scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A curated dataset of modern and ancient high-coverage shotgun human genomes

Pierpaolo Maisano Delser[1,2]✉, Eppie R. Jones[1,3], Anahit Hovhannisyan[4], Lara Cassidy[2], Ron Pinhasi[5] & Andrea Manica[1]✉

Over the last few years, genome-wide data for a large number of ancient human samples have been collected. Whilst datasets of captured SNPs have been collated, high coverage shotgun genomes (which are relatively few but allow certain types of analyses not possible with ascertained captured SNPs) have to be reprocessed by individual groups from raw reads. This task is computationally intensive. Here, we release a dataset including 35 whole-genome sequenced samples, previously published and distributed worldwide, together with the genetic pipeline used to process them. The dataset contains 72,041,355 sites called across 19 ancient and 16 modern individuals and includes sequence data from four previously published ancient samples which we sequenced to higher coverage (10–18x). Such a resource will allow researchers to analyse their new samples with the same genetic pipeline and directly compare them to the reference dataset without re-processing published samples. Moreover, this dataset can be easily expanded to increase the sample distribution both across time and space.

Background & Summary

The number of ancient humans with genome-wide data available has increased from less than five a decade ago to more than 3,000 thanks to advancements in extraction and sequencing methods for ancient DNA (aDNA)[1]. However, there are just a few high-quality (coverage >10x) shotgun whole-genome sequenced ancient samples[2]. While genetic pipelines have been previously published[3–6], combining data processed with different approaches is hard and time consuming. Therefore, researchers have to download raw reads of published samples and reprocess them to create a dataset to compare their new samples against without pipeline-associated biases. This problem is less pronounced for modern DNA samples as the higher quality of DNA and sequencing coverage partially reduce the biases introduced by the usage of different bioinformatic tools.

Panels including shotgun data for modern samples distributed worldwide have been previously published, such as the Simons Genome Diversity Program[7], 1000 Genome Project[8] and Human Genome Diversity Project (HGDP-CEPH panel)[9]. However, the same concept has not yet been applied to ancient samples or a mix of modern and ancient samples. This study aims to start filling this gap by creating a dataset including both modern and ancient samples distributed across all continents. Therefore, we fully reprocessed 15 high-quality shotgun sequenced ancient samples downloaded from the literature, generated additional new data for previously published 4 ancient samples and merged them with 16 modern samples. The final dataset includes 35 individuals and researchers can use it to quickly compare their new samples against a set of individuals distributed across time and space (Fig. 1). Moreover, we hope that researchers will add additional data processed with the pipeline that we released to increase the sample resolution both in time and space.

[1]Department of Zoology, University of Cambridge, Cambridge, CB2 3EJ, UK. [2]Smurfit Institute of Genetics, Trinity College Dublin, Dublin, 2, Ireland. [3]Genomics Medicine Ireland, Dublin, Ireland. [4]Institute of Molecular Biology, National Academy of Sciences, 7 Hasratyan Street, 0014, Yerevan, Armenia. [5]Department of Evolutionary Anthropology, University of Vienna, 1090, Vienna, Austria. ✉e-mail: pm604@cam.ac.uk; am315@cam.ac.uk

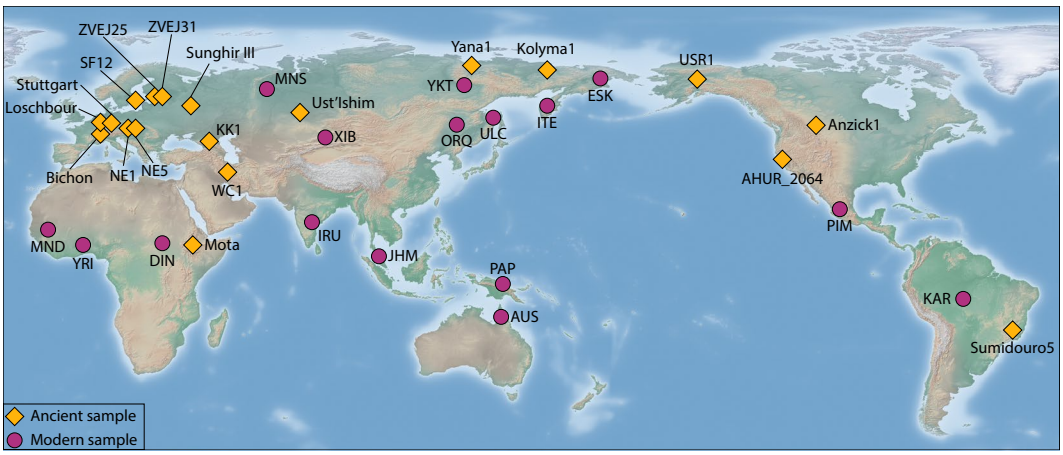

Fig. 1 Geographic distribution of samples included in the dataset. Population acronyms are reported in Table 2.

