers that can be used to differentiate ancient populations whose population structure was shown to vary over time [7].

Single nucleotide polymorphism genotyping microarrays were originally developed to detect phenotype–genotype associations in association mapping, admixture mapping, identity by descent mapping, and similar studies. It was not until later that SNP microarrays were employed in population genetic studies aimed at inferring population structure through principal component analyses (PCAs), ADMIXTURE-like programs, and other tools aimed at predicting group membership. It soon became clear that the allele frequency spectrum obtained through microarrays is more skewed for some populations than for other ones due to the choice of SNP panels [8]. The Human Origins and various Illumina microarrays (including the Illumina Human 660W-Quad, which is very similar to the Illumina 610-Quad array) were shown to distort the population structure for modern day populations compared to larger genomic databases and underreport the biodiversity compared to microarrays customized for population genetics [9,10], which results in an ascertainment bias.

1.3. The Problems of Ascertainment Bias and Population Stratification in Ancient DNA

Any inference of identity in archeological studies is fraught with difficulties. Carbon dating requires extracting organic material from fossil bones and authenticating it as composed of degraded proteins; this process is highly susceptible to contamination, which yields erroneous estimates [11]. The identification of 'cultures' from archaeological remains and their association with past population groups is also inadequate [12]. Population genetic studies suffer from similar problems due to ascertainment bias, which can distort measures of human diversity, bias population genetic inferences, and alter the conclusions in unexpected ways [13]. Ascertainment bias is a major concern in genetic, biomedical, and evolutionary studies, particularly in the absence of an established population structure model for either modern-day or ancient populations.

The difficulties related to establishing an acceptable population model are partially due to our incomplete knowledge of human population biodiversity in the past and present. Often, modern-day populations are assumed to be the parental populations of the modern-day population of interest, which results in population stratification. This problem arises due to differences in the allele frequencies of unknown case/control subpopulations due to separate demographic histories (not biological processes). A misunderstanding of the population structure necessitates mismatched cases and controls, which introduces genetic heterogeneity into the analysis that can lead to spurious associations and obscure the true association [14]. Thereby, the phenotypes of interest (e.g., risk loci or drug response) may differ between these subpopulations and bias the association analyses by generating false positives [15]. These problems have been well-known for a long time, and statistical remedies have been proposed [16]; however, they were all tailored for modern-day data, and do not address the conceptual problems. It is now clear that population models should consider aDNA data and the unique challenges they pose, such as, haploidy that reduces the biodiversity of the samples and high missingness, which precludes comparing individuals on the same marker set [1].

1.4. The Use of Ancestry Informative Markers in Genetics

Past studies have resolved, to a large extent, the problems faced in DNA analyses with ancestry informative markers (AIMs). Ancestry informative markers are SNPs that exhibit large variation

in minor allele frequencies (MAF) among populations. Over the past two decades, geneticists have scoured genomes for these patterns, and to date produced numerous AIM sets to determine an individual's ancestry, detect stratification in biomedical studies, infer geographic structure, find risk loci in a candidate region, and localize biogeographical origins (e.g., [9,10,17–19]). Ancestry informative marker panels can delineate population structure in a cost-effective manner by detecting variation in individual ancestry that can confound methods such as Mendelian randomization trials, association analyses, and forensic investigations by increasing false positive results or reducing power [20].

Although initially preferred due to the high cost of sequencing (which has decreased with time) AIMs are still frequently used in forensics, carrier screening, and biogeography in both microarrays (e.g., [9,21]) and whole genomic data [22]. Admixture mapping is another powerful method to map phenotypic variation or diseases that show differential risk by ancestry. The mapping takes advantage of higher densities of genetic variants and extensions to admixed populations, which exhibit strong differences in prevalence across populations [23]. Therefore, it is necessary to have a large number of AIMs throughout the genome to allow for the inference of local chromosomal ancestry blocks.

Despite their high prevalence, it has never been clear which AIMs should be used. All AIM panels have limitations [24] and it is unknown whether the established AIMs would be informative for aDNA studies. The characteristics of ideal AIMs remain contentious, with some authors preferring common SNPs (minor allele frequency >1%) [25], SNPs with high fixation index (F_{ST}) [26], SNPs with high pairwise MAF between populations [24], or SNPs that satisfy several criteria. Consequently, AIMs do not overlap across studies. Of the 21 AIM datasets reviewed by Pakstis et al. [27], the union of SNPs consisted of 1397 AIMs appearedof which only 46 occurred in three to six panels. Finally, studies typically show that AIMs can separate populations or broadly classify individuals into subcontinental populations rather than showing that AIMs can capture the population structure of the complete SNP set or allow fine-population mapping. Given the uncertainties surrounding AIMs, their potential incompatibility for capturing ancient structure and admixtures, and the challenges imposed by aDNA data, it is unclear whether, if at all, AIM-finding methods or AIMs can be utilized to study ancient population structures.

1.5. Ancient Ancestry Informative Markers to Define Ancient Population Structure

aDNA, aDNA allows for the construction of AIM panels from the parental populations of modern-day people and can refine population structure estimates. To overcome some of the aforementioned problems with aDNA data, 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 in order to improve the inference of ancient population structure. In the absence of aAIM-finding tools, we selected several established AIM-finding tools: two existing AIM-finding algorithms (Infocalc [28] and F_{ST}-based algorithm [29] that employed Wright's F_{ST} [30]) and developed three novel admixture and PCA-based algorithms. Both methods have characteristics that have been reported to be beneficial in measuring genetic distances in population genetic studies (e.g., [16,31,32]), and were expected to be useful in identifying AIMs. Since AIM-finding tools were never tested on aDNA, it is necessary to first compare their ability in finding informative markers, which can differentiate ancient populations. For that, we interrogated a comprehensive dataset of 302 ancient genomes from Europe, the Middle East, and North Eurasia, spanning time periods from 14,000 through to 1500 years ago, and that were sequenced using both microarray and whole genome technology. These genomes were grouped into 21 populations based on geographical and temporal information (Table S1). Each population was then further divided into subpopulations based on the genetic similarity between the genomes in terms of their admixture profile. To test how well the aAIM candidates, as identified by various tools, capture the population structure and identify admixed individuals, we first derived summary statistics using these aAIM candidates. Then, we compared the performances of the best aAIM set with the complete SNP set in classifying individuals to populations and identifying two-way admixed individuals (Figure 2). Our current study offers a methodological

framework to evaluate AIMs, contrasts different AIM-finding strategies, reports the first set of aAIMs, and demonstrates that in some cases, they provide more reliable predictions than the complete SNP set.

