A SNP panel for co-analysis of capture and shotgun ancient DNA data

Romain Fournier[1,*], Alice Pearson Fulton[1] and David Reich[1,2,3,4,*]

1 Department of Human Evolutionary Biology, Harvard University, 02138, Cambridge, MA, USA
2 Department of Genetics, Harvard Medical School, 02115, Boston, MA, USA
3 Broad Institute of MIT and Harvard, 02142, Cambridge, MA, USA
4 Howard Hughes Medical Institute, 02138, Boston, MA, USA
* corresponding authors: romainfournier@fas.harvard.edu and reich@genetics.med.harvard.edu

Abstract

Advances in technology have decreased the cost of generating genetic variation data from ancient people, resulting in exponentially increasing numbers of individuals with whole genome data. However, each technology comes with platform-specific biases, limiting co-analyzability of individuals sequenced with different technologies as well as joint analysis of modern and ancient individuals. We present a method to identify single nucleotide polymorphisms (SNPs) with minimal technology-specific bias. Leveraging data from over 17,000 ancient individuals, we apply this method to identify a set of more than a million SNPs that we call the "Compatibility" panel, and which has been effectively assayed in a large fraction of ancient human DNA experiments published to date. We also identify a subset of these SNPs, the "Compatibility-HO" panel, which further restricts to positions that have been assayed in more than ten thousand modern individuals from more than a thousand diverse populations using the Affymetrix Human Origins (HO) genotyping array. The Compatibility panel reduces spurious Z-scores due to different sequencing platforms by nearly an order of magnitude, while retaining around 70-90% of statistical power for *f*-statistic analysis.

Introduction

The number of ancient individuals with genome-wide ancient DNA data has expanded exponentially over the past decade and a half [1]. These datasets, which crossed the threshold of 10,000 humans with genome-wide data in 2024, have offered granular insights into ancient population history and cultural practices, and have also enabled the detection of natural selection through time-series analysis [2,3]. The increase in sample size has been turbo-charged by in-solution enrichment for targeted subsets of more than a million single nucleotide polymorphisms (SNPs) informative about history and biology, which reduces the required amount of DNA sequencing by one to two orders of magnitude per individual. The main in-solution SNP enrichment methods used in the field are the "Agilent 1240k reagent" (1240k or AG), accounting for about 65% of all published individuals, and the "Twist Ancient DNA" assay (Twist or TW), accounting for 5% of published individuals. While they have greatly decreased the cost of ancient DNA analysis, in-solution enrichment methods have the drawback that they are more prone to biases toward one allele or another at analyzed positions, and yield more variable coverage across targeted SNPs than the approach of shotgun sequencing without an enrichment step [4]. These differences have meant that in co-analysis of data produced using different technological platforms, individuals with data produced using the same platform often share alleles at an elevated rate with each other, leading to false-positive inferences of shared ancestry. As a result of these biases, many researchers have restricted many of their analyses to individuals sequenced with the same technology. But since key individuals have been analyzed using different technologies, such restriction significantly limits the scientific questions that can be asked. Datasets from ancient individuals also tend to share alleles at a higher rate than with modern individuals, not due to more shared ancestry, but instead due

to the propensity of ancient DNA methods to better capture some alleles than others, making it difficult to co-analyze modern and ancient individuals.

Two approaches have been suggested for combining data sequenced using different technologies. The first consists of using imputed data [5]. Imputation works by copying haplotype segments from a reference panel, and then selecting the path that best represents the sequencing data. By copying segments from an external panel, this method can leverage neighboring sites to correct for local biases that are due to sequencing. This method has been effectively used to increase the power to discover signals of directive selection in a data set comprising thousands of individuals sequenced with various sequencing platforms [3]. However, the approach is restricted to individuals with sufficient sequencing depth (>0.5x genome-wide coverage) and requires the individual's haplotype variation to be well represented in the reference panel [5]. The second approach focuses on selecting a subset of SNPs that are unbiased across technologies. By selecting SNPs where the ratio of reads mapping to the reference allele at heterozygous sites differed by less than 4% across technologies, one study identified a panel of SNPs comprising 42% of the original 1240k [4], increasing co-analyzability but greatly reducing the number of SNPs and power of downstream analyses. The previously reported SNP panel was also not optimized to remove SNPs that show bias relative to datasets of modern individuals, making co-analysis of modern and ancient individuals with this panel unreliable [6].