Sample ID	Mass sampled (g)	Average autosomal coverage
Kotias (KK1)	0.101	12.03
Latvia_HG2 (ZVEJ25)	0.092	18.17
NE5 (14.6)	0.18	15.99
ZVEJ31	0.102	9.97

Table 1. Data statistics for newly sequenced samples. Average autosomal coverage was estimated on bam files after mapping quality filtering (mq20), duplicates removal, indel realignment and 2 bp softclipping.

Sample_ID	Sample_acronym	Population_ID	Country	Latitude	Longitude	Study
SS6004477	AUS	Australian	Australia	−13	143	SGDP – Mallick *et al.*, 2016
LP6005443-DNA_B09	DIN	Dinka	Sudan	8.8	27.4	SGDP – Mallick *et al.*, 2016
LP6005443-DNA_B03	ESK	Eskimo_Sireniki	Russia	64.4	173.9	SGDP – Mallick *et al.*, 2016
LP6005519-DNA_D05	IRU	Irula	India	13.5	80	SGDP – Mallick *et al.*, 2016
LP6005443-DNA_D04	ITE	Itelman	Russia	57	157	SGDP – Mallick *et al.*, 2016
LP6005441-DNA_G06	KAR	Karitiana	Brazil	−10	−63	SGDP – Mallick *et al.*, 2016
LP6005441-DNA_E07	MND	Mandenka	Senegal	12	−12	SGDP – Mallick *et al.*, 2016
LP6005443-DNA_G04	MNS	Mansi	Russia	63.65	62.1	SGDP – Mallick *et al.*, 2016
LP6005441-DNA_F09	ORQ	Oroqen	China	50.4	126.5	SGDP – Mallick *et al.*, 2016
LP6005443-DNA_D08	PAP	Papuan	PapuaNewGuinea	−4	143	SGDP – Mallick *et al.*, 2016
LP6005441-DNA_F10	PIM	Pima	Mexico	29	−108	SGDP – Mallick *et al.*, 2016
LP6005442-DNA_H12	ULC	Ulchi	Russia	52.43	140.42	SGDP – Mallick *et al.*, 2016
LP6005442-DNA_D01	XIB	Xibo	China	43.5	81.5	SGDP – Mallick *et al.*, 2016
LP6005442-DNA_F01	YKT	Yakut	Russia	63	129.5	SGDP – Mallick *et al.*, 2016
LP6005442-DNA_B02	YRI	Yoruba	Nigeria	7.4	3.9	SGDP – Mallick *et al.*, 2016
JHM06	JHM	Jehai	Malaysia	5.25	101.17	McColl *et al.*, 2018

Table 2. Metadata for modern samples. SGDP: Simons Genome Diversity Panel.

Methods

Sample collection. Additional sequence data were generated for four ancient samples which were previously collected and described in the following original publications: ZVEJ25 and ZVEJ31 were published in Jones *et al.*[10], KK1 in Jones *et al.*[11] and NE5 in Gamba *et al.*[12]. Furthermore, 15 additional ancient samples and 16 modern samples have been downloaded from the literature (see Online-only Tables 1 and 2). The final dataset includes 35 samples consisting of 19 ancient and 16 modern samples.

DNA extraction, Library preparation and next-generation sequencing. DNA was extracted and libraries were prepared for ZVEJ25, ZVEJ31, KK1 and NE5 (Table 1), following protocols described in the original publications, with the exception that DNA extracts were incubated with USER enzyme (5 µl enzyme: 16.50 µl of extract) for 3 hours at 37 °C prior to library preparation in order to repair post-mortem molecular damage. The libraries were sequenced across 31 lanes of a HiSeq. 2,500.