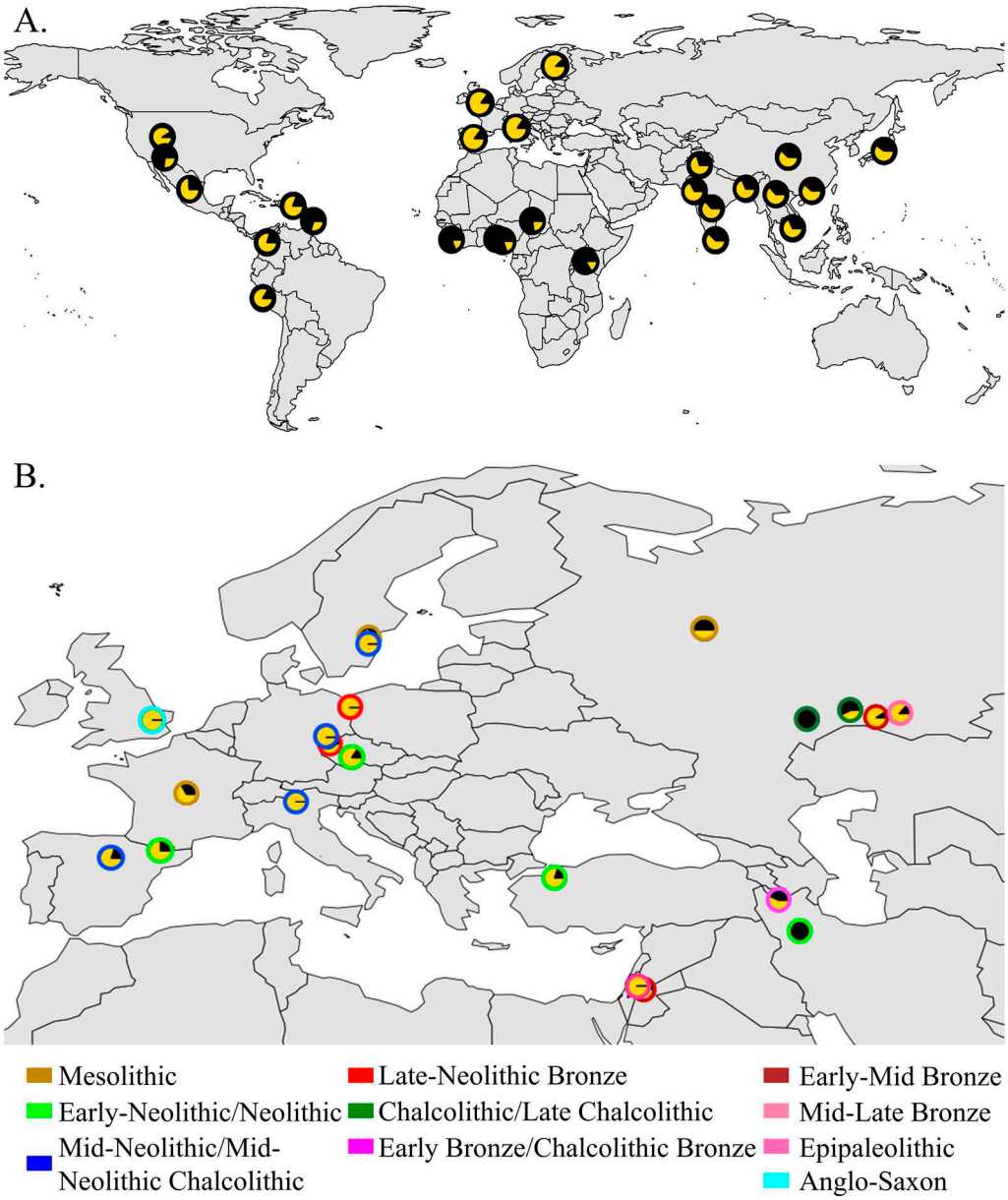

Figure 1. Geographic distribution of the highly differentiated rs7896530 in modern-day (**A**) and ancient (**B**) populations. The geographic distributions of the T (black) and G (yellow) alleles were obtained from the Geography of Genetic Variants Browser [33] and Table S1, respectively.

2. Materials and Methods

2.1. Ancient Data Collection

Genomic data were obtained from 11 publications depicting 302 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) [34], 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) [35]. Picard (version 2.1.1) (http://picard.sourceforge.net/) and MarkDuplicates were used to remove reads with identical outer mapping coordinates

(which are likely PCA artifacts). The Genome Analysis Toolkit RealignerTargetCreator module (GATK version 3.6) [36,37] was used to generate SNP and small insertion/deletion (InDel) calls for the data within the targeted regions only. GATK InDelRealigner/BaseRecalibrator was then used for local read realignment around known InDels and for the base quality score recalibration of predicted variant sites based on dbSNP 138 and 1000 Genomes known sites, resulting in corrections for base reported quality. The recalibration was followed by SNP/InDel calling with the GATK HaplotypeCaller. Variants were filtered for a minimum confidence score of 30 and a minimum mapping quality of 40. At the genotype level, all of the genotypes that had a genotype depth of less than four (GD < 4) or a genotype quality score less than 10 (GQ < 10) were removed from the dataset by setting them as missing in the VCF. GATK DepthofCoverage was used to re-examine coverage following the realignment. VCFtools (version 0.1.14) [38] were used to convert the VCF file to PLINK format [39]. The final dataset comprised of 150,278 autosomal SNPs from 302 aDNA genomes (Table S1; Additional file 1 in Supplementary Materials). Eight aDNA genomes (I0247, I1584, ATP9, IR1, Kostenki14, MA1, and Ust Ishim) without any country/region designation were omitted in the closest population determination calculations. The genomes were classified into 21 populations, based on their sampling country/region and era.

2.2. Data Analyses

2.2.1. The Genetic Structure Canvas of Ancient Eurasian Genomes

The population structure of the ancient genomes was described using PCA implemented in PLINK v1.9 [39]. Ancient genomes and SNPs with over 90% missingness were removed. We also applied the model-based clustering methods implemented in ADMIXTURE v1.3 [40]. Minor allele frequency was calculated using PLINK (–maf command). The MAF for modern-day populations was calculated from the 1000 Genomes populations (ALL.2of4intersection.20100804.genotypes) [41]. The percentage of rare and novel variants and other functional information were obtained through the Variant Effect Predictor (VEP).

2.2.2. Identifying aAIMs Using Multiple Methods

We applied two established and three novel methods to detect aAIM candidates as follows:

1. **Infocalc** v1.1 [28], determines the amount of information that multiallelic markers provide of an individual's ancestry by calculating the informativeness (I) of each marker separately and ranking the SNPs by their informativeness. Infocalc determines I based on the mathematical expression described in Rosenberg et al. (2003). We compared the performances (Figure 2) of the top 5000, 10,000, 15,000, and 20,000 most informative markers (results not shown). The 15,000 dataset outperformed all of the other datasets, and was selected for further analyses.
2. $\boldsymbol{F_{ST}}$. Wright's fixation indices (F_{ST}) [30] measures the degree of differentiation among populations, which was potentially arising due to the genetic structure within populations. Given a set of populations (Table S1), we employed PLINK v1.9 [39] to estimate F_{ST} separately for all the markers using the –fst command alongside –within flag. Due to the high fragmentation of the data, F_{ST} values could only be calculated for 46% of the dataset. We compared the performances (Figure 2) of 5000, 10,000, 15,000, and 20,000 SNPs with the highest F_{ST} values (results not shown). The 15,000 dataset outperformed all of the other datasets, and was selected for further analyses.
3. **Admixture$_1$**. This method assumes that aAIMs have high allelic frequencies in certain subpopulations that drive the differentiation of admixture components. Analyzing ADMIXTURE's output file (P file) for K = 10, we identified the markers (rows) that had high allele frequency (>0.9) in only one admixture component (columns). Comparing the number of high-MAF SNPs in all of the columns, we selected 9309 from the five columns with the highest number of such SNPs.
4. **Admixture$_2$**. This method assumes that aAIMs embody both high allelic frequencies in certain subpopulations, and that the high variance between these allelic frequencies differentiates the admixture components. Analyzing ADMIXTURE's output file for K of 10, we identified

11,418 SNPs showing high variance ($\geq$0.04) and a high allele frequency range (maxima–minima $\geq$ 0.65) between the admixture components.