We introduce a method that improves upon the previous approach [4]. First, we detect biased SNPs using an increased sample size, leveraging over 17,000 individuals with >1x coverage, compared to fewer than 500 for the previously reported co-analysis SNP panel. Second, we use a different statistical test that accounts for the variance in read counts, and does not require the same individuals to be sequenced on different technologies. Third, the statistical test generalizes to an arbitrary number of analysis platforms, and we leverage this to report a "Compatibility" SNP panel that allows unbiased co-analysis of ancient and modern DNA data generated using 1240k, Twist, low-coverage shotgun, or diploid sequencing platforms (Supplementary Table 9). We also identify a subset of these SNPs that are also targeted by the Affymetrix Human Origins (HO) genotyping array, the "Compatibility-HO" panel, and that we verify as producing co-analyzable data between modern and ancient data (Supplementary Table 10).

The Compatibility panel is a subset of a previously described "2M" panel consisting of the union of positions targeted by the 1240k and Twist assays, along with SNPs that were effectively captured using either of these assays due to being within 50 base pairs of targeted positions [4]. It comprises 66% of the autosomal SNPs, to which we added the unfiltered 63,994 positions on the X chromosome and 199,881 positions on the Y chromosome from the 2M panel. We tested the new panel using f_4 statistics on sets of two individuals sequenced with different pairs of technologies, and found that technology-related inflation in Z-scores was reduced by up to an order of magnitude, greatly facilitating reliable co-analysis of heterogeneous data. We also assessed the loss of power resulting from the reduced number of SNPs, and found that Z-scores decreased by 10% on average for ancient individuals sequenced at coverage above 0.8x.

Results

Annotating SNPs based on the strength of evidence for incompatibility across platforms

To identify SNPs for the Compatibility panel, we started with autosomal SNPs from the 2M SNP set introduced by Rohland et al.[4], and then removed SNPs based on a series of criteria. First, we removed all SNPs that were not polymorphic in the 1000 Genomes Project (1KGP) [7] or GnomAD [8] datasets. Next, we divided the individuals with sufficient coverage from our internal database according to their sequencing technology (Table 1). For each sequencing technology, we calculated the ratio of the coverage of each SNP to the genome-wide average. We discarded SNPs whose ratio was smaller than 5% in each of the AG, TW, and SG datasets. Then, we discarded SNPs with significant differences in their reference bias across technologies. To achieve this, we counted, for each SNP, the sequences mapping to the reference and alternate alleles in individuals confidently imputed as heterozygous (genotype posterior >0.9). We discarded SNPs based on the Z-score under the null hypothesis that the proportion of reads mapping to each allele is independent of the sequencing technology (see Methods). Additionally, we annotated the SNPs targeted by the Affymetrix Human Origins array based on their relative mismatch rate between sequencing and genotyping data (see Methods).

Technology	Number of individuals	Minimum coverage on 1240k	Included in Min SNP Coverage Criteria	Present in validation set
1240k (Boston)	5697	1x	Yes	Yes
1240k (Max Planck Institute)	1371	1x	No	No
SG	2054	1x	Yes	Yes
TW	7154	2x	Yes	Yes
Ancient Diploid Genomes	74	1x	No	No
Modern Diploid Shotgun Genomes	1416	1x	No	No

Table 1. Summary of technologies and sample sizes used in constructing and validating the Compatibility panel. *For each technology, we report the number of individuals analyzed, minimum required coverage for inclusion, whether the group was used to define SNP coverage criteria, and whether it was included in the validation set. Summary statistics about the number of reads mapping to reference and alternate alleles at each locus are in supplementary tables 2-7.*