Sample	Total Bases	Read Count	GC (%)	Q20 (%)	Q30 (%)	Reads Aligned	Endogenous DNA
KK1_1	32,085,537,489	317,678,589	49.3	96.6	94.5	226,739,842	0.71
KK1_2	31,821,488,543	315,064,243	49.7	96.9	94.8	221,241,435	0.70
KK1_3	30,903,010,501	305,970,401	47.8	96.6	94.4	218,378,529	0.71
KK1_4	28,374,056,452	280,931,252	48.5	96.6	94.5	200,616,589	0.71
KK1_5	27,051,061,997	267,832,297	47.4	96.8	94.8	187,070,443	0.70
KK1_6	26,428,490,321	261,668,221	49.7	96.7	94.5	182,602,757	0.70
NE5_1	15,230,188,243	150,793,943	48.4	96.7	94.6	113,866,866	0.76
NE5_2	22,443,822,868	222,216,068	47.8	96.7	94.6	167,444,317	0.75
NE5_3	19,414,144,957	192,219,257	47.7	96.7	94.6	145,145,785	0.76
NE5_4	35,602,627,361	352,501,261	48.9	96.8	94.7	257,297,424	0.73
NE5_5	39,509,022,440	391,178,440	49.5	96.7	94.5	285,303,006	0.73
NE5_6	38,119,633,918	377,422,118	47.7	96.8	94.7	275,284,926	0.73
ZVEJ25_1	22,502,142,793	222,793,493	48.2	96.8	94.6	173,630,441	0.78
ZVEJ25_2	26,264,479,451	260,044,351	47.5	96.8	94.6	202,756,810	0.78
ZVEJ25_3	19,884,007,259	196,871,359	48.1	96.8	94.6	153,807,348	0.78
ZVEJ25_4	30,314,118,184	300,139,784	47.0	96.9	94.8	234,102,091	0.78
ZVEJ25_5	34,172,785,511	338,344,411	48.2	96.9	94.7	264,070,011	0.78
ZVEJ25_6	32,515,172,804	321,932,404	48.2	96.9	94.7	251,187,453	0.78
ZVEJ31_1	42,951,382,412	425,261,212	52.0	96.9	94.7	215,656,479	0.51
ZVEJ31_2	41,717,115,447	413,040,747	50.7	96.9	94.8	209,910,986	0.51
ZVEJ31_3	36,806,312,233	364,418,933	53.8	96.7	94.4	185,131,989	0.51
ZVEJ31_4	34,986,764,509	346,403,609	51.3	96.9	94.6	166,115,737	0.48
ZVEJ31_5	34,797,229,121	344,527,021	53.8	96.8	94.5	164,914,158	0.48
ZVEJ31_6	39,275,860,102	388,869,902	52.0	96.8	94.6	185,999,314	0.48

Table 3. Raw data statistics for the newly sequenced libraries.

Bioinformatics analysis. *Ancient samples.* The following approach was used for both the newly sequenced ancient samples and downloaded raw fastq files from previously published ancient samples.

Adapters were trimmed with cutadapt v1.9.1[13] and then raw reads were aligned to human reference sequence hg19/GRCh37 with the rCRS mitochondrial sequence using bwa aln v0.7.12[14] with seeding disabled (-l 1000), maximum edit distance set to -n 0.01 and maximum number of gap opens set to -o 2. These parameters are recommended for aDNA as they allow for more mismatches to the reference genome[15]. Sai files were converted into sam files using bwa samse v0.7.12 and the read group line was also added. Bam files were generated using samtools view v1.9[16]. Reads from multiple libraries belonging to the same sample were merged with the module MergeSamFiles within Picard v2.9.2[17]. Aligned reads were filtered for minimum mapping quality 20 with samtools view v1.9. Indexing, sorting and duplicate removal (rmdup) were performed with samtools v1.9. Indels were realigned using The Genome Analysis Toolkit v3.7[18] (module RealignerTargetCreator and IndelRealigner) and 2 bp were softclipped (phred quality score reduced to 2) at the start and ends of reads using a custom python script. Final bam files were split by chromosome using samtools view v1.9 and variant calling was performed with UnifiedGenotyper from The Genome Analysis Toolkit v3.7. All calls were filtered for minimum base quality 20 (−mbq 20) and reference-bias free priors were used (−inputPrior 0.0010 -inputPrior 0.4995). The same priors have been used for modern samples in the Simons Genome Diversity Panel[7].

