Insights from the Analysis of Ancient and Modern DNAs with Population-specific SNPs

Gang Shi (gshi@xidian.edu.cn)
Xidian University

Article

Keywords:

Posted Date: December 11th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3447042/v1

Additional Declarations: No competing interests reported.

Insights from the Analysis of Ancient and Modern DNA with Population-specific SNPs

Gang Shi[1*]

[1]School of Telecommunications Engineering, Xidian University, 2 South Taibai Road, Xi'an, Shaanxi, 710071, China, gshi@xidian.edu.cn

[*]Corresponding author

Abstract

Studies of ancient and modern DNA have substantially improved our understanding of the early history of human populations. Despite the advancement of whole-genome sequencing technologies, present studies of ancient DNA (aDNA) are largely based on a panel of preselected genomic variants; thus, valuable genetic information in aDNA should be further explored. In this work, we analyze genotype data from 19 ancient and 16 modern high-coverage shotgun human genomes. We used modern populations from the 1000 Genomes Project and the Human Genome Diversity Project as reference populations and selected SNPs that were polymorphic in one reference population and monomorphic in the others. With the population-specific SNPs, we conducted ancestral spectrum analyses on the 19 aDNA and the 16 modern DNA to determine their coancestries with the modern reference populations. We show that ancestral spectrum analyses effectively reveal the genetic affinity between aDNA and modern populations, which is also true for modern DNA. Regarding the 11 aDNA with expected transition to transversion ratios, the results agree with previous analyses. The other 8 aDNA with excessive transition to transversion ratios revealed ancestral spectra indicative of a high level of DNA damage that cannot be fully explained by postmortem cytosine deamination. UDG treatment or bioinformatics treatments seem necessary for the meaningful study of such aDNA.

Introduction

Numerous ancient and modern human DNA have been collected, sequenced and analyzed to study the early histories of human populations, including population expansion, divergences, admixture events, migrations and introgressions from Neanderthals and Denisovans [1, 2]. In particular, ancient DNA (aDNA) from archaeological sites provide information not only on the origins of modern human ancestry but also about the geographic distributions of human populations in history. Nevertheless, genetic information in aDNA should be further explored. First, present studies typically focus on a set of genomic positions that are polymorphic among a worldwide panel of modern human populations [3]. This leaves a large number of less frequent or rare variants unexplored. Most of the less frequent or rare variants in modern populations are population-specific [4, 5] and hence are informative for understanding the affinity between aDNA and modern populations. Second, existing analyses of aDNA are largely based on principal component (PC) analyses [6, 7]. Genotypes of aDNA are routinely projected onto the principal directions established by modern populations, which serve as reference populations, and the genetic relationships between the aDNA and the modern populations are examined in the PC space. In the case of two-way admixture, the connection between eigenvectors and ancestral proportions has been established, which is more complicated and less intuitive for high-dimension admixtures [8]. Therefore, PC plots are useful for demonstrating genetic distances between the aDNA and the modern populations, but not the precise quantification of their ancestral relationships or genetic affinity.

Here, we conducted ancestral spectrum analyses [5] on 19 ancient and 16 modern high-coverage shotgun human genomes [9]. Genotype data were obtained by whole-genome shotgun sequencing and thus are not limited to the genomic variants commonly used in existing studies [3] and include many less frequent or rare variants that have not been analyzed previously. We used modern populations from the 1000 Genomes Project (1kGP) [4] and the Human Genome Diversity Project (HGDP) [10] as reference populations: the first panel of reference populations included 5 populations from the 1kGP and HGDP, and second comprised 7 populations from HGDP. Then, we selected two panels of single nucleotide polymorphisms (SNPs) that were polymorphic in one reference population and monomorphic in the others using the two panels of reference populations. The population-specific SNPs were used to conduct ancestral spectrum analyses [5] on the 19 aDNA and the 16 modern DNA.

It has been shown that the estimated ancestral information corresponding to a reference population is the best linear unbiased estimator of the ancestral proportion of a sample being analyzed if the reference population approximates an ancestral population well [5]. Otherwise, this approach provides a quantitative measure of coancestry between the sample and the corresponding reference population [5]. In this paper, the ancestral spectrum of an aDNA or a modern DNA is a set of ancestral information measuring the coancestries between the sample and the reference populations that should not be considered ancestral proportions of the sample. Details on the ancestral spectrum analysis were previously described in [5].

Methods

The panel of population-specific SNPs with 5 reference populations

Among the 2504 unrelated individuals from the 1kGP, we excluded those from American (AMR) and African American populations, which are known to be recently admixed [4]. We selected African (AFR), East Asian (EAS), European (EUR) and South Asian (SAS) populations in the 1kGP as the reference populations with sample sizes of 504, 504, 503 and 489, respectively. With the genotype data in phase 3 [4], we extracted diallelic SNPs without multicharacter alleles using PLINK 1.9 [11] and retained those in neutral genomic regions [12] that was used in the pipeline to process the 19 ancient and 16 modern high-coverage shotgun human genomes [9]. We retained autosomal SNPs with reference SNP numbers in dbSNP build 151 [13] and 16,575,155 SNPs passed all filters.