We iteratively removed SNPs with the highest bias from the 2M set, and then quantified the remaining bias by computing an f_4 statistic on held-out individuals assessed using two different technologies:

$$\text{bias} = f_4(\text{ind}_1\text{tech}_1, \text{ind}_1\text{tech}_2; \text{ind}_2\text{tech}_1, \text{ind}_2\text{tech}_2) \qquad \text{Eq. 1}$$

We stopped removing SNPs when the magnitude of this f_4 statistic fell to 30% of that of the full panel, corresponding to the asymptotic low point of the AG-SG curve (Figure 1a) (the AG-SG comparison showed the highest bias statistic among the pairs of technologies. The bias after discarding 30% of the SNPs ($f_4(AG, SG) = 2.5 \times 10^{-4}$) is lower than the original bias between TW and SG ($f_4(TW, SG) = 3.4 \times 10^{-4}$), a pair of technologies considered to be nearly fully co-analyzable according to the literature [4,9].

We identified the SNPs in the Compatibility-HO panel by starting with the Compatibility panel, then further restricting to SNPs genotyped on the HO array, and finally filtering based on the f_4 statistic between more than one thousand modern individuals for which we had both shotgun sequencing data and HO genotyping data. We computed the bias using Eq. 1 with individuals from both the HGDP and 1KGP datasets (Figure 1b) (the HGDP individuals were not used for training, so for the HGDP individuals the validation is out-of-sample). We stopped removing SNPs when the magnitude of this bias statistic in the HGDP data fell below 20% of that of the full panel, and the magnitude of the bias statistic in the 1KGP data fell to zero (Figure 1).

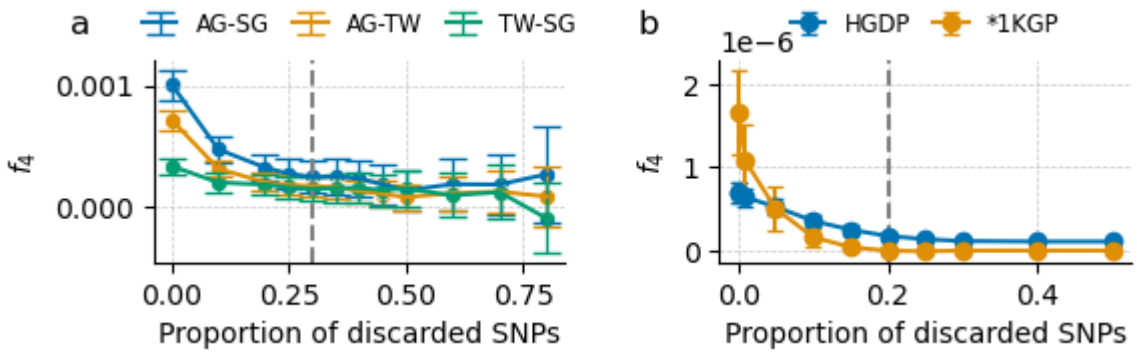

Figure 1: Selection of a threshold for inclusion in the panels (a) *Bias, as defined by Eq. 1, for held-out individuals sequenced using at least two technologies among AG, TW and SG. The average value of the f_4 statistics and 1.96 x s.e.m. are displayed.* ***(b)*** *Same quantity for individuals sequenced and genotyped in the Affymetrix Human Origins Array. The individuals from 1KGP were used for training, while the HGDP ones were not providing true out-of-sample validation.*

Properties of the Compatibility panel

We compared the original and unbiased SNP panels with regard to a series of metrics relevant to population genetic analysis.