5. **Principal Component-derived (PD)**. This method assumes that AIMs can replicate the population structure of subpopulations represented by the variation in the first two PCs. This is an interactive PC-based approach that identifies the smallest set of markers necessary to capture the population structure of a group of individuals, as captured by the complete SNP set (CSS). More specifically, for each population group (Table S1) in which at least 100 SNPs were available, we carried out PCA after all of the SNPs with high missingness (>0.05) were excised. If the population group had insufficient SNPs, we relaxed the missingness threshold by an additional 0.05, although 0.05 were sufficient for almost all of the groups. We then scored the SNPs by their informativeness, as in [42], and used the top 100 most informative SNPs to plot the individuals on a scatter plot using PC1 and PC2 as axes. We visually compared the plot to that obtained from the CSS (Figure S11). If the plots were dissimilar, we repeated the analysis using an additional 100 top-scored SNPs until either the plots exhibited high similarity or a threshold of 2000 SNPs was reached. In this manner, we identified the minimum number of the most informative SNPs that were needed to replicate the PCA results of the CSS. We were unable to complete the analyses for three populations due to the small number of individuals. The PD method is available on https://github.com/eelhaik/PCA-derived-aAIMs. On average, 861 SNPs were collated per population group. Overall, the dataset comprised 13,027 SNPs.

To compare the prediction accuracy of the aAIMs subsets, two control datasets ($Rand_{10k}$ and $Rand_{15k}$) were generated by randomly sampling 10,000 and 15,000 SNPs from the CSS, respectively. The aAIMs identified by all of these methods are available as Additional File 2 in Supplementary Materials.

2.2.3. Classifying Individuals into Populations from Genomic Data

Following the reported success of the admixture-based method, which employs AIMs to describe and classify individuals to populations [17,43–45], we sought to develop an analogous method that employs aAIMs.

Identifying ancient admixture components: To avoid over-fitting, and since some of the methods employ ADMIXTURE, we sought to identify admixture components in a small cohort of diverse individuals. For that, we selected 100 random ancient genomes and removed six because of insufficient data (>95% missingness). To those, we added 20 Han Chinese and 20 Yoruba modern genomes from the 1000 Genomes Project [41]. We then applied an *unsupervised* ADMIXTURE with the parameter *K* ranging from 8 to 13. Although we were unable to find a single *K* when culturally related genomes exhibited homogeneous admixture patterns, the most robust population substructure was found for the *K* value of 10. Two more admixture components were obtained by separately analyzing the Spanish and German genomes, which appeared indistinguishable in the original analysis, along with five Yoruba genomes. We observed very little admixture of the ancient individuals with the Han and Yoruba. Overall, we identified 10 admixture components in ancient genomes, corresponding to the allele frequencies of 10 hypothetical populations. Similar to Elhaik et al. [17], we simulated 15 samples to represent each hypothetical population by generating 30 alleles whose MAF values corresponded to the MAF of each population, and assigning those genotypes to the simulated individuals. The putative ancestral ancient populations are available in Additional File 3 in the Supplementary Materials.

Relabeling populations: Initially, the labels from the corresponding literature 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 relabeled 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 21 populations whose labels are of the form "area_time period". In the case of the Caucasus label, all of 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 labeled 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 assigned to that population, individuals were grouped into subclusters through an iterative process that uses the *k*-means clustering technique paired with multiple pairwise *F*-tests. Iterations ran over the number of *k* subclusters [$k = 2, ..., N_P/2$]. At each iteration *i*, *k*-means was used to identify the *k* subclusters; then, the *F*-test was applied on each pair of subclusters to test whether they were significantly different ($p < 0.05$). If two clusters were significantly different from all of the pairs at iteration *i*, the process repeats for $i + 1$ until at least one pair violates the condition, in which case the optimal number of *k* subclusters or reference populations within that population is the number of subclusters that did not violate the condition.

Assigning individuals to populations: We developed an admixture-based classifier that was not sensitive to the exclusion of random groups of individuals or the inclusion of large numbers of individuals from admixed groups, and was trained on a third of the data. Using a *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 populations. The assignment accuracy was measured against the population classification (Table S1).

2.2.4. Assessing Admixture Mapping

Creating hybrid individuals: We selected 15 individuals from five populations that showed homogeneity in their admixture components (Figure S4) and randomly sampled 120 pairs. Since selecting random alleles from each parent was problematic due to the high missingness of the data, we randomly selected half the genotypes of each parent to form 120 "offspring" or hybrid genomes. Each hybrid had three SNP sets: the CSS, PD aAIMs, and a random SNP set of the size of PD aAIMs with SNPs selected randomly for each hybrid.

Assessing admixture accuracy: Following [43,45,46], we applied a *supervised* ADMIXTURE to the three SNP sets of each hybrid.

3. Results

3.1. Depicting Ancient Population Structure

We constructed a dataset of 150,278 autosomal SNPs from 302 ancient genomes classified into 21 populations from Europe, the Middle East, and North Eurasia, and dated to time periods spanning from 14,000 years ago through to 1500 years ago (Figure 3, Table S1). These samples were chosen in order to obtain a broad temporal and geographical coverage. Nonetheless, due to the limited availability of ancient genomes, our dataset was not uniform over time and space. For instance, there were 57 Central European genomes from the Late Neolithic to the Bronze Age, but populations such as Mesolithic Central and Western Europeans, Bronze Age Jordanians, Chalcolithic Russians, and Mesolithic Russians comprised only of three genomes each. The population labels that we used corresponded directly with those from the published papers; in some cases, they were left unchanged, while in others cases we merged groups with similar admixture profiles in order to create broader, but homogenous populations.

