A High-Coverage Genome Sequence from an Archaic Denisovan Individual

Matthias Meyer,1‡‡ Martin Kircher,1‡‡ Marie-Theres Gansauge,1 Heng Li,2 Fernando Racimo,1 Swapnanillick,2,3 Joshua G. Schraiber,4 Flora Jay,5 Kay Prüfer,1 Cesare de Filippo,1 Peter H. Sudmant,6 Can Alkan,4,5 Qiaomei Fu,1,7 Ron Do,2 Nadin Rohland,2,3 Arti Tandon,2,3 Michael Siebauer,1 Richard E. Green,8 Katarzyna Bryc,9 Adrian W. Briggs,1 Udo Stenzel,1 Jesse Dabney,1 Jay Shendure,4 Jacob Kitzman,5 Michael F. Hammer,9 Michael V. Shukov,10 Anatoli P. Derevianko,10 Nick Patterson,2 Aida M. Andrés,1 Evan E. Eichler,6,11 Montgomery Slatkin,1 David Reich,6,12‡‡‡‡ Janet Kelso,1 Svante Pääbo1‡

1Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, D-04103 Leipzig, Germany. 2Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, MA 02142, USA. 3Department of Genetics, Harvard Medical School, Boston, MA 02115, USA. 4Department of Integrative Biology, University of California, Berkeley, CA 94720, USA. 5Department of Computer Engineering, Bilkent University, 06800 Ankara, Turkey. 6Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA 98195, USA. 7CAS-MPS Joint Laboratory for Human Evolution, Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, 100044 Beijing, China. 8Jack Baskin School of Engineering, University of California, Santa Cruz, CA 95064, USA. 9Arizona Research Laboratories, Division of Biotechnology, University of Arizona, Tucson, AZ 85721, USA. 10Paleolithic Department, Institute of Archaeology and Ethnography, Russian Academy of Sciences, Siberian Branch, 630090 Novosibirsk, Russia. 11Howard Hughes Medical Institute, Seattle, WA 98195, USA.

‡These authors contributed equally to this work. 
‡‡§To whom correspondence should be addressed. E-mail: mmeyer@eva.mpg.de (M.M.); reich@genetics.med.harvard.edu (D.R.); paabo@eva.mpg.de (S.P.)

We present a DNA library preparation method that has allowed us to reconstruct a high coverage (30X) genome sequence of a Denisovan, an extinct relative of Neandertals. The quality of this genome allows a direct estimation of Denisovan heterozygosity indicating that genetic diversity in these archaic hominins was extremely low. It also allows tentative dating of the specimen on the basis of “missing evolution” in its genome, detailed measurements of Denisovan and Neandertal admixture into present-day human populations, and the generation of a near-complete catalog of genetic changes that swept to high frequency in modern humans since their divergence from Denisovans.

Draft genome sequences have been recovered from two archaic human groups, Neandertals (1) and Denisovans (2). While Neandertals are defined by distinct morphological features and occur in the fossil record of Europe, Western and Central Asia from at least 230,000 until about 30,000 years ago (3), Denisovans are known only from a distal manual phalanx and two molars, all excavated at Denisova Cave in the Altai Mountains in southern Siberia (2, 4, 5). The draft nuclear genome sequence retrieved from the Denisovan phalanx (2) revealed that Denisovans are a sister group to Neandertals (2), with the Denisovan nuclear genome sequence falling outside Neandertal genetic diversity, suggesting an independent population history that differs from that of Neandertals. Also, whereas a genetic contribution from Neandertal to the present-day human gene pool is present in all populations outside Africa, a contribution from Denisovans is found exclusively in island Southeast Asia and Oceania (6).

Both published archaic genome sequences are of low coverage: 1.9-fold genomic coverage from the Denisovan phalanx and a total of 1.3-fold derived from three Crocean Neandertals. As a consequence, many changes in the genomes are affected by sequencing errors or nucleotide misincorporations caused by DNA damage. Previous attempts to generate a genome sequence of high coverage from an archaic human have been hampered by the high levels of environmental contamination. The fraction of hominin endogenous DNA is commonly smaller than 1% and rarely approaches 5% (7, 8), making shotgun sequencing of the entire genome economically and logistically impractical. The only known exception is the Denisovan phalanx, which contains ~70% endogenous DNA. However, an extremely small fragment of this specimen is available to us, and the absolute number of endogenous molecules that could be recovered from the sample was too low to generate high genomic coverage.