Raw data was not available for four previously published samples included in this dataset and so alignment data was processed instead (Loschbour, Stuttgart_LBK, Ust_Ishim and WC1). The data for Loschbour, Stuttgart_LBK and Ust_Ishim had been aligned to GRCh37 with additional decoy sequences (hs37d5) using the same non-default bwa aln parameters. We removed reads aligning to these decoys and updated the bam file headers accordingly, before proceeding with the processing pipeline outlined above. The available alignment data from WC1 was mapped using bwa aln with default parameters and had a mapping quality filter of 25 already applied. We realigned these reads using the non-default parameters and proceeded with the processing pipeline.

For those who wish to follow this pipeline with newly produced ancient DNA data, we recommend a final data authentication step. Characteristic patterns of aDNA post-mortem damage (e.g. short read lengths and cytosine deamination) can be verified using mapDamage software[19]. A number of methods exist to estimate contamination levels on the basis of these damage patterns, as well as other measures, including heterozygosity at haploid loci and the breakdown of linkage disequilibrium[20–23]

We focused on selecting a subset of the genome representing neutral genomic variation for demographic inferences[24,25]. Therefore, specific filters were applied to discard: recombination hotspots (filter_hotspot1000g), poor mapping quality regions (filter_Map20), recent duplication (recent duplications, RepeatMasker score <20), recent segmental duplication (filter_segDups), simple repeats (filter_simpleRepeat), gene exons together with 1000 bp flanking and conserved elements together 100 bp flanking (filter_selection_10000_100) and positions with systematic sequencing errors (filter_SysErrHCB and filter_SysErr.starch). All CpG sites were removed as well as C and G sites with an adjacent missing genotype. Genotypes were filtered by minimum coverage 8x and