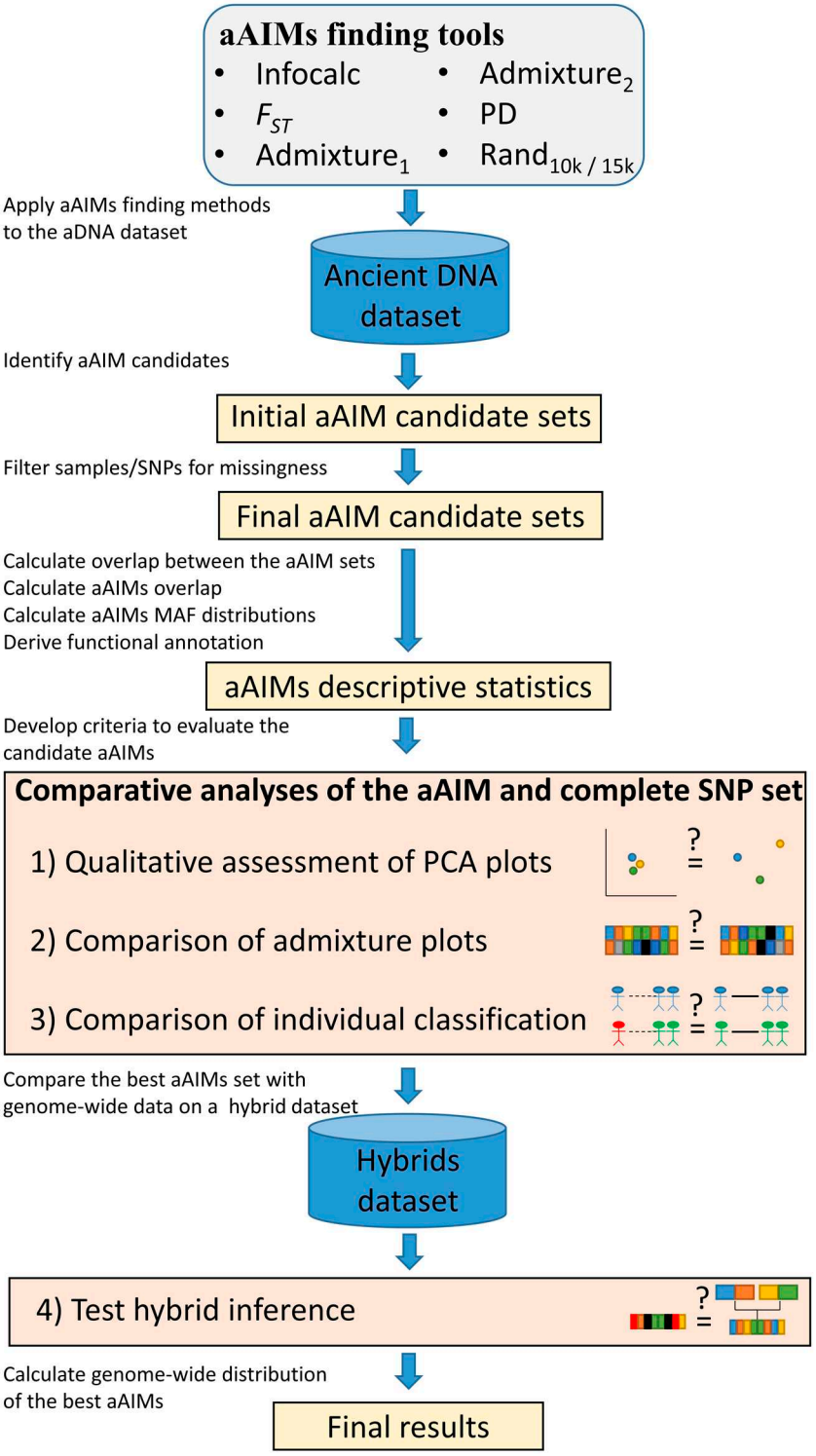

Figure 2. A workflow to identify and evaluate the accuracy of ancient ancestry informative markers (aAIM)-finding algorithms compared to each other as well as to the complete single nucleotide polymorphism (SNP) (CSS) set. We adopted four criteria to evaluate how well the aAIM candidates captured the population structure depicted by the CSS. First, we qualitatively compared the dispersal of genomes obtained from a principal component analysis (PCA) to that of the CSS. Second, we compared 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, we tested which aAIMs classified individuals to populations most accurately. Finally, we evaluated the ability of the top performing method to identify admixed individuals against the CSS. aDNA: ancient DNA.

Missingness varied greatly within the samples, as well as within the markers. The sample-based missingness ranged from 0.05% (KK1) to 99.2% (I1951), with a mean of 54%. Similarly, missingness also varied among the populations, with Levantine Epipaleolithic and Neolithic genomes having the highest missingness (n = 19, μ = 90, σ = 16) and Mesolithic Swedish genomes having the lowest (n = 8, μ = 29, σ = 27). The SNP-based missingness ranged from 30% to 98%, with an mean of 54%.

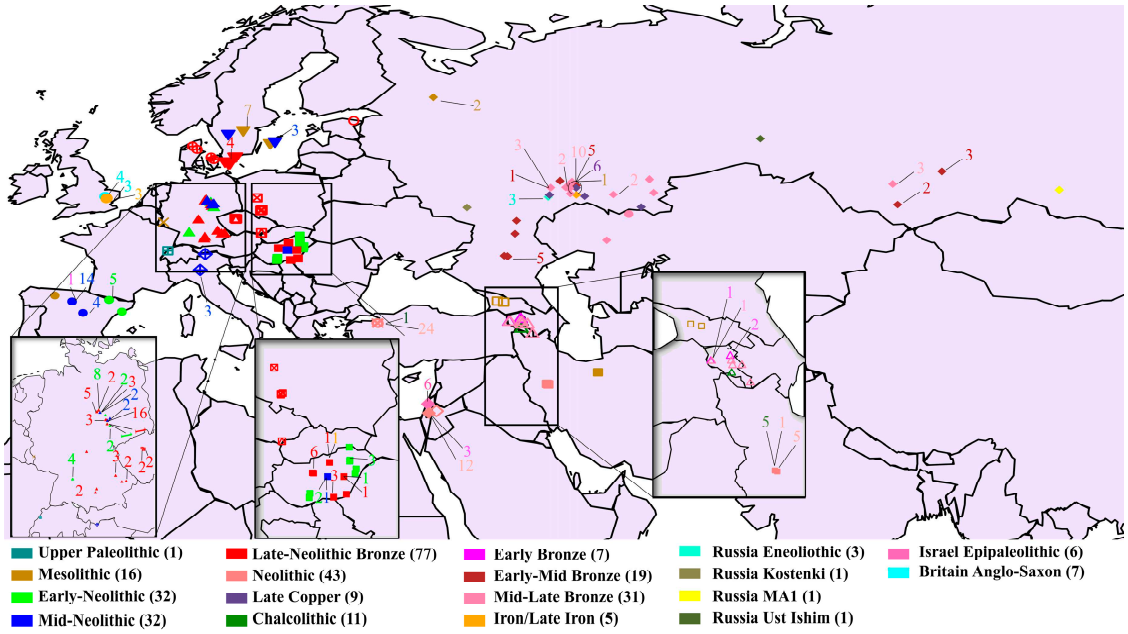

Figure 3. Geographical locations of the ancient genomes. The shapes designate the country of origin of the genomes and their colors designate the era. The total number of ancient genomes from each era is noted. Insets show densely sampled regions.

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 S1). 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 were clustered between Central European genomes. While ancient Spanish, Armenian, Central European Union (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.

Our *unsupervised* ADMIXTURE analysis with a range of splits (K) (Figure S2) found that no choice of K minimized the cross-validation error (CVE) (Figure S3), as expected in the analysis of monder-day populations, probably because the high noise and missingness in the data prevented the CVE from stabilizing. At K = 10 (Figure S4), multiple genomes (e.g., Britain Iron Saxon, Mesolithic Neolithic Caucasus population, Bronze Age Jordanian, Epipaleolithic Levantine, Chalcolithic, Mesolithic and Early Mid Bronze Russian, Early Neolithic Spanish, Mesolithic and Mid Neolithic Swedish, and Neolithic Turkish) appeared to be homogeneous in relation to their population and exhibited a distinct allelic frequency profile of admixture components. For these reasons, we decided to choose K = 10 as the optimal value. Furthermore, in this case, putative ancient ancestral components, such as the *Early Neolithic European* (brown) and the *Russia Mid Late Bronze* (magenta), which were predominantly found among European genomes, could be identified. Except for their predominance in Neolithic Turkish genomes, these two 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, which was likely due to noise.

3.2. Identifying and Describing the Ancient Ancestry Informative Markers Candidates

We developed a framework to identify and evaluate the efficacy of aAIM candidates in capturing ancient population structure and allowing admixture mapping (Figure 2). Ancient ancestry informative marker candidates were identified using five methods (Figure 2). Similar to the CSS, genomes and SNPs with over 90% missingness were removed, leaving each dataset with 223–263 genomes (Table S2). Furthermore, 310 SNPs without data were removed from the $Rand_{10k}$ dataset. The final number of aAIM candidates is shown in Table S3. Overlapping aAIMs between the methods are remarkably small and range from 560 ($Rand_{10k}$ and $Admixture_1$) to 2160 ($Admixture_1$ and $Admixture_2$). Interestingly, Infocalc and F_{ST}, which are often used together, share only ~10% of their aAIM candidates. The PD method shares 13.7% of its aAIMs with F_{ST} and ~10% with Infocalc.

Comparing the properties of the aAIM candidates (Figure S5a), we found that Infocalc prioritized SNPs with the lowest MAF (45% of the aAIMs have MAF < 0.1) and F_{ST} captured the aAIMs with a high frequency of low–mid MAFs. By contrast, PD and the admixture-based methods exhibited higher frequencies of high MAF SNPs, with $Admixture_2$ having the highest proportion of high MAF aAIMs (75% of the aAIMS have MAF >0.4). Remarkably, the MAF distributions exhibited a similarity with modern populations (Figure S5b), though, with fewer alleles in the lowest MAF bins for all the methods. Unsurprisingly, most of the aAIM variants were non-functional (94.6–96.3%) and varied little from the CSS's annotation (Table S4).