A single-stranded library preparation method. DNA libraries for sequencing are normally prepared from double-stranded DNA (Fig. 1). However, for ancient DNA the use of single-stranded DNA may be advantageous as it will double its representation in the library. Furthermore, in a single-stranded DNA library, double-stranded molecules that carry modifications on one strand that prevent their incorporation into double-stranded DNA libraries could still be represented by the unmodified strand. We therefore devised a single-stranded library preparation method wherein the ancient DNA is dephosphorylated, heat denatured, and ligated to a biotinylated adaptor oligonucleotide, which allows its immobilization on streptavidin-coated beads (Fig. 1). A primer hybridized to the adaptor is then used to copy the original strand with a DNA polymerase. Finally, a second adaptor is joined to the copied strand by blunt-end ligation and the library molecules are released from the beads. The entire protocol is devoid of DNA purification steps, which inevitably cause loss of material.

We applied this method to aliquots of the two DNA extracts (as well as side fractions) that were previously generated from the 40 mg of bone that comprised the entire inner part of the phalanx (2, 8). Comparisons of these newly generated libraries to the two libraries generated in the previous study (2) show at least a 6-fold and 22-fold increase in the recovery of library molecules (8), which is particularly pronounced for longer molecules (fig. S4).

In addition to improved sequence yield, the single-strand library protocol reveals new aspects of DNA fragmentation and modification patterns (8). Since the ends of both DNA strands are left intact, it reveals that strand breakage occurs preferentially before and after guanine residues (fig. S6), suggesting that guanine nucleotides are frequently lost from ancient DNA, possibly as the result of depurination. It also reveals that deamination of cytosine residues occurs with almost equal frequencies at both ends of the ancient DNA molecules. Since deamination is hypothesized to be frequent in single-stranded DNA overhangs (9, 10), this suggests that 5′- and 3′-overhangs occur at similar lengths and fre-
Genome sequencing. We sequenced these libraries from both ends using Illumina’s Genome Analyzer IIx and included reads for two indexes (11), which were added in the clean room to exclude the possibility of downstream contamination with modern DNA libraries (1). Sequences longer than 35 bp were aligned to the human reference genome (GRCh37/1000 Genome project release) and the chimpanzee genome (CGSC 2.1/UCSC pantro2 release) with BWA (12). After removal of PCR duplicates, insertions/deletions and genotypes were called with GATK (8, 13). The three Denisovan libraries yielded 82.2 gigabases of non-duplicate sequence aligned to the human genome (8). Together with previous data (2) this provides about 31-fold coverage of the ~1.86 gigabases of the human autosomal genome to which short sequences can be confidently mapped (8). We also sequenced the genomes of eleven present-day individuals: a San, Mbuti, Mandenka, Yoruba and Dinka from Africa; a French and Sardinian from Europe; a Han, Dai and Papuan from Asia; and a Karitiana from South America. DNA from these individuals was bar-coded, pooled and sequenced to ~24- to 33-fold genomic coverage (8). Because the samples were pooled, sequencing errors are the same across samples and are not expected to bias inferences about population relationships.

Genome quality. We used three independent measures to estimate human contamination in the Denisovan genome sequence (8). First, on the basis of a ~4,100-fold coverage of the Denisovan mitochondrial (mt) genome we estimate that 0.35% (95% confidence interval (C.I.) 0.33%–0.36%) of fragments that overlap positions where the Denisovan mtDNA differs from most present-day humans show the modern human variant. Second, using the fact that the Denisovan phalanx comes from a female (2), we infer male human DNA contamination to be 0.07% (C.I. 0.05%–0.09%) from alignments to the Y-chromosome. Third, a maximum-likelihood quantification of autosomal contamination gives an estimate of 0.22% (C.I. 0.22–0.23%). We conclude that less than 0.5% of the hominin sequences determined are extraneous to the bone (i.e., contamination from present-day humans).

Coverage of the genome is fairly uniform with 99.93% of the ‘mappable’ positions covered by at least one, 99.43% by at least ten, and 92.93% by at least 20 independent DNA sequences (8). High-quality genotypes (genotype quality ≥ 40) could be determined for 97.64% of the positions. While coverage in libraries prepared from ancient samples with previous methods are biased toward GC-rich sequences (14), the coverage of the libraries prepared with the single-stranded method from the Denisovan individual is similar to the eleven present-day human genomes (prepared from double-stranded DNA) in that coverage is positively correlated with AT-content (fig. S12).