We used 61 indigenous AMRs from the HGDP as the reference population for the AMR population. We extracted diallelic SNPs without multicharacter alleles using PLINK 1.9 from the high-coverage genotype dataset [14] and retained autosomal SNPs with reference SNP numbers in dbSNP build 151. The genomic positions were converted to GRCh37 coordinates and those outside of the neutral genomic regions [12] were removed. After filtering, 10,421,194 SNPs remained.

We removed 47,468 SNPs with inconsistent alleles between the two datasets and merged the genotype data of the 5 reference populations from the 1kGP and HGDP, leaving 19,855,165 unique SNPs in the combined dataset. Population-specific SNPs polymorphic in one of the five reference populations were selected using PSNPS, a utility enclosed in ASA (https://github.com/eat1000/ASA) [5]. For AFR, EAS, EUR

and SAS, we screened population-specific SNPs with minor allele frequencies (MAFs) between 0.01 and 0.05 and for AMR with MAFs between 0.02 and 0.2, because of the small sample size of the AMR population. There were 747,070, 13,953, 74,877, 24,920 and 82,455 SNPs specific to the AFR, AMR, EAS, EUR and SAS populations, respectively. For AFR, EAS and SAS, we randomly chose 25,000 population-specific SNPs. The first panel of population-specific SNPs with the 5 reference populations comprises 113,873 SNPs, details of which are presented in Supplementary Table S1.

The panel of population-specific SNPs with 7 reference populations

The second panel of reference populations was based on populations in the HGDP. The 929 individuals from 54 worldwide populations were grouped into 7 reference populations according to their geographic regions [14]: Africa (*N*=104), America (*N*=61), Central and South Asia (*N*=197), East Asia (*N*=223), Europe (*N*=155), Middle East (*N*=161) and Oceania (*N*=28).

Population-specific SNPs were screened using PSNPS among the 10,421,194 SNPs after filtering as described previously. Population-specific SNPs for the reference populations from America and Oceania were restricted to those with MAFs between 0.02 and 0.2 due to their small sample sizes. For the other 5 reference populations, the MAFs of the population-specific SNPs were between 0.01 and 0.05. There were 267,951, 12,391, 24,389, 28,932, 7170, 14,519 and 32,758 SNPs specific to populations from Africa, America, Central and South Asia, East Asia, Europe, Middle East and Oceania, respectively. For the populations with over 25,000 population-specific SNPs, we randomly chose 25,000 SNPs. The second panel of population-specific SNPs with

the 7 reference populations comprises 133,469 SNPs, shown in Supplementary Table S2.

Analysis of the 19 aDNA and the 16 modern DNA

We used genotype data from the 19 aDNA and the 16 modern DNA obtained by whole-genome sequencing, which included 509,351,727 sites in neutral regions before filtering [9]. Details on the samples and the genetic pipeline used to process them were described previously, see [9] and the references therein. We conducted the same filtering used for the 1kGP genotype data and 99,754,278 SNPs remained. Based on the two panels of population-specific SNPs, coancestries between the study samples and the two panels of reference populations were analyzed by computing their ancestral information vectors with Ancestral Spectral Analyzer (ASA) version 1.1.0 [5].

Analysis of 15 modern DNA from the Simons Genome Diversity Project

Among the 16 modern DNA, 15 were from the Simons Genome Diversity Project (SGDP) [15]. We extracted the genotypes of the 15 samples from the dataset released by SGDP [15]. After filtering as described for the 1kGP dataset, 7,613,259 SNPs remained. Coancestries between the 15 SGDP samples and the two panels of reference populations were analyzed by computing their ancestral information vectors with the two panels of population-specific SNPs, using ASA version 1.1.0.

Results

Ancestral spectra with the 5 reference populations

In the first panel of 113,873 population-specific SNPs, 86,931 were available in the genotype dataset of the 19 aDNA and 16 modern DNA. The estimated ancestral

information vectors of the 35 samples are shown in Supplementary Table S3.

The ancestral spectra of 11 aDNA with expected transition to transversion (ts/tv) ratios [9] are presented in Figure 1. ZVEJ25, ZVEJ31, SF12, Loschbour, Stuttgart, NE5 and KK1 have almost exclusive ancestral information from EUR, indicating that they share common ancestry with modern EURs. This finding is consistent with the fact that all 7 samples were collected in Europe and their estimated dates range from 9712 to 5965 before present (BP). AHUR_2064 and USR1 are largely coancestral with AMR, while USR1 also shows some coancestries with EAS and EUR. This finding indicates that AHUR_2064 shares common ancestry with modern indigenous AMRs. Since AHUR_2064 was from Nevada, USA, and USR1 was from Alaska, USA, and their estimated dates, which were 10,970 and 11,435 BP, respectively, were close, this finding confirms that ancestors of the modern indigenous AMR population split from the ancient AMR population and migrated further south [2, 16]. Ust'Ishim and Sunghir III were the most ancient genomes among the samples, dated to 45,000 and 34,093 BP, respectively. Because the population-specific variants based on modern populations are often recent mutations, the two aDNA contained the lowest total level of ancestral information. Ust'Ishim shows a small degree of coancestry with EAS, EUR and SAS populations and Sunghir III with EUR and SAS populations.