We first calculated the fraction of SNPs intersecting the classic 1240k set, and noticed a significant overrepresentation of discarded SNPs in this set (Supplementary Table 1, p-value $p < 10^{-5}$). This plausibly reflects the fact that these were SNPs for which we had particularly large amounts of data providing increased sensitivity to detect differences in bias across platforms. Second, we compared the minor allele frequency (MAF) spectra of both panels using individuals labeled with African (Figure 2a) or European (Figure 2b) ancestry. In both instances, we observed an overrepresentation of SNPs with

large MAF (above ~15%) among the discarded SNPs, which is plausibly due to these being ones where it is easier to detect bias. Third and finally, we compared the distribution of mutation profiles, noticing a depletion of transitions in the discarded SNPs (Figure 2c). This might reflect lower imputation genotype posterior due to postmortem degradation, reducing the number of individuals passing the filtering criteria for these loci.

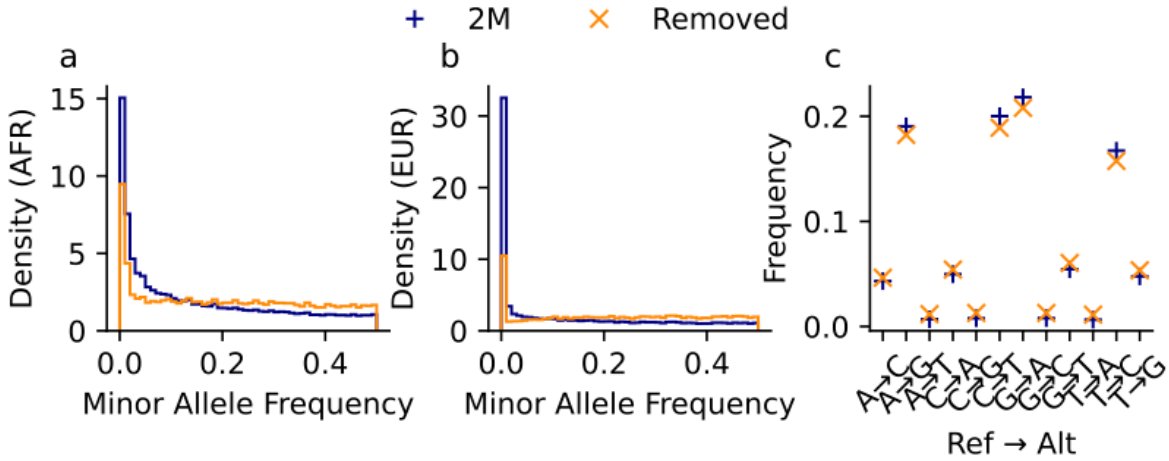

Figure 2: Properties of the Compatibility Panel SNPs. a-b) *Minor allele frequency spectrum in the 1KGP AFR (a) and EUR (b) continental labels. c) Distribution of the mutation profiles in the 2M and Compatibility panels.*

Negligible technical bias in practical settings

We evaluated the bias reduction provided by our SNP panels in realistic scenarios. To achieve this, we calculated three f_4(pop1, pop2; pop3, pop4) statistics which are not expected to differ significantly from 0 given the relationships among these groups. In each of the three f_4 tests, populations 1 and 3 included only 1240k data, while populations 2 and 4 consisted of data sequenced using the same alternative technology (Twist, Shotgun, Diploid Modern, or HO) (Table 1). This setup corresponds to a worst-case scenario, where any technology-specific biases can be expected to lead to a particularly significant deviation from 0. Indeed, all f_4 tests produced results significantly different from 0 in this setting (p-value $< 10^{-17}$, Z-score>8.7) (Table 2). Using the recommended SNP panel to compute these f_4 statistics showed up to an order of magnitude reduction in the Z-scores, with the remaining Z-scores reaching values below what is typically reported in such analyses (Z<3), except for our test with the 1KGP data set, which contained the largest sample size. In real analyses, a more uniform distribution of the sequencing technologies across the populations would decrease the bias. Overall, these results suggest that our proposed SNP panels effectively reduce biases.