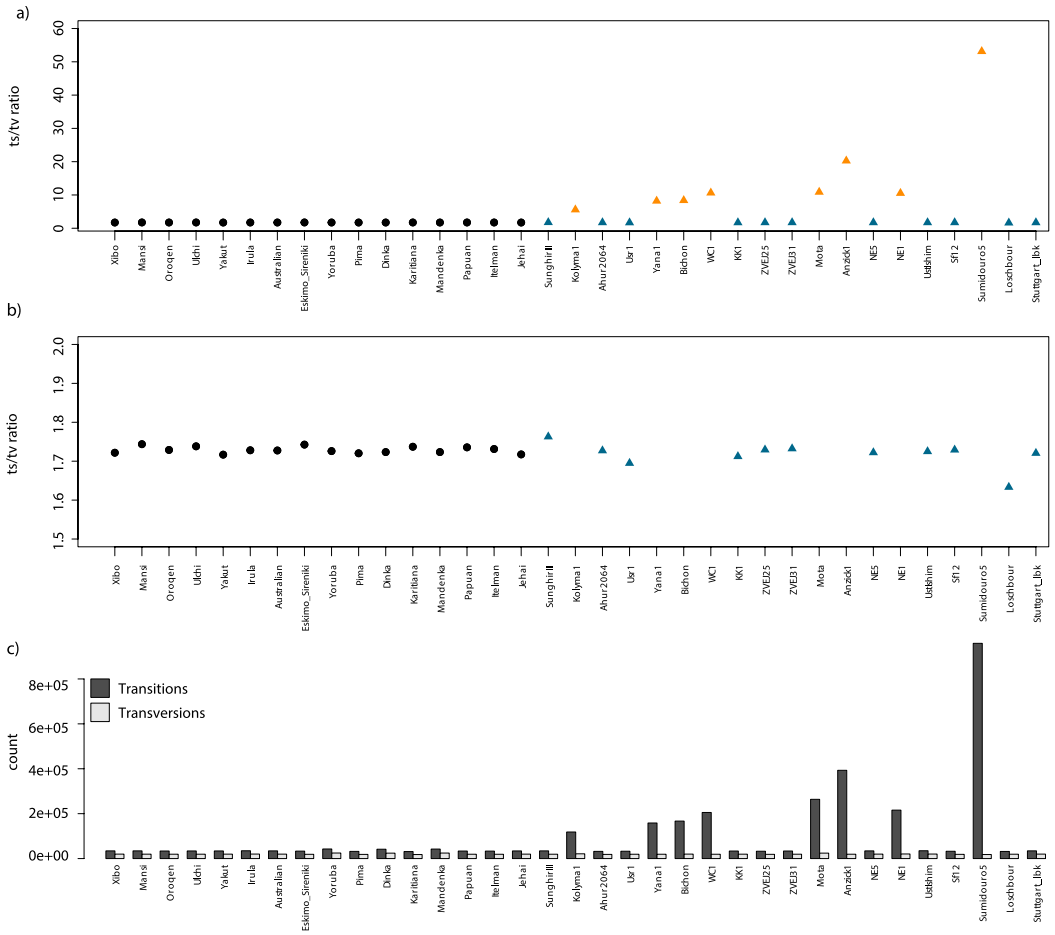

Fig. 2 (**a**) Transitions/Transversions ratio (ts/tv) per sample. Ancient and modern samples are represented by triangles and circles respectively. UDG and non-UDG treated samples are in blue and orange respectively. (**b**) same as in a) but with a different y axis to focus on the ts/tv ratio among modern and UDG-treated ancient samples. (**c**) Number of transitions (ts) and transversions (tv) per sample.

maximum coverage defined as twice the average coverage. Vcf files per chromosome belonging to the same sample were concatenated using vcf-concat from vcftools v0.1.15[2,26].

Modern samples. Bam files were downloaded from the Simons Genome Diversity Panel[7] and from McColl *et al.*[27]. (Table 2). Bam files were split by chromosome and variant calling, filtering for GC sites and coverage were performed as described above for the ancient samples with the same options and thresholds.

Final dataset. Per sample vcf files were compressed with bgzip and indexed with tabix from htslib v1.6[16]. The final dataset was assembled by merging filtered compressed vcf files for all modern and ancient samples with bcftools merge v1.6[16]. Only sites with called genotypes for all samples were kept using vcftools v0.1.15 (--max-missing 1). Tri-allelic sites were also discarded using bcftools view v1.6 (-m1 -M2). Final vcf statistics were generated with bcftools stats v1.6. Downstream analysis and plotting were performed in R v3.6.3[28].

Data Records

All newly generated sequencing raw reads have been deposited in the NCBI Sequence Read Archive Bioproject PRJNA670050[29]. Both filtered and unfiltered vcf files have been uploaded to figshare[30].

Technical Validation

Summary of newly generated data. DNA was extracted for four previously published samples (ZVEJ25, ZVEJ31, KK1 and NE5) and sequence data were generated with an average coverage between 10x and 18x (Table 1). Endogenous DNA was estimated between 0.48 and 0.71 across all libraries (Table 3). Each library generated between 150 and 425 millions of reads corresponding to 15.2 and 42.9 Gb respectively (Table 3).

Summary of the whole dataset including ancient and modern samples. The final dataset includes 35 samples with 509,351,727 sites in neutral regions before filtering (see Methods section for a detailed description of which regions were considered for variant calling). Sites not called across all samples (0% missing data

allowed) were then discarded and 72,045,170 were retained. Multi-allelic sites (3815) were also removed bringing the final number of filtered sites to 72,041,355 (Online-only Table 2). Minimum and maximum coverage per sample within the final dataset is 11.3x and 55x respectively (within filtered intervals) with an average coverage across all samples of 29.7x (Online-only Table 2). We calculated the number of transitions (ts), transversions (tv) and the ts/tv ratio per sample (Online-only Table 2). As expected, all eight ancient samples that were not subjected to UDG-treatment showed a higher ts/tv ratio than their UDG-treated counterparts (see Fig. 2), consistent with higher levels of DNA damage in these samples. The Brazilian sample Sumidouro 5 shows the highest excess of transition, possibly due to poor DNA preservation caused by environmental conditions. All other samples (both modern and UDG-treated ancient) showed similar ts/tv ratio with an average of 1.72, maximum and minimum of 1.76 and 1.63 respectively (see Online-only Table 2, Fig. 2).

Code availability

All newly generated sequencing raw reads (see Table 3) have been deposited in the NCBI Sequence Read Archive (SRR12854172, SRR12854173, SRR12854174, SRR12854175). Six compressed fastq files per sample were uploaded. The fastq files have the same names as the libraries described in Table 3.

The genetic pipeline used to process the data is available at https://github.com/EvolEcolGroup/data_paper_genetic_pipeline.

The filtered compressed vcf file used for the analyses has been uploaded to figshare[30] with the title "A curated dataset of modern and ancient high-coverage shotgun human genomes".