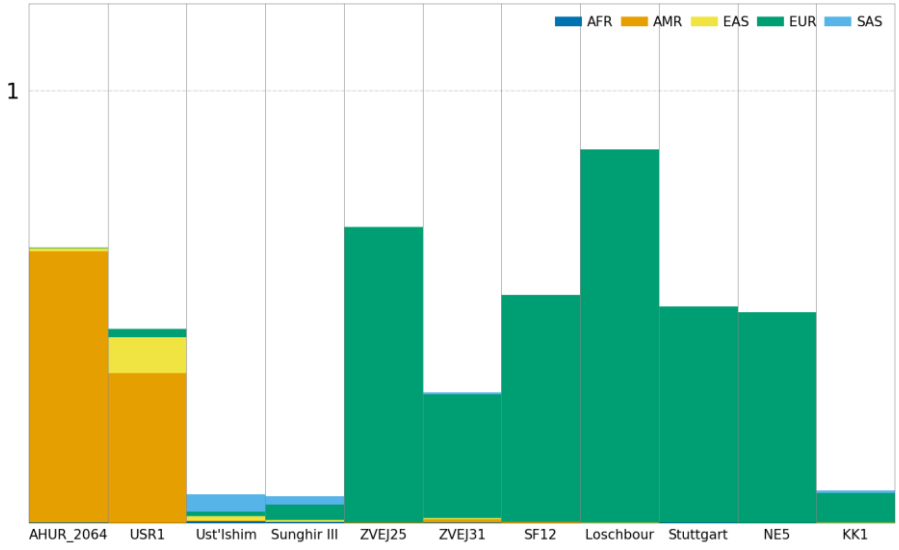

Figure 1. Ancestral spectra of the 11 aDNA with expected ts/tv ratios; 5 reference populations.

The ancestral spectra of the other 8 aDNA with excessive ts/tv ratios [9] are presented in Figure 2. These samples all demonstrate different levels of coancestries with 4 or 5 reference populations. A high ts/tv ratio usually suggests a large number of errors due to postmortem cytosine deamination [3]. We constructed another SNP panel restricted to transversion substitutions, including 25,000, 4767, 25,000, 7871 and 25,000 SNPs specific to the AFR, AMR, EAS, EUR and SAS populations, respectively. The ancestral spectra are shown in Figure 3 and are similar to those in Figure 2. This suggests that high ts/tv ratios indicate not only excessive transitions due to postmortem cytosine deamination, but also transversion errors, possibly due to the high level of DNA damage in the samples [9]. Comparing the ancestral spectra of NE1 and NE5, which have ts/tv ratios of 10.53 and 1.72, respectively, NE5 shares exclusive ancestry with EUR, while NE1 demonstrates additional coancestries with AFR, EAS, and SAS and a small level of coancestry with AMR. Considering the geographical and temporal

proximity of the two samples, the ancestral spectrum of NE1 is probably misleading due to the quality issue with the sample. Accordingly, the results of other samples with high ts/tv ratios are also unlikely to be reliable.

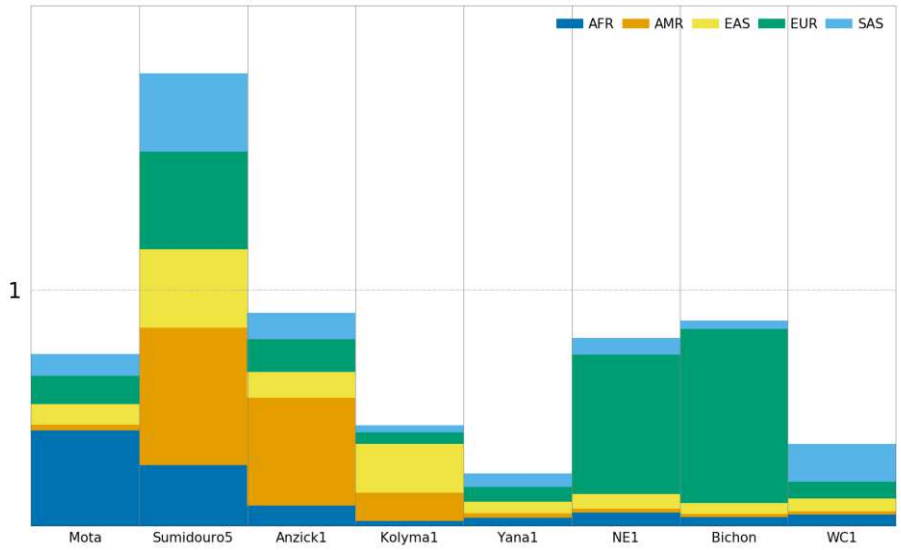

Figure 2. Ancestral spectra of the 8 aDNA with high ts/tv ratios; 5 reference populations.