Population 1	Population 2	Population 3	Population 4	Z (original)	Z (new)
England Neolithic TW (n=16)	England Neolithic AG (n=25)	Yamnaya TW (n=25)	Yamnaya AG (n=25)	10.8	2.8
England Medieval AG (n=65)	England Medieval SG (n=8)	Denmark Medieval AG (n=14)	Viking SG (n=27)	12.2	2.2
YRI (n=101)	Kenya Makwasinyi AG (n=10)	GBR (n=90)	England Medieval AG (n=65)	11.3	3.2
Kenya Makwasinyi AG (n=10)	Kikuyu HO (n=4)	Belgium Medieval AG (n=36)	French HO (n=61)	8.7	1.4

Table 2: f_4 ***statistics obtained using the original and new SNP panels.*** *Data from different technologies were used to calculate* f_4 *statistics. The first three rows were computed using the Compatibility panel, and the last row the Compatibility-HO panel. Individuals from 1KGP (GBR and YRI) were used during the training phase, whereas the other ones were held out.*

Minimal reduction in statistical power from use of the Compatibility panel

We quantified the reduction in statistical power resulting from the lower number of SNPs in the Compatibility panel. To perform this computation, we re-examined a dataset of 291 individuals assigned to different groups across Iron Age Britain [10]. Following the original study, we first tested for differences in ancestry between pairs of groups by calculating statistics of the form f_4(group 1, group 2; Turkey Neolithic, Old Steppe). We observed a high correlation between the Z-scores obtained using the original SNP panel and the new one ($R^2 = 0.94$), with a regression coefficient of 0.89 (Figure 3a).

We analyzed statistics of the same form, but replacing British Iron Age groups with British Iron Age individuals, to assess how coverage impacts the standard errors of the f_4 statistics. To do so, we calculated f_4(ind 1, ind 2; Turkey Neolithic, Old Steppe) for pairs of individuals belonging to pairs of groups that were detected as being significantly different in their ancestry in the original study (reported absolute Z-scores above 3). For each pair, we recorded the number of overlapping SNPs, as well as the Z-scores obtained when calculating the f_4 statistics with the new and original panels. We then split all results according to the number of overlapping sites, creating 10 sets of the same size. Within each group, we estimated the regression coefficient of the Z-scores calculated on both panels, as we did in the previous analysis (see Methods). The regression coefficient (Figure 3b) stays at 0.9 for individuals with coverage > 0.8× before progressively dropping to 0.7. The greater loss of power for lower coverage individuals makes intuitive sense, since for such individuals there is expected to be little linkage disequilibrium among SNPs for which data are available, and hence nearby SNPs are not filling in information gaps as is the case for higher coverage individuals.