3.3. Comparative Testing of Ancient Ancestry Informative Marker Candidates

The accuracy of the aAIMs was evaluated using four criteria and by comparing each method against both CSS and two random SNP sets of sizes that approximated the number of aAIM candidates. We first calculated the PCA for each SNP set and compared the population dispersion along the primary two axes. Similarly to the CSS (Figure S1), all the methods depicted the Europe–Middle East contrast (PC1) and parallel clines (PC2) in the European genomes so that ancient Jordanian, Levantine, Turkic, and Spanish genomes clustered at one extreme of the Near East–Europe cline, whereas the genomes from Russia and Sweden clustered at the other end (Figure S6). However, similar as with the random sets, Infocalc and F_{ST} did not separate Levantine and Turkic individuals from each other. Infocalc also merged the Caucasus individuals with central Europeans. The admixture-based methods and PD clearly separated all of the ancient populations, similar to the CSS or more discernably, in the case of Scandinavians and Russians.

Secondly, we quantitatively assessed which dataset produced the closest admixture signature to that of the CSS (Figure S4). For that, we calculated the admixture proportions in relation to the 10 putatively ancient ancestral populations that we obtained with the CSS (Figure S7), and then computed their Euclidean distances (Figure S8) to their counterparts obtained with the CSS (Figure 4). The PD aAIMs led to significantly shorter Euclidean distances ($\mu = 0.13$, $\sigma = 0.1$, $n = 302$) compared to those obtained from the other aAIMs (Welch t-test: Infocalc ($t = 2.99$, p-value = 0.002), F_{ST} ($t = 7.32$, p-value = 8.5×10^{-13}), $Admixture_1$ ($t = 8.71$, p-value = 2.2×10^{-16}), $Admixture_2$ ($t = 9.89$, p-value = 2×10^{-16}), $Rand_{10k}$ ($t = 4.59$, p-value = 5×10^{-6}), and $Rand_{15k}$ ($t = 3.27$, p-value = 0.001)). Infocalc's aAIMs produced the second-shortest distances from the CSS ($\mu = 0.17$, $\sigma = 0.15$); however, these differences in the distances compared to those obtained with the two random datasets were not statistically significant (Welch t-test: $Rand_{10k}$ ($t = 1.56$, p-value = 0.12) and $Rand_{15k}$ ($t = 0.33$, p-value = 0.77), respectively), suggesting that Infocalc was unable to capture the population structure. F_{ST}-derived AIMs ($\mu = 0.2$, $\sigma = 0.13$) performed significantly worse than the $Rand_{15k}$ aAIMs (Welch t-test, $t = 2.89$, p-value 0.004), and similar to the $Rand_{10k}$ aAIMs (Welch t-test, $t = 1.5$, p-value = 0.13). Finally, the two admixture-based datasets performed the worst out of all the methods ($\mu_1 = 0.22$, $\sigma_1 = 0.15$ and $\mu_2 = 0.24$, $\sigma_1 = 0.16$) and significantly worse than the two random datasets (Welch t-test: $Admixture_1$ [$Rand_{10k}$ $t = 2.99$, p-value = 0.002] and [$Rand_{15k}$ $t = 4.35$, p-value = 1.6×10^{-5}]; $Admixture_2$ [$Rand_{10k}$ $t = 4.34$, p-value = 1.7×10^{-5}] and [$Rand_{15k}$ $t = 5.65$, p-value = 2.5×10^{-8}]).

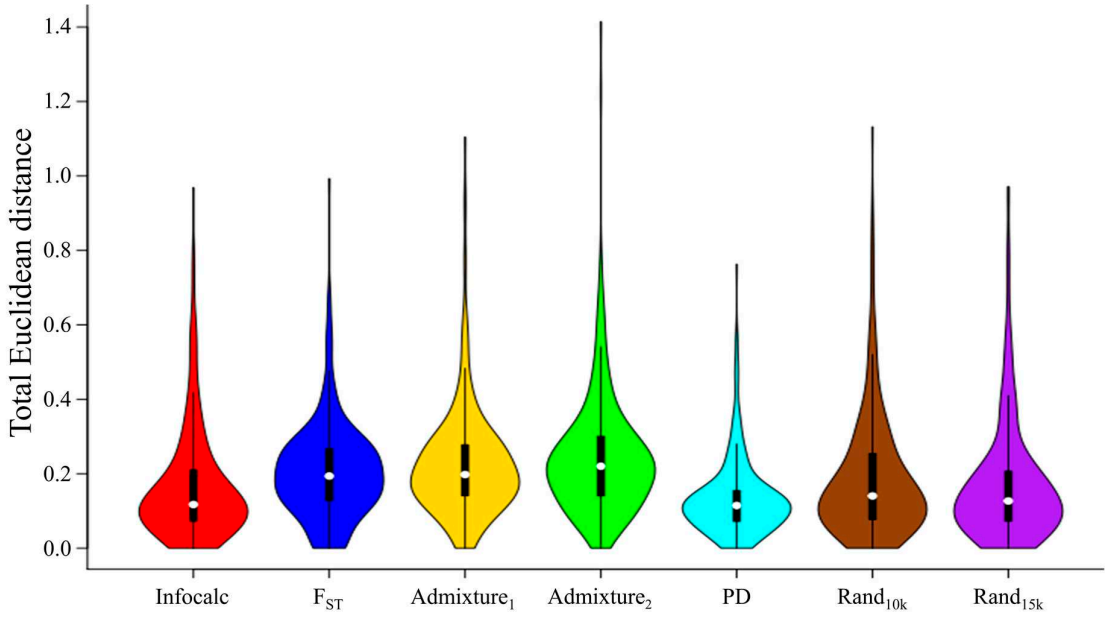

Figure 4. A comparison of the Euclidean distances (Δ) between the admixture proportions of the ancient genomes obtained from the CSS and those obtained from the aAIM sets using violin plots. Lower distances indicate high genetic similarity between the admixture proportions obtained using two different SNP sets.

Thirdly, we assessed which aAIMs dataset allowed classification of individuals into population groups most accurately. An admixture-based population classifier was applied to the admixture proportions produced by all of the datasets, and their accuracy was compared to that of the CSS ($76 \pm 5\%$) 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 of the 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 two random SNP sets ($\text{Rand}_{10k} = 42 \pm 5\%$ and $\text{Rand}_{15k} = 50 \pm 5\%$), with Infocalc ($50 \pm 6\%$) outperforming the remaining methods. Of note is the poor performance of F_{ST} aAIMs, which indicates its unsuitability for aDNA data. As expected, high missingness was associated with incorrect predictions (Figure S9). For example, the low-coverage, low-quality Britain Anglo-Saxon genomes proved challenging for all of the methods (0–40%), but predicted correctly with the CSS (100%).

3.4. Inference of Admixed Samples

The last criterion used to evaluate the accuracy of the aAIMs was to test whether they can identify hybrid individuals. Due to the high accuracy of the PD aAIMs in classifying individuals into populations, when compared to the alternative datasets, we decided to focus on aAIMs identified by the PD. Figure S10 illustrates the genome-wide distribution of PD aAIMs. To assess whether these aAIMs can identify hybrid individuals, ancient individuals were hybridized to form 120 mixed individuals who were represented in three datasets: CSS, PD aAIMs, and a random SNP set of the size of PD aAIMs (Table 2).