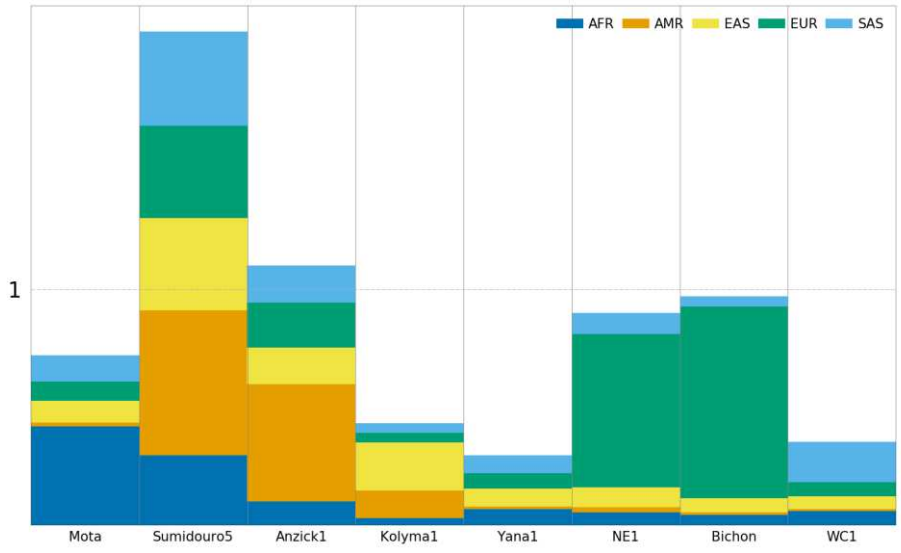

Figure 3. Ancestral spectra of the 8 aDNA with high ts/tv ratios using transversion SNPs only; 5 reference populations.

The ancestral spectra of the 16 modern DNA are presented in Figure 4. Not surprisingly, YRI, MND and DIN are largely coancestral with AFR. Because KAR and PIM are in the HGDP and belong to the AMR reference population, they share exclusive ancestries with AMR in this analysis. ESK, ITE, ULC, YKT and ORQ were obtained from East Asia and are primarily coancestral with EAS. In addition, they have a small level of coancestries with EUR, and their coancestries with AMR increase as their locations become closer to eastern Siberia. The results are consistent with the ancestral spectra of East Asians in the HGDP [5]. XIB is mostly of coancestry with EAS, while MNS largely shares ancestries with EAS and EUR. AUS and PAP revealed small total ancestral information, showing coancestries with AFR, EAS and SAS. Their ancestors were believed to have migrated from Eurasia approximately 50,000 BP after the divergence between the ancestors of EUR and EAS [16]. Some alleles specific to modern AFR populations might be carried in populations peopling South and East Asia at the time. This is also consistent with the ancestral spectra of other Oceanians in the HGDP [5]. Note that AUS has additional conancestry with EUR, possibly due to recent admixtures. As expected, IRU is mainly of coancestry with SAS and JHM shares ancestries with EAS and SAS.

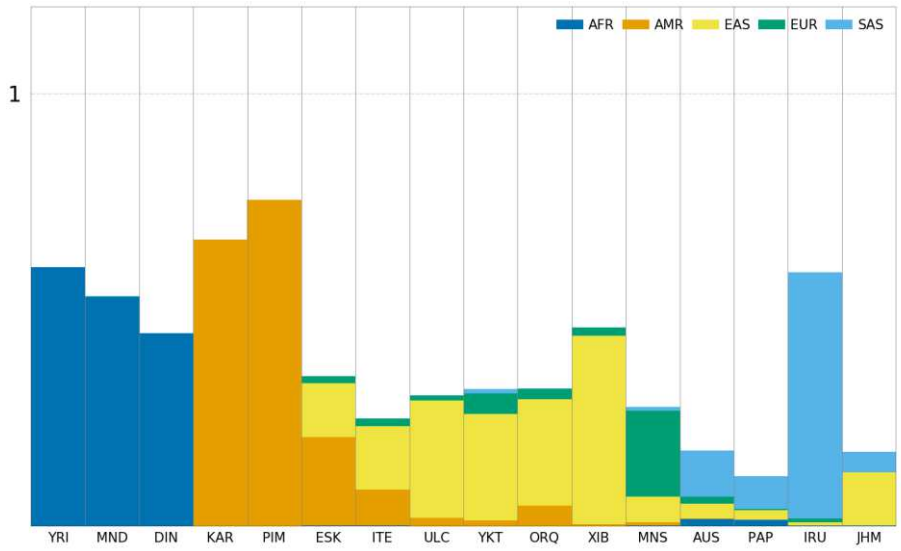

Figure 4. Ancestral spectra of the 16 modern DNA; 5 reference populations.

Of the 16 modern DNA, 15 were from the SGDP and were previously sequenced. With the genotype data from the SGDP, the results of the ancestral spectrum analysis are shown in Figure 5 and detailed in Supplementary Table S4. Comparing this figure with Figure 4, we can see that the absolute values of the ancestral information are much larger, while general patterns of ancestral components remain. This is because the genotype data used in Figure 4 were called by the genetic pipeline specially designed for studying aDNA, which is robust for calling genotypes of degraded aDNA samples. This approach is possibly at the expense of sensitivity to a certain extent, especially for calling genotypes of high-quality modern DNA samples.