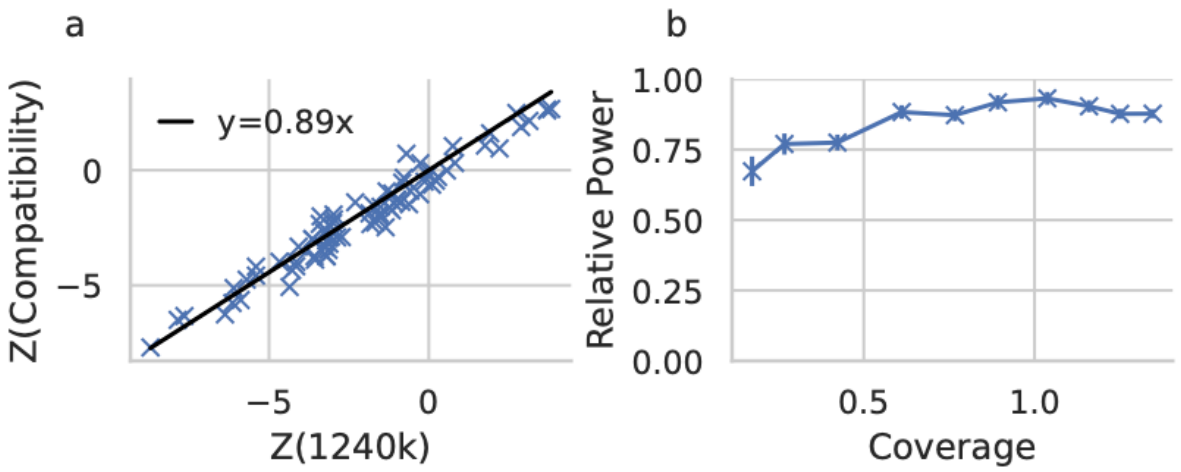

Figure 3: Power reduction in the new SNP set. (a) *Differences in Z-scores obtained using the original SNP set (1240k) and the Compatibility set. The regression line y=0.89 is shown to illustrate the reduction in power.* ***(b)*** *Same analysis but stratifying the data by coverage. The relative power corresponds to the regression coefficient fitted on Z(Compatibility) = a Z(1240k). Error bars correspond to the 95% CI around the parameter a, assuming independent Z-scores (Methods).*

Discussion

By reducing biases across technologies, our curated SNP panels can be used for multiple downstream analyses. One application is preserving compatibility with data generated using earlier sequencing technologies. As a result, new individuals can be analyzed without needing to match the platforms used in previous studies. Second, the panels enable joint modeling of ancient and modern individuals in f_4-based methods, such as qpAdm and qpWave [11]. Finally, methods beyond f_4 statistics may benefit from the curated panels. This is particularly relevant for frequency-based approaches and principal component analyses, where platform-specific biases can lead to spurious results.
The power to detect technology-specific bias is directly linked to the number of heterozygous individuals sequenced at each locus. As a result, our method is underpowered to detect problematic SNPs that have minor allele frequencies under 10%, or that are only present in underrepresented ancestries. An important caveat is that the new SNP panel, despite significantly reducing bias, does not fully eliminate it. Therefore, we recommend that analysts be cautious when interpreting Z-scores that are close to the rejection threshold. Finally, the compatibility panel has non-homogeneous coverage across technologies. For some analyses, it might be recommended to further impose a coverage threshold on the remaining SNPs.

As larger and more diverse datasets become available, we expect that future updates of this panel could further reduce biases. We also release a script and current read counts that researchers can use to build their own compatibility panel.

Methods

Removing SNPs based on differences in reference biases

Our analysis relied on imputed genotypes as in [3]. For each SNP in the 2M SNP panel [4] that is polymorphic in the 1KGP or Gnomad datasets, we retained individuals imputed as heterozygous with a genotype posterior probability above 90%. We then built a contingency table with sequencing technologies as rows and the number of reads mapping to the reference and alternate alleles as columns. We performed a chi-squared test of independence on these tables and stored the corresponding p-values for each SNP. We discarded SNPs with the lowest p-values.

Removing SNPs for the Compatibility-HO panel

We leveraged 1028 individuals with both sequencing and genotyping data [7,12–19]. For each SNP, we calculated the relative number of mismatches over the total number of heterozygous individuals at that site. We discarded SNPs according to this metric.

Calculating f_4 statistics

We calculated f_4 statistics using Admixtools2 [20] with the `maxmiss` parameter set to 0.

Power Analysis

We analyzed the same set of individuals and their corresponding population labels as in Patterson et al. 2022 [10] (Supplementary Table 8). We fit a linear regression without an intercept to the resulting Z-scores to estimate the power reduction. For the coverage analysis, we proceeded in the same way but this time using pairs of individuals belonging to the different groups with large Z-scores (>3). We separated the pairs of individuals according to the number of overlapping sites. If p denotes the proportion of overlapping sites on the 1240k panel, and c denote the coverage, we have:

$$p \approx (1 - e^{-c})^2$$

We inverted this relation to obtain c. This equation assumes that the probability of a SNP being missing is independent of other reads and uniform, which means that the actual value of c is likely to be lower. In other words, the power is likely to remain high at lower coverages than the reported 0.8x.