The genetic admixture distances between the hybrid individuals that were generated using the CSS and PD aAIMs were significantly smaller ($\mu = 0.05$, $\sigma = 0.04$) than the genetic admixture distances between the CSS and the random SNP set ($\mu = 0.45$, $\sigma = 0.15$, Welch t-test p-values $= 2.2 \times 10^{-8}$) and those between the PD and the random SNP set ($\mu = 0.43$, $\sigma = 0.15$, Welch t-test p-values $= 1.9 \times 10^{-8}$). Thus, we demonstrated that PD aAIMs can be used for studying admixed individuals and can be potentially used in future admixture mapping involving aDNA.

Table 1. Accuracy in classifying individuals to populations using the aAIM candidates. The total number of individuals (*n*) per population are reported in column two. Columns three to eight show the number of individuals correctly predicted to their populations and, in brackets, the corresponding population percentage. Columns seven and eight effectively represent a random number of 10000 and 15000 SNPs, respectively. Mean and standard error 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 Chalc	6	2 (33)	3 (50)	0 (0)	3 (50)	3 (50)	3 (50)	2 (33)	2 (33)
Central North EU Late Neol Bronz	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 ± 5	61 ± 5	3 ± 2	50 ± 6	21 ± 5	33 ± 7	42 ± 5	50 ± 5

EU: Europe. CSS: Complete single nucleotide polymorphism (SNP) set; PD: Principal component analysis (PCA)-derived.

Table 2. Accuracy of inferring hybrid individuals using the PD's aAIMs. The six parental populations and the number of hybrid individuals generated from them are shown. Each hybrid was represented by three datasets: CSS, PD aAIMs, and a random SNP set. The mean genetic distances (*d*) between the admixture components of these datasets per population are shown. Short distances indicate high genetic similarity.

Parental Population A	Parental Population B	# Hybrids	$\overline{d(\mathbf{CSS},\ \mathbf{PD})}$	$\overline{d(\mathbf{CSS},\ \boldsymbol{random\ set})}$	$\overline{d(\mathbf{PD},\ \boldsymbol{random\ set})}$
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

CSS: Complete single nucleotide polymorphism (SNP) set; PD: Principal component analysis (PCA)-derived.

4. Discussion

Questions of identity and primordialism are at the center of scientific and public debate. Until recently, charting the emergence of agriculture, the spread of languages, and the rise and decline of cultures were topics dominated by archeologists. The emergence of aDNA allows paleogeneticists to delve into this debate with a discordant set of assumptions about biology and identity [47]. This was not unforeseen, as population genetic analyses excel at identifying individual differences, which can inform archeologically contended subjects such as migration and the degree of admixture or population replacements. However, aDNA analyses also require destroying genetic material, sometimes irrevocably, which makes them impossible to replicate. It is therefore crucial to develop a robust genetic methodology that uses population genetic principles to examine the assumptions made by both archeologists and paleogeneticists. It is reasonable to expect that many of the tools employed to study modern-day genomes will need to be adapted to the four-dimensional environment facilitated by aDNA.

Ancestry informative markers are some of the most useful tools in addressing population, biomedical, forensics, and evolutionary questions that remain in use today [9,48–50]. However, it is unclear to what extent known AIMs are applicable to ancient genomic data, which are characterized by high missingness and haploidy [1].

In this study, we defined aAIMs (Figure 1) and sought to identify them using various methods. The number of aAIM candidates detected by each method ranged from 9,000 to 15,000. These numbers are of the same magnitude as large AIMs studies (e.g., [51,52]) and reasonable, provided that there is potential relatedness of the ancient Eurasian populations and the near absence of heterozygote markers in the data. To find which of the aAIM candidate sets produced by each method best represent the true population structure, we used the CSS as a benchmark for qualitative and quantitative comparisons.

Identifying the ideal AIM set that would be both small and include redundancies (in the case of sequencing failure), capture the population structure, and allow the identification of admixed individuals is one of the challenges of population genetics. We showed that the aAIMs identified through the PD method outperformed all other methods, in agreement with previous studies that tested PCA-based methods [25]. In forty percent of the populations, classifications made by the PD method were more accurate than those made using the CSS (Table 1), which highlights the limitations of using markers indiscriminately. This is not surprising, since not all the markers are equally informative, and less informative markers (e.g., exonic markers) may mask the population structure, resulting in the misclassification of populations. The notion of "more is better" is, hence particularly misguided with aDNA that harbors a multi-layered population structure in a poor set of markers. The application of the PD aAIMs for admixture mapping, combined with tools that can homogenize cases and controls [16], enables the carrying out of future association studies on aDNA samples (e.g., [3]). Further investigations with additional data may identify formerly common markers associated with those disease that with time became rare and undetectable.

The use of PCA to infer population structure is controversial [53–56], and its use as a clustering method has been criticized [16]. We note that the PD method employs PCA only to produce and replicate a population structure profile of certain subpopulations based on various sets of markers and does not make claims that the PCA-derived profiles represent the true genetic distances between individuals.

Surprisingly, Infocalc and F_{ST} that are commonly used to identify AIMs [18] and are reported to perform well [57], have oftentimes underperformed random SNP selections. Not only was F_{ST} already shown to be particularly small within continental populations [58], but these methods may be particularly sensitive to aDNA data that are both haploid and have high missingness (Figure S9). We also found no relationships between the performances of MAF and aAIMs (Figure S5). Enrichment for high or low MAF SNPs did not guarantee success, although the PD harbored more common SNPs than most of the underperforming methods.

Our study has several limitations. We studied an uneven number of Eurasian populations from various times and locations, causing a skew toward markers that predict central European populations from the Late Neolithic and Bronze Age. A modest attempt to reduce this bias was made by including modern-day African and Asian populations; however, more comprehensive analyses should be made when more global genomes are available from consecutives eras. 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 the high missingness of the data, it is likely that our study missed informative markers that could improve the classification accuracy in newly sequenced populations. Therefore, our framework and methods must be applied again when more comprehensive aDNA datasets are available.

5. Conclusions

The use of ancient genomes in research is in its infancy, and is expected to intensify and expand to new fields as more data become available. One of the main advantages of aDNA is that it widens the number of ancestry types and makes them multi-faceted, requiring fine-tuned molecular utilities to depict ancestry over time. AIMs are some of the most effective tools that have spear-headed population genetics over the past two decades and are ancillary to the challenge of understanding population structure. Here, we defined aAIMs, proposed a framework to evaluate AIM-finding methods, demonstrated the accuracy of a novel aAIM-finding method, and reported the most successful set of aAIMs. Future analyses may benefit from using our framework, methods, and aAIMs in order to refine ancient population structure models and examine primordialist theories.

Supplementary Materials: The following are available online at http://www.mdpi.com/2073-4425/9/12/625/s1, Figure S1: Scatter plot of all ancient populations along the first two principal components, Figure S2: ADMIXTURE results of ancient genomes at K=2 through K=12, Figure S3: Cross-validation (CV) error and standard error of ADMIXTURE runs of K ranging from 1 to 15, Figure S4: Ancient population structure inferred by ADMIXTURE analysis, Figure S5: Minor allele frequency distributions for aAIMs identified with various methods, Figure S6: PCA plots of the aAIMs candidates identified by various methods, Figure S7: A supervised ADMIXUTRE analysis of the aAIMs candidates identified by various methods, Figure S8: A distribution of the Euclidean distances between the admixture proportions of the ancient genomes obtained from the CSS and those obtained by the aAIMs of each method, Figure S9: The effect of data missingness in predictions made using the CSS, Figure S10: Genome wide distribution of SNPs in the CSS (dots) and PD (red bars) datasets, Figure S11: Illustrating how PCA-derived (PD) aAIMs are obtained for Caucasus populations (Chalcolithic – Bronze), Table S1: Summary of aDNA samples used in this study, Table S2: Genomes retained in each method after removing samples with high missingness for the dataset of 302 genomes, Table S3: Overlapping SNPs with rs# between different methods, Table S4: Functional and frequency annotation of the studied SNPs, Additional File 1: aDNA dataset, Additional File 2: aAIM candidates produced by various tools, Additional File 3: ancient putative ancestral populations.