To estimate average per-base error rates in the Denisovan genome we counted differences between the sequenced DNA fragments and regions of the human genome that are highly conserved within primates [approximately 5.6 million bases, (8)]. The error rate is 0.13% for the Denisovan genome, 0.17% to 0.19% for the genome sequences from the eleven present-day humans, and 1.2–1.7% for the two trios sequenced by the 1000 Genomes Pilot project (table S11). The lower Denisovan error rate is likely due to consensus-calling from duplicate reads representing the same DNA fragments, and from overlap-merging of paired-end reads.
Molecular estimates of divergence and fossil age. We estimated the average DNA sequence divergence of all pair-wise combinations of the Denisovan genome and the 11 present-day humans as a fraction of the branch leading from the human-chimpanzee ancestor to present-day humans (Fig. 2, (8)). Assuming a human-chimpanzee average DNA sequence divergence of 6.5 million years ago (15), the Denisova-present-day human divergence is approximately 800,000 years, close to our previous estimate (2).

We next estimated the divergence of the archaic and modern human populations, which must be more recent than the DNA sequence divergence. To do this, we identified sites that are variable in a present-day West African individual, who is not affected by Denisovan or Neandertal gene flow, and counted how often the Denisovan and Neandertal genomes carry derived alleles not present in chimpanzee (1). From this, we estimate the population divergence between Denisovans and present-day humans to be 170,000–700,000 years (8). This is wider than our previous estimate (1), largely because it takes into account recent studies that broaden the range of plausible estimates for human mutation rates and thus the human-chimpanzee divergence date.

When comparing the number of substitutions inferred to have occurred between the human-chimpanzee ancestor and the Denisovan and present-day human genomes, the number for the Denisovan genome is 1.16% lower (1.13–1.27%) (Fig. 2) (8). This presumably reflects the age of the Denisovan bone, which had less time to accumulate changes than present-day humans. Assuming 6.5 million years of sequence divergence between humans and chimpanzees, the shortening of the Denisovan branch allows the bone to be tentatively dated to between 74,000 and 82,000 years before present, in general agreement with the archaeological dates (2). However, we caution that multiple sources of error may affect this estimate (8). For example, the numbers of substitutions inferred to have occurred to the present-day human sequences vary by up to one-fifth of the reduction estimated for the Denisovan bone. Nevertheless, the results suggest that in the future it will be possible to determine dates of fossils based on genome sequences.

Denisovan and Neandertal gene flow. To visualize the relationship between Denisova and the eleven present-day humans, we used TreeMix, which simultaneously infers a tree of relationships and “migration events” (16) (Fig. 3). This method estimates that 6.0% of the genomes of present-day Papuans derive from Denisovans (8). While this procedure does not provide a perfect fit to the data (for example, it does not model Neandertal admixture), it agrees with our previous finding that Denisovans have contributed to the genomes of present-day Melanesians, Australian Aborigines, and other South-East Asian islanders (2, 6).

We tested whether Denisovans share more derived alleles with any of the 11 present-day humans (8). To increase the power to detect gene flow, we used a new approach, ‘enhanced’ D-statistics, which restricts the analysis to alleles that are not present in 35 African genomes and are thus more likely to come from archaic humans. This confirms that Denisovans share more alleles with Papuans than with mainland Eurasians (Fig. 4A, table S24). However, in contrast to a recent study proposing more allele sharing between Denisovans and populations from southern China, such as the Dai, than with populations from northern China, such as the Han (17), we find less Denisovan allele sharing with the Dai than with the Han (although non-significantly so, Z = −0.9) (Fig. 4B) (table S25). Further analysis shows that if Denisovans contributed any DNA to the Dai, it represents less than 0.1% of their genomes today (table S26).

Interestingly, we find that Denisovans share more alleles with the three populations from eastern Asia and South America (Dai, Han, and Karitiana) than with the two European populations (French and Sardinian) (Z = 5.3). However, this does not appear to be due to Denisovan gene flow into the ancestors of present-day Asians, since the excess archaic material is more closely related to Neandertals than to Denisovans (table S27). We estimate that the proportion of Neandertal ancestry in Europe is 24% lower than in eastern Asia and South America (95% C.I. 12–36%). One possible explanation is that there were at least two independent Neandertal gene flow events into modern humans (18). An alternative explanation is a single Neandertal gene flow event followed by dilution of the Neandertal proportion in the ancestors of Europeans due to later migration out of Africa. However, this would require about 24% of the present-day European gene pool to be derived from African migrations subsequent to the Neandertal admixture.

Strikingly, Papuans share more alleles with the Denisovan genome on the autosomes than on the X chromosome (P = 0.01 by a two-sided test) (table S28). One possible explanation for this finding is that the gene flow into Papuan ancestors involved primarily Denisovan males. Another explanation is population substructure combined with predominantly female migration among the ancestors of modern humans as they encountered Denisovans (thus diluting the Denisovan component on chromosome X) (19). A third possibility is natural selection against hybrid incompatibility alleles, which are known to be concentrated on chromosome X (20). We note that some autosome (e.g., chromosome 11) also have less Denisovan ancestry (table S30), suggesting that factors such as hybrid incompatibility may be at play.