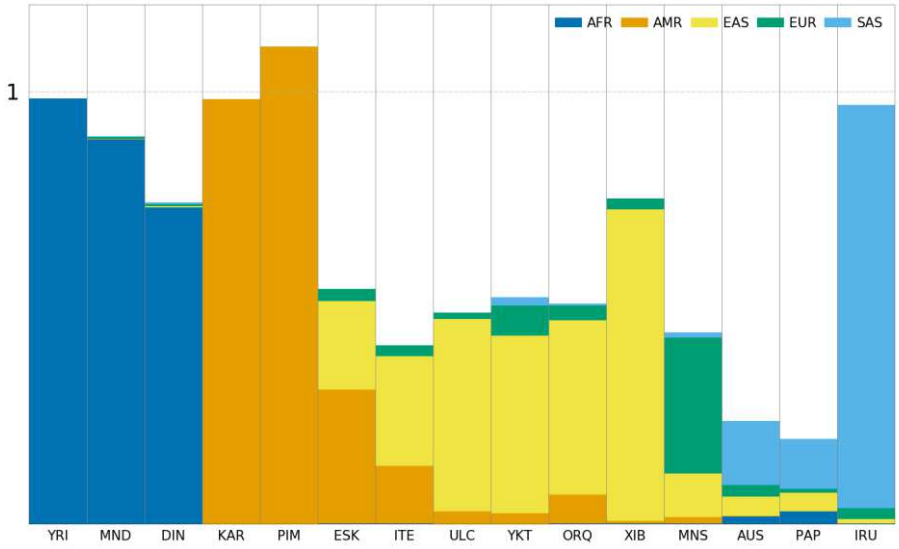

Figure 5. Ancestral spectra of the 15 modern DNA with SGDP genotype data; 5 reference populations.

Ancestral spectra with the 7 reference populations

Using the second panel of SNPs specific to the 7 reference populations in HGDP, the ancestral spectra of the 19 aDNA and the 16 modern DNA are presented in Supplementary Table S5. The results of the 11 aDNA with expected ts/tv ratios are shown in Figure 6. Because different reference populations were used, in particular for those from Eurasia, the results are slightly different from those obtained with the panel of 5 reference populations. For instance, KK1, which was sampled at Kotias Klde, Georgia, and dated to 9712 BP, shows almost exclusive coancestry with EUR in Figure 1 and a mixture of coancestries with populations from Europe, Central and South Asia and Middle East in Figure 6. Other aDNA that share ancestries with EUR also show various levels of coancestries with the three reference populations. We analyzed the 5 reference populations in the first panel by using the second panel of population-specific SNPs associated with the 7 reference populations, and the results are shown in Figure

7. EURs from the 1kGP indeed show coancestries with Europeans, Central and South Asians and Middle Easterns from the HGDP, and the average ancestral information is 0.53, 0.03 and 0.07, respectively.

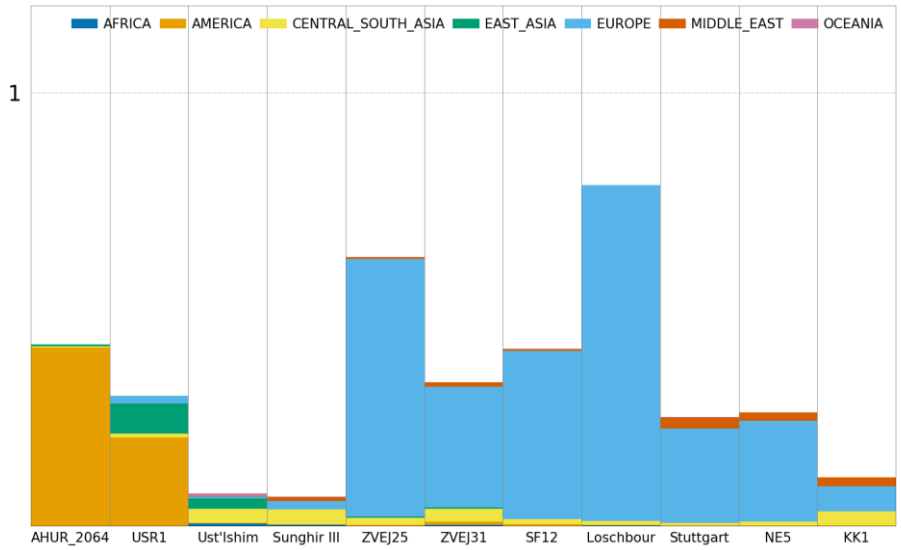

Figure 6. Ancestral spectra of the 11 aDNA with expected ts/tv ratios; 7 reference populations.