We fitted the regression line assuming all Z-scores are independent. This means that the confidence intervals provided around the slope are tighter than what is expected if points were truly independent.

Editing

We wrote a rough draft of each paragraph, and once completed, used Large Language Models to edit at the sentence level. We typically asked for three versions of some sentences. We then selected the version we preferred, or we mixed different elements of different answers. This tool was only used to edit the style of this manuscript, not its structure nor its content. We then carried out multiple additional rounds of editing (not assisted by these tools).

Acknowledgments

We thank Shop Mallick, Bárbara Sousa da Mota, and Nick Patterson for valuable discussions. We thank the bioinformatics team in our laboratory for making available the datasets we used for identifying SNPs susceptible to bias. This work was supported by John Templeton Foundation grant 61220; by National Institutes of Health grant HG012287; and by the Howard Hughes Medical Institute.

Bibliography

1. Mallick S, Micco A, Mah M, Ringbauer H, Lazaridis I, Olalde I, et al. The Allen Ancient DNA Resource (AADR) a curated compendium of ancient human genomes. Sci Data. 2024;11: 182.

2. Mathieson I, Terhorst J. Direct detection of natural selection in Bronze Age Britain. Genome Res. 2022;32: 2057–2067.

3. Akbari A, Barton AR, Gazal S, Li Z, Kariminejad M, Perry A, et al. Pervasive findings of directional selection realize the promise of ancient DNA to elucidate human adaptation. bioRxiv. 2024. doi:10.1101/2024.09.14.613021

4. Rohland N, Mallick S, Mah M, Maier R, Patterson N, Reich D. Three assays for in-solution enrichment of ancient human DNA at more than a million SNPs. Genome Res. 2022;32: 2068–2078.

5. Sousa da Mota B, Rubinacci S, Cruz Dávalos DI, G. Amorim CE, Sikora M, Johannsen NN, et al. Imputation of ancient human genomes. Nature Communications. 2023;14: 1–17.

6. Pickrell JK, Patterson N, Barbieri C, Berthold F, Gerlach L, Güldemann T, et al. The genetic prehistory of southern Africa. Nat Commun. 2012;3: 1143.

7. A global reference for human genetic variation. Nature. 2015;526: 68–74.

8. Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, et al. A genomic mutational constraint map using variation in 76,156 human genomes. Nature. 2023;625: 92–100.

9. Davidson R, Roca-Rada X, Ravishankar S, Taufik L, Haarkötter C, Collen E, et al. Optimized in-solution enrichment of over a million ancient human SNPs. Genome Biology. 2025;26: 1–20.

10. Patterson N, Isakov M, Booth T, Büster L, Fischer C-E, Olalde I, et al. Large-scale migration into Britain during the Middle to Late Bronze Age. Nature. 2022;601: 588–594.

11. Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, et al. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature. 2015;522: 207–211.

12. Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature. 2014;513: 409–413.

13. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature. 2016;538: 201–206.

14. Skoglund P, Posth C, Sirak K, Spriggs M, Valentin F, Bedford S, et al. Genomic insights into the peopling of the Southwest Pacific. Nature. 2016;538: 510–513.

15. Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature. 2014;505: 43–49.

16. Fan S, Kelly DE, Beltrame MH, Hansen MEB, Mallick S, Ranciaro A, et al. African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations. Genome Biol. 2019;20: 82.

17. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient admixture in human history. Genetics. 2012;192: 1065–1093.

18. Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, et al. Insights into human genetic variation and population history from 929 diverse genomes. Science. 2020;367. doi:10.1126/science.aay5012

19. Nakatsuka N, Moorjani P, Rai N, Sarkar B, Tandon A, Patterson N, et al. The promise of discovering population-specific disease-associated genes in South Asia. Nat Genet. 2017;49: 1403–1407.

20. Maier R, Flegontov P, Flegontova O, Işıldak U, Changmai P, Reich D. On the limits of fitting complex models of population history to f-statistics. Elife. 2023;12. doi:10.7554/eLife.85492