Denisovan genetic diversity. The high quality of the Denisovan genome allowed us to measure its heterozygosity, i.e., the fraction of nucleotide sites that are different between a person’s maternal and paternal genomes (Fig. 5A). Several methods indicate that the Denisovan heterozygosity is about 0.022% (8). This is ~20% of the heterozygosity seen in the Africans, ~26–33% of that in the Eurasians, and 36% of that in the Karitiana, a South American population with extremely low heterozygosity (21). Since we find no evidence for unusually long stretches of homozygosity in the Denisovan genome (8), this is not due to inbreeding among the immediate ancestors of the Denisovan individual. We thus conclude that genetic diversity of the population to which the Denisovan individual belonged was very low compared to present-day humans.

To estimate how Denisovan and modern human population sizes have changed over time we applied a Markovian coalescent model (22) to all genomes analyzed. This shows that present-day human genomes share similar population size changes, in particular a more than two-fold increase in size before 125,000–250,000 years ago (depending on the mutation rates assumed (23), Fig. 5B). Denisovans, in contrast, show a drastic decline in size at the time when the modern human population
Of more relevance may be examination of aspects of the Denisovan karyotype. The great apes have 24 pairs of chromosomes while humans have 23. This difference is caused by a fusion of two acrocentric chromosomes that formed the metacentric human chromosome 2 (25), and resulted in the unique head-to-head joining of the telomeric hexamer repeat GGGGTT. A difference in karyotype would likely have reduced the fertility of any offspring of Denisovans and modern humans. We searched all DNA fragments sequenced from the Denisovan individual and identified twelve fragments containing joined repeats. By contrast, reads from several chimpanzees and bonobos failed to yield any such fragments (8). We conclude that Denisovans and modern humans (and presumably Neandertals) shared a karyotype consisting of 46 chromosomes.

**Features unique to modern humans.** Genome sequences of archaic human genomes allow the identification of derived genomic features that became fixed or nearly fixed in modern humans after the divergence from their archaic relatives. The previous Denisovan and Neandertal genomes (1, 2) allowed less than half of all such features to be assessed with confidence. The current Denisovan genome enables us to generate an essentially complete catalog of recent changes in the human genome accessible with short read technology (26). In total, we identified 111,812 single nucleotide changes (SNCs) and 9,499 insertions and deletions where modern humans are fixed for the derived state while the Denisovan individual carried the ancestral, i.e. ape-like, variant (8). This is a relatively small number. We identified 260 human-specific SNCs that cause fixed amino acid substitutions in well-defined human genes, 72 fixed SNCs that affect splice sites, and 35 SNCs that affect well-defined motifs inside regulatory regions.

One way to identify changes that may have functional consequences is to focus on sites that are highly conserved among primates and that have changed on the modern human lineage after separation from Denisovan ancestors. We note that among the 23 most conserved positions affected by amino acid changes (primate conservation score \( \geq 0.95 \)), eight affect genes that are associated with brain function or nervous system development (\( \text{NOVA1, SLITRK1, KATNA1, LUZP1, ARHGAP32, ADSL, HTR2B, CBTPAP2} \)). Four of these are involved in axonal and dendritic growth (\( \text{SLITRK1, KATNA1} \)) and synaptic transmission (\( \text{ARHGAP32, HTR2B} \)) and two have been implicated in autism (\( \text{ADSL, CBTPAP2} \)). \( \text{CBTPAP2} \) is also associated with susceptibility to language disorders (27) and is particularly noteworthy as it is one of the few genes known to be regulated by \( \text{FOXP2} \), a transcription factor involved in
language and speech development as well as synaptic plasticity (28). It is thus tempting to speculate that crucial aspects of synaptic transmission may have changed in modern humans.

Our limited understanding of how genes relate to phenotypes makes it impossible to predict the functional consequences of these changes. However, diseases caused by mutations in genes offer clues as to which organ systems particular genes may affect. Of the 34 genes with clear associations with human diseases that carry fixed substitutions changing the encoded amino acids in present-day humans, four (HPS5, GGCX, ERCC5, ZMPSTE24) affect the skin and six (RP1L1, GGCX, FRMD7, ABCA4, VCAN, CRYBB3) affect the eye. Thus, particular aspects of the physiology of the skin and the eye may have changed recently in human history. Another fixed difference occurs in EVC2, which when mutated causes Ellis-Van Creveld syndrome. Among other symptoms, this syndrome includes taurodontism, an enlargement of the dental pulp cavity and fusion of the roots, a trait that is common in teeth of Neandertals and other archaic humans. A Denisovan molar found in the cave has an enlarged pulp cavity but lacks fused roots (2). This suggests that the mutation in EVC2, perhaps in conjunction with mutations in other genes, has caused a change in dental morphology in modern humans.