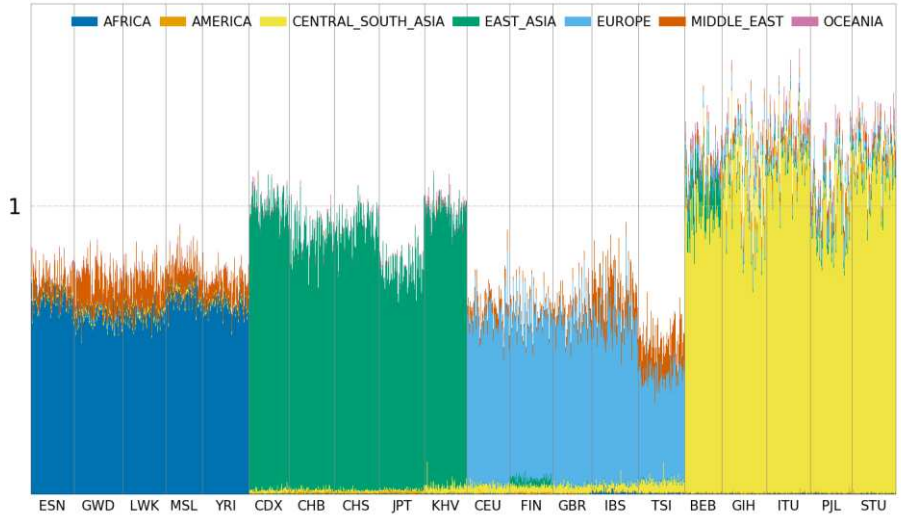

Figure 7. Ancestral spectra of the 5 reference populations in the first panel using the second panel of population-specific SNPs; 7 reference populations.

The ancestral spectra of the 16 modern DNA using the 7 reference populations are

displayed in Figure 8. YRI and MND are of exclusive coancestry with Africans; similarly, KAR and PIM with Americans; YKT, ORQ and XIB with East Asians; and PAP with Oceanians. This is because the 8 modern samples are also in the HGDP and belong to the corresponding reference populations. DIN, which shares ancestry with AFR in the analysis with the 5 reference populations, shows coancestries with populations in Africa, Central and South Asia and Middle East. This observation is consistent with the AFRs in the 1kGP revealing average ancestral information of 0.64, 0.01, and 0.10 with the three reference populations. As observed in the results of aDNA, ESK, ITE, ULC, MNS, AUS and IRU, which share ancestries with EUR in the analysis using the 5 reference populations, have a small degree of coancestries with populations from Central and South Asia and Middle East in the analysis with the 7 reference populations. Note that IRU and JHM show a small level of coancestries with Oceanian populations. This is because they have coancestries with SAS in the analysis with the 5 reference populations, and SASs in the 1kGP share ancestral information of 0.02 with Oceanians, on average.

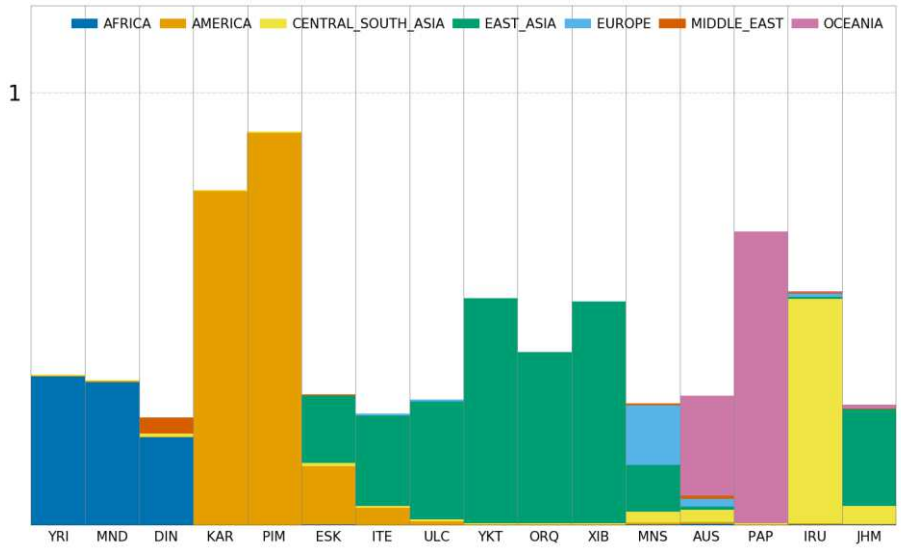

Figure 8. Ancestral spectra of the 16 modern DNA; 7 reference populations.

The 15 modern DNA from the SGDP were also analyzed with the 7 reference populations and genotype data from the SGDP [15], and the results are shown in Figure 9 and Supplementary Table S6. The general patterns of ancestral spectra are approximately the same as those shown in Figure 8; however, the absolute values of the ancestral information with SGDP genotypes are much larger than those with the genotypes called together with aDNA.