Author Contributions: Conceptualization, E.E.; methodology, E.E., U.E., R.D. and M.P.; software, U.E. and R.D.; validation, E.E. and U.E. formal analysis, E.E., U.E., R.D. and M.P.; investigation, E.E., U.E., R.D. and M.P.; data curation, U.E., R.D. and M.P.; writing—original draft preparation, E.E. and U.E.; writing—review and editing, E.E., U.E. and S.S.; visualization, E.E., U.E. and R.D.; supervision, E.E.; project administration, E.E.; funding acquisition, E.E.

Funding: This study was partially supported by the MRC Confidence in Concept Scheme award 2014-University of Sheffield to E.E. (Ref: MC_PC_14115), MRC (MR/R025126/1) to E.E., and funding from the DNA Diagnostics Center, Inc.

Acknowledgments: 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.

Conflicts of Interest: EE is a consultant to DNA Diagnostic Centre and DNA Consultants. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Morozova, I.; Flegontov, P.; Mikheyev, A.S.; 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. [CrossRef] [PubMed]

2. Marciniak, S.; Perry, G.H. Harnessing ancient genomes to study the history of human adaptation. *Nat. Rev. Genet.* **2017**, *18*, 659–674. [CrossRef] [PubMed]
3. Cassidy, L.M.; Martiniano, R.; Murphy, E.M.; Teasdale, M.D.; Mallory, J.; Hartwell, B.; Bradley, D.G. Neolithic and Bronze Age migration to Ireland and establishment of the insular Atlantic genome. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 368–373. [CrossRef] [PubMed]
4. Patterson, N.J.; Moorjani, P.; Luo, Y.; Mallick, S.; Rohland, N.; Zhan, Y.; Genschoreck, T.; Webster, T.; Reich, D. Ancient admixture in Human history. *Genetics* **2012**, *192*, 1065–1093. [CrossRef] [PubMed]
5. Mathieson, I.; Lazaridis, I.; Rohland, N.; Mallick, S.; Patterson, N.; Roodenberg, S.A.; Harney, E.; Stewardson, K.; Fernandes, D.; Novak, M.; et al. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **2015**, *528*, 499–503. [CrossRef] [PubMed]
6. Fu, Q.; Hajdinjak, M.; Moldovan, O.T.; Constantin, S.; Mallick, S.; Skoglund, P.; Patterson, N.; Rohland, N.; Lazaridis, I.; Nickel, B.; et al. An early modern human from Romania with a recent Neanderthal ancestor. *Nature* **2015**, *524*, 216–219. [CrossRef] [PubMed]
7. Lazaridis, I.; Nadel, D.; Rollefson, G.; Merrett, D.C.; Rohland, N.; Mallick, S.; Fernandes, D.; Novak, M.; Gamarra, B.; Sirak, K.; et al. Genomic insights into the origin of farming in the ancient Near East. *Nature* **2016**, *536*, 419–424. [CrossRef] [PubMed]
8. Li, J.Z.; Absher, D.M.; Tang, H.; Southwick, A.M.; Casto, A.M.; Ramachandran, S.; Cann, H.M.; Barsh, G.S.; Feldman, M.; Cavalli-Sforza, L.L.; et al. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **2008**, *319*, 1100–1104. [CrossRef] [PubMed]
9. Elhaik, E.; Yusuf, L.; Anderson, A.I.J.; Pirooznia, M.; Arnellos, D.; Vilshansky, G.; Ercal, G.; Lu, Y.; Webster, T.; Baird, M.L.; et al. 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. [CrossRef] [PubMed]
10. 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. [CrossRef]
11. Hublin, J.-J. The last Neanderthal. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 10520–10522. [CrossRef] [PubMed]
12. Jones, S. *The Archaeology of Ethnicity: Constructing Identities in the Past and Present*; Routledge: London, UK, 1997.
13. Albrechtsen, A.; Nielsen, F.C.; Nielsen, R. Ascertainment biases in SNP chips affect measures of population divergence. *Mol. Biol. Evol.* **2010**, *27*, 2534–2547. [CrossRef] [PubMed]
14. Marchini, J.; Cardon, L.R.; Phillips, M.S.; Donnelly, P. The effects of human population structure on large genetic association studies. *Nat. Genet.* **2004**, *36*, 512–517. [CrossRef]
15. Yusuf, S.; Wittes, J. Interpreting geographic variations in results of randomized, controlled trials. *N. Engl. J. Med.* **2016**, *375*, 2263–2271. [CrossRef]
16. Elhaik, E.; Ryan, D.M. Pair Matcher (PaM): Fast model-based optimisation of treatment/case-control matches. *Bioinformatics* **2018**. [CrossRef]
17. Elhaik, E.; Tatarinova, T.; Chebotarev, D.; Piras, I.S.; 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. [CrossRef]
18. 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. [CrossRef] [PubMed]
19. Kosoy, R.; Nassir, R.; Tian, C.; White, P.A.; Butler, L.M.; Silva, G.; Kittles, R.; Alarcon-Riquelme, M.E.; Gregersen, P.K.; Belmont, J.W.; et al. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum. Mutat.* **2009**, *30*, 69–78. [CrossRef]
20. Qin, H.; Zhu, X. Power comparison of admixture mapping and direct association analysis in genome-wide association studies. *Genet. Epidemiol.* **2012**, *36*, 235–243. [CrossRef] [PubMed]
21. Barbosa, F.B.; Cagnin, N.F.; Simioni, M.; Farias, A.A.; Torres, F.R.; Molck, M.C.; Araujo, T.K.; Gil-Da-Silva-Lopes, V.L.; Donadi, E.A.; Simões, A.L. Ancestry informative marker panel to estimate population stratification using genome-wide human array. *Ann. Hum. Genet.* **2017**, *81*, 225–233. [CrossRef] [PubMed]