We also examined duplicated regions larger than 9 kilobase pairs (kbp) in the Denisovan and the present-day human genomes, and found the majority of them to be shared (8). However, we find ten regions that are expanded in all present-day humans but not in the Denisovan ge-
Nome. Notably, one of these overlaps a segmental duplication associated with a pericentric inversion of chromosome 18. In contrast to humans, the Denisovan genome harbors only a partial duplication of this region, suggesting that a deletion occurred in the Denisovan lineage. However, we are unable to resolve if the pericentric inversion is indeed present in Denisovans.

**Implications for archaic and modern human history.** It is striking that genetic diversity among Denisovans was low although they were present in Siberia as well as presumably in Southeast Asia where they interacted with the ancestors of present-day Melanesians (6). Only future research can show how wide their geographic range was at any one time in their history. However, it is likely that they have expanded from a small population size with not enough time elapsing for genetic diversity to correspondingly increase. When technical improvements such as the ones presented here will make it possible to sequence a Neandertal genome to a quality comparable to the Denisovan and modern genomes, it will be important to clarify whether the temporal trajectory of Neandertal effective population size matches that of the Denisovans. If that is the case, it is likely that the low Denisovan diversity reflects the expansion out of Africa of a population ancestral to both Denisovans and Neandertals, a possibility that seems compatible with the dates for population divergences and population size changes presented.

By providing a comprehensive catalog of features that became fixed in modern humans after their separation from their closest archaic relatives, this work will eventually lead to a better understanding of the biological differences that existed between the groups. This should ultimately aid in determining how it was that modern humans came to expand dramatically in population size as well as cultural complexity while archaic humans eventually dwindled in numbers and became physically extinct.

**References and Notes**

8. Materials and methods are available as supplementary materials on Science Online.

A. Scally et al., Insights into hominid evolution from the gorilla genome sequence. *Nature* 483, 169 (2012). doi:10.1038/nature10842 Medline


M. F. Hammer et al., The ratio of human X chromosome to autosomal diversity is positively correlated with genetic distance from genes. *Nat. Genet.* 42, 830 (2010). doi:10.1038/ng.651 Medline


178. S. P. David et al., Bupropion efficacy for smoking cessation is influenced by the DRD2 Taq1A polymorphism: Analysis of pooled data from two clinical trials. Nicotine Tob. Res. 9, 1251 (2007). doi:10.1080/14622030701705027

179. P. Burton et al., Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447, 661 (2007). doi:10.1038/nature05911


191. C. A. Mulcare et al., The T allele of a single-nucleotide polymorphism 13.9 kb upstream of the lactase gene (LCT) (C-13.9kbT) does not predict or cause the lactase-persistence phenotype in Africans. Am. J. Hum. Genet. 74, 1102 (2004). doi:10.1086/420510


Acknowledgments: Raw Denisovan sequence data are available from the European Nucleotide Archive (ENA) under study accession ERP001519. Raw Denisovan sequence and alignment data are available at http://cdn.eva.mpg.de/denisova/ and as a public data set via Amazon Web Services (AWS) at http://aws.amazon.com/datasets/2357. The present-day human sequences are available from the Short Read Archive under accession SRA047577. We thank D. Falush, P. Johnson, J. Krause, M. Lachmann, S. A. Aximu, B. Höber, B. Höffner, A. Weihmann, T. Kratzer, and R. Roesch for expert technical assistance; R. Schultz for help with data management; and M. Schreiber for improvement of graphics. The Presidential Innovation Fund of the Max Planck Society made this project possible. The NSF provided a HOMINID grant #1032255 to N.P. and D.R. P.H.S. is supported by an HHMI scholar. E.E.E. is on the scientific advisory boards for Pacific Biosciences, Inc., SynapDx Corp, and DNAnexus, Inc.

Supplementary Materials
www.sciencemag.org/cgi/content/full/science.1224344/DC1
Materials and Methods
Figs. S1 to S38
Tables S1 to S58
Materials and Methods
Figs. S1 to S38
Tables S1 to S58
References (29–196)

7 May 2012; accepted 14 August 2012
Published online 30 August 2012; 10.1126/science.1224344