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

Pierpaolo Maisano Delser\textsuperscript{1,2,}✉, Eppie R. Jones\textsuperscript{1,3}, Anahit Hovhannisyan\textsuperscript{4}, Lara Cassidy\textsuperscript{2}, Ron Pinhasi\textsuperscript{5} & Andrea Manica\textsuperscript{1,6}✉

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)\textsuperscript{1}. However, there are just a few high-quality (coverage >10x) shotgun whole-genome sequenced ancient samples\textsuperscript{2}. While genetic pipelines have been previously published\textsuperscript{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\textsuperscript{7}, 1000 Genome Project\textsuperscript{8} and Human Genome Diversity Project (HGDP-CEPH panel)\textsuperscript{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.

\textsuperscript{1}Department of Zoology, University of Cambridge, Cambridge, CB2 3EJ, UK. \textsuperscript{2}Smurfit Institute of Genetics, Trinity College Dublin, Dublin, 2, Ireland. \textsuperscript{3}Genomics Medicine Ireland, Dublin, Ireland. \textsuperscript{4}Institute of Molecular Biology, National Academy of Sciences, 7 Hasratyan Street, 0014, Yerevan, Armenia. \textsuperscript{5}Department of Evolutionary Anthropology, University of Vienna, 1090, Vienna, Austria. \textsuperscript{6}e-mail: pm604@cam.ac.uk; am315@cam.ac.uk
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.\textsuperscript{10}, KK1 in Jones et al.\textsuperscript{11} and NE5 in Gamba et al.\textsuperscript{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.

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

| Sample ID  | Sample acronym | Population ID | Country | Latitude | Longitude | Study                        |
|------------|----------------|---------------|---------|----------|-----------|------------------------------|
| SS6004477  | AUS            | Australian    |         | −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        | Mandinka      | 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   | 56.4     | 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          |
Ancient samples and downloaded raw fastq files from previously published ancient samples were realigned using The Genome Analysis Toolkit v3.7\(^{18}\) (module RealignerTargetCreator and IndelRealigner) tools view v1.9. Indexing, sorting and duplicate removal (rmdup) were performed with samtools v1.9. Indels 1000 bp flanking and conserved elements together 100 bp flanking (filter_selection_10000_100) and positions recent segmental duplication (filter_segDups), simple repeats (filter_simpleRepeat), gene exons together with poor mapping quality regions (filter_Map20), recent duplication (recent duplications, RepeatMasker score hg19/GRCh37 with the rCRS mitochondrial sequence using bwa aln v0.7.12\(^{14}\) with seeding disabled (-l 1000), recommended for aDNA as they allow for more mismatches to the reference genome\(^{15}\). Sai files were converted with UnifiedGenotyper from The Genome Analysis Toolkit v3.7. All calls were filtered for minimum base quality mbq 20) and reference-bias free priors were used (-inputPrior 0.0010 -inputPrior 0.4995). The same priors were used when determining the likelihood of specific variants \(2\). 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 well suited to process ancient DNA with systematic sequencing errors (filter_SysErrHCB and filter_SysErr.starch). All CpG sites were removed as deamination (e.g. 5-methylcytosine) 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 realigned these reads using the non-default parameters and proceeded with the processing pipeline. 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. WC1 was mapped using bwa aln with default parameters and had a mapping quality filter of 25 already applied. WC1 was mapped using bwa aln with default parameters and had a mapping quality filter of 25 already applied. 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\(^{20–23}\).

Table 3. Raw data statistics for the newly sequenced libraries.
maximum coverage defined as twice the average coverage. Vcf files per chromosome belonging to the same sample were concatenated using vcf-concat from vcftools v0.1.152. Modern samples. Bam files were downloaded from the Simons Genome Diversity Panel7 and from McColl et al. (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. The final dataset was assembled by merging filtered compressed vcf files for all modern and ancient samples with bcftools merge v1.6. 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.

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 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\textsuperscript{30} with the title "A curated dataset of modern and ancient high-coverage gunshot 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/10.1101/105346/v1 (2017).
4. Peltzer, A. et al. EAGER: efficient ancient genome reconstruction. Genome Biol. 17, 60 (2016).
5. Schulert, M. et al. Characterization of ancient and modern genomes by SNP detection and phylogenetic 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, 2527 (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. Schulert, 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. Bioinformat. 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. Jonsson, 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. Bioinformat. 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. Peyrigne, S. & Petes, 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., Gukov, B., Danko, C. G. & Snieple, 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:SRR287922 (2021).
30. Maissano Delser, P. et al. A curated dataset of modern and ancient high-coverage gunshot 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