Received: 26 October 2020; Accepted: 10 June 2021;
Published online: 04 August 2021

References

1. Racimo, F. & Sikora, M. Vander Linden, M., Schroeder, H. & Lalueza-Fox, C. Beyond broad strokes: sociocultural insights from the study of ancient genomes. *Nat. Rev. Genet.* **21**, 355–366 (2020).
2. Downloadable genotypes of present-day and ancient DNA data (compiled from published papers). https://reich.hms.harvard.edu/downloadable-genotypes-present-day-and-ancient-dna-data-compiled-published-papers (2020).
3. Link, V. *et al. ATLAS: Analysis Tools for Low-depth and Ancient Samples*. Preprint at https://www.biorxiv.org/content/10.1101/105346v1 (2017).
4. Peltzer, A. *et al.* EAGER: efficient ancient genome reconstruction. *Genome Biol.* **17**, 60 (2016).
5. Schubert, M. *et al.* Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nat. Protoc.* **9**, 1056–1082 (2014).
6. Yates, J. A. F. *et al.* Reproducible, portable, and efficient ancient genome reconstruction with nf-core/eager. *Peer J* **9**, e10947 (2021).
7. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
8. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
9. Bergström, A. *et al.* Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367** (2020).
10. Jones, E. R. *et al.* The Neolithic Transition in the Baltic Was Not Driven by Admixture with Early European Farmers. *Curr. Biol.* **27**, 576–582 (2017).
11. Jones, E. R. *et al.* Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nat. Commun.* **6**, 8912 (2015).
12. Gamba, C. *et al.* Genome flux and stasis in a five millennium transect of European prehistory. *Nat. Commun.* **5**, 5257 (2014).
13. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
14. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
15. Schubert, M. *et al.* Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics* **13**, 178 (2012).
16. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* **25**, 2078–2079 (2009).
17. Picard Tools - By Broad Institute. http://broadinstitute.github.io/picard/.
18. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
19. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F. & Orlando, L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29**, 1682–1684 (2013).
20. Moreno-Mayar, J. V. *et al.* A likelihood method for estimating present-day human contamination in ancient male samples using low-depth X-chromosome data. *Bioinforma. Oxf. Engl.* **36**, 828–841 (2020).
21. Nakatsuka, N. *et al.* ContamLD: estimation of ancient nuclear DNA contamination using breakdown of linkage disequilibrium. *Genome Biol.* **21**, 199 (2020).
22. Peyrégne, S. & Peter, B. M. AuthentiCT: a model of ancient DNA damage to estimate the proportion of present-day DNA contamination. *Genome Biol.* **21**, 246 (2020).
23. Renaud, G., Slon, V., Duggan, A. T. & Kelso, J. Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. *Genome Biol.* **16**, 224 (2015).
24. Kuhlwilm, M. *et al.* Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature* **530**, 429–433 (2016).
25. Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G. & Siepel, A. Bayesian inference of ancient human demography from individual genome sequences. *Nat. Genet.* **43**, 1031–1034 (2011).
26. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
27. McColl, H. *et al.* The prehistoric peopling of Southeast Asia. *Science* **361**, 88–92 (2018).
28. R Core Team. R: A Language and Environment for Statistical Computing. https://www.R-project.org/ (2020).
29. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP287922 (2021).
30. Maisano Delser, P. *et al.* A curated dataset of modern and ancient high-coverage shotgun human genomes. *figshare* https://doi.org/10.6084/m9.figshare.c.5183474 (2021).

Acknowledgements

PMD was supported by funding from the HERA Joint Research Programme "Uses of the Past" (CitiGen), the European Union's Horizon 2020 research and innovation programme under Grant Agreement 649307. PMD and AM were supported by ERC Consolidator Grant 647797 'LocalAdaptation'. E.R.J. was supported by a Herchel Smith Research Fellowship. RP was supported by ERC starting grant ADNABIOARC (263441).

Author contributions

A.M. designed the project. P.M.D., L.C., E.J. and A.H. performed the analyses. R.P. provided the samples. A.M. and P.M.D. wrote the manuscript. All authors had input in the manuscript and approved the final version.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to P.M.D. or A.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

The Creative Commons Public Domain Dedication waiver http://creativecommons.org/publicdomain/zero/1.0/ applies to the metadata files associated with this article.

© The Author(s) 2021