22. Peng, Q.; Schork, N.J.; Wilhelmsen, K.C.; Ehlers, C.L. Whole genome sequence association and ancestry-informed polygenic profile of EEG alpha in a Native American population. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **2017**, *174*, 435–450. [CrossRef] [PubMed]
23. Shriner, D. Overview of admixture mapping. *Curr. Protoc. Hum. Genet.* **2013**, *76*, 1–23. [CrossRef] [PubMed]
24. Kidd, K.K.; Speed, W.C.; Pakstis, A.J.; Furtado, M.R.; Fang, R.; Madbouly, A.; Maiers, M.; Middha, M.; Friedlaender, F.R.; Kidd, J.R. Progress toward an efficient panel of SNPs for ancestry inference. *Forensic Sci. Int. Genet.* **2014**, *10*, 23–32. [CrossRef] [PubMed]
25. Huckins, L.M.; Boraska, V.; Franklin, C.S.; Floyd, J.A.; Southam, L.; Boraska, V.; Franklin, C.S.; Floyd, J.A.; Thornton, L.M.; Huckins, L.M.; et al. Using ancestry-informative markers to identify fine structure across 15 populations of European origin. *Eur. J. Hum. Genet.* **2014**, *22*, 1190–1200. [CrossRef]
26. 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. [CrossRef] [PubMed]
27. Pakstis, A.J.; Kang, L.; Liu, L.; Zhang, Z.; Jin, T.; Grigorenko, E.L.; Wendt, F.R.; Budowle, B.; Hadi, S.; Al Qahtani, M.S.; et al. Increasing the reference populations for the 55 AISNP panel: The need and benefits. *Int. J. Leg. Med.* **2017**, *131*, 913–917. [CrossRef] [PubMed]
28. Rosenberg, N.A.; Li, L.M.; Ward, R.; Pritchard, J.K. Informativeness of genetic markers for inference of ancestry. *Am. J. Hum. Genet.* **2003**, *73*, 1402–1422. [CrossRef]
29. Kidd, J.R.; Friedlaender, F.R.; Speed, W.C.; Pakstis, A.J.; De La Vega, F.M.; Kidd, K.K. Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples. *Investig. Genet.* **2011**, *2*, 1. [CrossRef]
30. Wright, S. *Evolution and the Genetics of Populations. A Treatise in Three Volumes*; University of Chicago Press: Chicago, IL, USA, 1968.
31. Marshall, S.; Das, R.; Pirooznia, M.; Elhaik, E. Reconstructing Druze population history. *Sci. Rep.* **2016**, *6*, 35837. [CrossRef]
32. Lazaridis, I.; Patterson, N.; Mittnik, A.; Renaud, G.; Mallick, S.; Kirsanow, K.; Sudmant, P.H.; Schraiber, J.G.; Castellano, S.; Lipson, M. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **2014**, *513*, 409–413. [CrossRef]
33. Marcus, J.H.; Novembre, J. Visualizing the geography of genetic variants. *Bioinformatics* **2017**, *33*, 594–595. [CrossRef] [PubMed]
34. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [CrossRef] [PubMed]
35. 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. [CrossRef] [PubMed]
36. McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernytsky, A.; Garimella, K.; Altshuler, D.; Gabriel, S.; Daly, M.; et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **2010**, *20*, 1297–1303. [CrossRef] [PubMed]
37. DePristo, M.A.; Banks, E.; Poplin, R.; Garimella, K.V.; Maguire, J.R.; Hartl, C.; Philippakis, A.A.; del Angel, G.; Rivas, M.A.; Hanna, M.; et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **2011**, *43*, 491–498. [CrossRef] [PubMed]
38. Danecek, P.; Auton, A.; Abecasis, G.; Albers, C.A.; Banks, E.; DePristo, M.A.; Handsaker, R.E.; Lunter, G.; Marth, G.T.; Sherry, S.T.; et al. The variant call format and VCFtools. *Bioinformatics* **2011**, *27*, 2156–2158. [CrossRef] [PubMed]
39. Purcell, S.; Neale, B.; Todd-Brown, K.; Thomas, L.; Ferreira, M.A.; Bender, D.; Maller, J.; Sklar, P.; de Bakker, P.I.; Daly, M.J.; et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **2007**, *81*, 559–575. [CrossRef]
40. Alexander, D.H.; Novembre, J.; Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **2009**, *19*, 1655–1664. [CrossRef]
41. Durbin, R.M.; Abecasis, G.R.; Altshuler, D.L.; Auton, A.; Brooks, L.D.; Gibbs, R.A.; Hurles, M.E.; McVean, G.A. A map of human genome variation from population-scale sequencing. *Nature* **2010**, *467*, 1061–1073. [CrossRef]
42. Paschou, P.; Ziv, E.; Burchard, E.G.; Choudhry, S.; Rodriguez-Cintron, W.; Mahoney, M.W.; Drineas, P. PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS Genet.* **2007**, *3*, 1672–1686. [CrossRef]

43. Das, R.; Wexler, P.; Pirooznia, M.; Elhaik, E. Localizing Ashkenazic Jews to primeval villages in the ancient Iranian lands of Ashkenaz. *Genome Biol. Evol.* **2016**, *8*, 1132–1149. [CrossRef] [PubMed]
44. Das, R.; Wexler, P.; Pirooznia, M.; Elhaik, E. The Origins of Ashkenaz, Ashkenazic Jews, and Yiddish. *Front. Genet.* **2017**, *8*, 87. [CrossRef] [PubMed]
45. Baughn, L.B.; Pearce, K.; Larson, D.; Polley, M.-Y.; Elhaik, E.; Baird, M.; Colby, C.; Benson, J.; Li, Z.; Asmann, Y.; et al. Differences in genomic abnormalities among African individuals with monoclonal gammopathies using calculated ancestry. *Blood Cancer J.* **2018**, *8*, 1–10. [CrossRef] [PubMed]
46. Elhaik, E. In search of the *Jüdische Typus*: A proposed benchmark to test the genetic basis of Jewishness challenges notions of "Jewish biomarkers". *Front. Genet.* **2016**, *7*, 141. [CrossRef] [PubMed]
47. Callaway, E. Divided by DNA: The uneasy relationship between archaeology and ancient genomics. *Nature* **2018**, *555*, 573–576. [CrossRef]
48. Bose, N.; Carlberg, K.; Sensabaugh, G.; Erlich, H.; Calloway, C. Target capture enrichment of nuclear SNP markers for massively parallel sequencing of degraded and mixed samples. *Forensic Sci. Int. Genet.* **2018**, *34*, 186–196. [CrossRef] [PubMed]
49. Bulbul, O.; Speed, W.C.; Gurkan, C.; Soundararajan, U.; Rajeevan, H.; Pakstis, A.J.; Kidd, K.K. Improving ancestry distinctions among Southwest Asian populations. *Forensic Sci. Int. Genet.* **2018**, *35*, 14–20. [CrossRef]
50. López-Cortés, A.; Echeverría-Garcés, G.; Burgos, G.; Zambrano, A.; Cabrera-Andrade, A.; García-Cárdenas, J.; Salazar, C.; Leone, P.; Paz-y-Miño, C. Molecular analysis of ancestry informative markers (AIMs-INDELs) in a high altitude Ecuadorian mestizo population affected with breast cancer. *Forensic Sci. Int. Genet. Suppl. Ser.* **2017**, *6*, e231–e232. [CrossRef]
51. Tian, C.; Hinds, D.A.; Shigeta, R.; Adler, S.G.; Lee, A.; Pahl, M.V.; Silva, G.; Belmont, J.W.; Hanson, R.L.; Knowler, W.C.; et al. A genomewide single-nucleotide-polymorphism panel for Mexican American admixture mapping. *Am. J. Hum. Genet.* **2007**, *80*, 1014–1023. [CrossRef] [PubMed]
52. 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. [CrossRef]
53. Arenas, M.; Francois, O.; Currat, M.; Ray, N.; Excoffier, L. Influence of admixture and paleolithic range contractions on current European diversity gradients. *Mol. Biol. Evol.* **2013**, *30*, 57–61. [CrossRef] [PubMed]
54. Elhaik, E. The missing link of Jewish European ancestry: Contrasting the Rhineland and the Khazarian hypotheses. *Genome Biol. Evol.* **2013**, *5*, 61–74. [CrossRef] [PubMed]
55. McVean, G. A genealogical interpretation of principal components analysis. *PLoS Genet.* **2009**, *5*, e1000686. [CrossRef] [PubMed]
56. Novembre, J.; Stephens, M. Interpreting principal component analyses of spatial population genetic variation. *Nat. Genet.* **2008**, *40*, 646–649. [CrossRef] [PubMed]
57. Ding, L.; Wiener, H.; Abebe, T.; Altaye, M.; Go, R.C.; Kercsmar, C.; Grabowski, G.; Martin, L.J.; Hershey, G.K.; Chakorborty, R.; et al. Comparison of measures of marker informativeness for ancestry and admixture mapping. *BMC Genom.* **2011**, *12*, 622. [CrossRef] [PubMed]
58. Elhaik, E. Empirical distributions of F_{ST} from large-scale Human polymorphism data. *PLoS ONE* **2012**, *7*, e49837. [CrossRef] [PubMed]

© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).