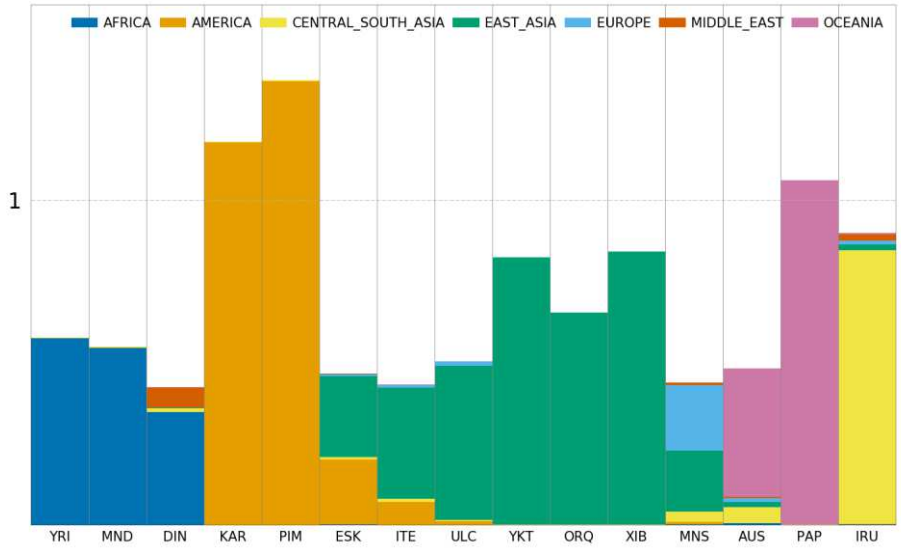

Figure 9. Ancestral spectra of the 15 modern DNA with SGDP genotype data; 7 reference populations.

Discussion

The ancestral spectrum of an individual comprises a set of ancestral information, each of which corresponds to one of the reference populations. Ancestral information is the best linear unbiased estimate of the ancestry proportion if the associated reference population approximates an ancestral population well [5]. When reference populations are modern populations, this is possible for study samples that are recently admixed. In

this analysis, our reference populations, however, are limited in approximating the ancestral populations of the aDNA. Hence, ancestral information simply measures coancestries between the aDNA and the modern reference populations, which can be interpreted in reverse. For instance, Loschbour shows a high level of coancestry with EUR and no coancestries with other reference populations in the first panel. This result suggests that Loschbour belongs to the ancestral population of modern EURs. It also confirms that at the time Loschbour lived, which was approximately 8055 BP, the ancestors of modern AFR, AMR, EAS, EUR and SAS populations had already separated.

Since we used modern populations as the reference populations in this study, SNPs specific to the reference populations are largely due to mutations that occurred after the separations of the ancestral populations of the reference populations; thus, these mutations are often recent. As a result, aDNA from the far past is unlikely to carry alleles that are recent and, therefore, demonstrate little ancestral information. For example, Ust'Ishim and Sunghir III are from 45,000 and 34,093 BP with total ancestral information of 0.07 and 0.06, respectively, in the analysis with the 5 reference populations and 0.08 and 0.07, respectively, in the analysis with the 7 reference populations. On the other hand, some variants specific to modern populations may be ancient. These variants could have been shared by ancestors of modern populations at the time of or before the separations and became specific to one population later due to genetic drift or different demographic events such as bottlenecks. For instance, AFR ancestral information of PAP in Figure 4 suggests that her ancestors may carry some

alleles specific to modern AFR populations at the time of their migration from Eurasia. This is also likely to be true for other populations inhabiting Eurasia at the same time. Nevertheless, modern EAS, EUR and SAS populations no longer carry these alleles.

Technical issues may confound the estimation of ancestral spectra [5]. As shown in Figures 4 and 5, the ancestral information of the 15 modern DNA with the genotype data from SGDP is much larger than that with the genotype data called together with the aDNA. A similar difference was observed between the ancestral spectra of the 1kGP samples with genotype data released in Phase 3 and those with genotype data obtained by deep sequencing technology [5]. In both cases, the types of ancestral information and their proportions in the same sample remain approximate. Therefore, caution must be taken when comparing the absolute values of ancestral information between samples obtained using different sequencing technologies or bioinformatics pipelines. Furthermore, unlike in the study of modern samples in which DNA quality is typically high, the quality of aDNA samples varies considerably. Thus, absolute values of ancestral information of aDNA cannot be compared meaningfully, even if the samples are assessed via standardized experimental and bioinformatics protocols.

We used two panels of reference populations in this study. Populations from Africa, America, East Asia, South Asia and Europe were represented in both panels, while the estimated ancestral information based on the two panels cannot be interpreted interchangeably. In terms of geographical distributions, populations from Southern Africa, Northeastern Asia and Eastern Europe were included in the second panel of reference populations but not in the first. Thus, populations from Africa, East Asia and

Europe in the two panels represent slightly different populations. Similarly, populations from Central Asia were aggregated with populations from South Asia in the second panel, but not in the first. Regarding indigenous AMRs, for which the same populations are present in both panels, the estimated ancestral information and its interpretation are not identical because variants specific to indigenous AMRs must be polymorphic in the AMRs and absent in the other reference populations. Note that the second panel includes populations from Northeastern Asia known to share some ancestry with indigenous AMRs [16]. SNPs specific to indigenous AMRs in the first panel of reference populations may not be specific to indigenous AMRs in the second panel of reference populations. Therefore, ancestral spectra must be interpreted together with the underlying reference populations. In general, the second panel comprises more reference populations and allows finer resolution for inferring the coancestries between the study samples and global populations. Because the sample sizes of the reference populations in the second panel are much smaller, the identified population-specific SNPs are subject to more misclassification errors [5].

Of the 19 aDNA analyzed in this study, 8 have excessively high ts/tv ratios and demonstrate various levels of coancestries with almost all reference populations, which is unlikely a true result. The ancestral spectra of the other 11 aDNA with expected ts/tv ratios appear to be reasonable. Whether aDNA have high ts/tv ratios is strongly associated with whether the samples were subject to UDG treatment [9]. These findings support the efficacy of UDG treatment for removing postmortem DNA damage that typically causes excessive transitions at genomic positions flanking read starts or ends

[3]. However, in our analysis of the 8 aDNA with transversions only, abnormal patterns of the ancestral spectra remain present (see Figures 2 and 3). This finding suggests that UDG treatment is highly desirable in the wet-laboratory processing of aDNA, and not only transition errors but also transversion errors may exist around the starts or ends of the sequencing reads. Regarding sequencing data generated without UDG treatment, customized bioinformatics pipelines that remove such errors could be particularly helpful.

Declarations

Ethical approval and consent to participate

Not applicable. No human or animal experiments were conducted in this study. All analyses were based on datasets publically available.

Consent for publication

Not applicable.

Ethical Guidelines

This study was carried out in accordance with relevant guidelines and regulations.

Availability of data and materials

Genotype datasets of the 19 ancient and 16 modern high-coverage shotgun human genomes are available at https://doi.org/10.6084/m9.figshare.c.5183474. Genotype datasets of the 1000 Genomes Project are available at http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/. Genotype datasets of the Human Genome Diversity Project are available at https://ngs.sanger.ac.uk//production/hgdp/hgdp_wgs.20190516/. Genotype datasets of

the Simons Genome Diversity Project are available at https://reichdata.hms.harvard.edu/pub/datasets/sgdp/. Software Ancestral Spectral Analyzer (ASA) is available at https://github.com/eat1000/ASA.

Funding

This work was supported by the national Thousand Youth Talents Plan.

Competing interest

The author declares that he has no competing interests.

References

1. Bergström A, Stringer C, Hajdinjak M, Scerri EML, Skoglund P. Origins of modern human ancestry. Nature. 2021;590(7845):229-237.
2. Liu Y, Mao X, Krause J, Fu Q. Insights into human history from the first decade of ancient human genomics. Science. 2021;373(6562):1479-1484.
3. Orlando L, Allaby R, Skoglund P, Der Sarkissian C, Stockhammer PW, Ávila-Arcos MC, Fu Q, Krause J, Willerslev E, Stone AC, Warinner C. Ancient DNA analysis. Nat Rev Methods Primers. 2021;1:14.
4. Auton A, Brooks LD, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.
5. Shi G, Kuang Q. Ancestral spectrum analysis with population-specific variants. Front Genet. 2021;12:724638.
6. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006;2(12):e190.

7. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38(8):904-9.
8. Ma J, Amos CI. Principal components analysis of population admixture. PLoS One. 2012;7(7):e40115.
9. Maisano Delser P, Jones ER, Hovhannisyan A, Cassidy L, Pinhasi R, Manica A. A curated dataset of modern and ancient high-coverage shotgun human genomes. Sci Data. 2021;8(1):202.
10. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. Genetic structure of human populations. Science. 2002;298(5602):2381-5.
11. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.
12. Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A. Bayesian inference of ancient human demography from individual genome sequences. Nat Genet. 2011;43(10):1031-4.
13. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308-11.
14. Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, Chen Y, Felkel S, Hallast P, Kamm J, Blanché H, Deleuze JF, Cann H, Mallick S, Reich D,

Sandhu MS, Skoglund P, Scally A, Xue Y, Durbin R, Tyler-Smith C. Insights into human genetic variation and population history from 929 diverse genomes. Science. 2020;367(6484):eaay5012.

15. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, Skoglund P, Lazaridis I, Sankararaman S, Fu Q, Rohland N, Renaud G, Erlich Y, Willems T, Gallo C, Spence JP, Song YS, Poletti G, Balloux F, van Driem G, de Knijff P, Romero IG, Jha AR, Behar DM, Bravi CM, Capelli C, Hervig T, Moreno-Estrada A, Posukh OL, Balanovska E, Balanovsky O, Karachanak-Yankova S, Sahakyan H, Toncheva D, Yepiskoposyan L, Tyler-Smith C, Xue Y, Abdullah MS, Ruiz-Linares A, Beall CM, Di Rienzo A, Jeong C, Starikovskaya EB, Metspalu E, Parik J, Villems R, Henn BM, Hodoglugil U, Mahley R, Sajantila A, Stamatoyannopoulos G, Wee JT, Khusainova R, Khusnutdinova E, Litvinov S, Ayodo G, Comas D, Hammer MF, Kivisild T, Klitz W, Winkler CA, Labuda D, Bamshad M, Jorde LB, Tishkoff SA, Watkins WS, Metspalu M, Dryomov S, Sukernik R, Singh L, Thangaraj K, Pääbo S, Kelso J, Patterson N, Reich D. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature. 2016;538(7624):201-206.

16. Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E. Tracing the peopling of the world through genomics. Nature. 2017;541(7637):302-310.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTables.xlsx