# Targeted Investigation of the Neandertal Genome by Array-Based Sequence Capture

Hernán A. Burbano,[1]* Emily Hodges,[2,3]* Richard E. Green,[1]† Adrian W. Briggs,[1] Johannes Krause,[1] Matthias Meyer,[1] Jeffrey M. Good,[1,4] Tomislav Maricic,[1] Philip L. F. Johnson,[5] Zhenyu Xuan,[2]‡ Michelle Rooks,[2,3] Arindam Bhattacharjee,[6] Leonardo Brizuela,[6] Frank W. Albert,[1] Marco de la Rasilla,[7] Javier Fortea,[7]§ Antonio Rosas,[8] Michael Lachmann,[1] Gregory J. Hannon,[2,3] Svante Pääbo[1]

It is now possible to perform whole-genome shotgun sequencing as well as capture of specific genomic regions for extinct organisms. However, targeted resequencing of large parts of nuclear genomes has yet to be demonstrated for ancient DNA. Here we show that hybridization capture on microarrays can successfully recover more than a megabase of target regions from Neandertal DNA even in the presence of ~99.8% microbial DNA. Using this approach, we have sequenced ~14,000 protein-coding positions inferred to have changed on the human lineage since the last common ancestor shared with chimpanzees. By generating the sequence of one Neandertal and 50 present-day humans at these positions, we have identified 88 amino acid substitutions that have become fixed in humans since our divergence from the Neandertals.

The fossil record provides a rough chronological overview of the major phenotypic changes during human evolution. However, the underlying genetic bases for most of these events remain elusive. This is partly because it is not known when most human-specific genetic changes, identified from genome comparisons to living relatives, occurred during the ~6.5 million years since the separation of the human and chimpanzee evolutionary lineages. However, shotgun sequencing of the Neandertal, a human form whose ancestors split from modern human ancestors 270,000 to 440,000 years ago, has been performed to ~1.3-fold coverage of the entire genome (*1*). Comparison of Neandertal and present-day human genomes can reveal information about whether genetic changes occurred before or after the ancestral population split of modern humans and Neandertals. However, low-coverage whole-genome shotgun sequencing inevitably leaves a substantial proportion of the genome uncovered. Although deeper shotgun sequencing of one or a few individuals may produce higher coverage across the whole genome, simple shotgun approaches cannot economically retrieve specific loci from multiple individuals, both due to the size of the mammalian genome per se and to the very high proportion (up to 99.9%) of microbial DNA in the vast majority of ancient tissue remains, with the exception of some instances of preservation in permafrost (*2*, *3*). Primer extension capture can isolate specific DNA sequences from multiple Neandertal individuals (*4*). However, although useful for capture of small target regions such as mitochondrial DNA (mtDNA) (*4*, *5*), this method is unlikely to be scalable up to megabase target regions, ruling out experiments such as the retrieval of exomes, large chromosomal regions, or validation of sites of interest identified in the low-coverage shotgun genome data.

Because microarrays can carry hundreds of thousands of probes, we investigated the use of massively parallel hybridization capture on glass slide microarrays (*6*, *7*) on Neandertal DNA at thousands of genomic positions where nucleotide substitutions changing amino acids (nonsynonymous substitutions) have occurred on the human lineage since its split from chimpanzees. For any substitution that is fixed, i.e., occurs in all present-day humans, it is currently impossible to judge how long ago either the original mutation or the subsequent fixation event occurred. However, by ascertaining the Neandertal state at these positions, we can separate fixed substitutions into two classes: (i) sites where a Neandertal carries the derived state, which indicates that the substitution must have occurred before the population split of modern humans and Neandertals; and (ii) sites where a Neandertal is ancestral, which indicates that fixation of a substitution in modern humans occurred after the population split with Neandertals (Fig. 1A).

To identify substitutions that occurred on the human lineage since the ancestral split with chimpanzee, we aligned human, chimpanzee, and orangutan protein sequence for all orthologous proteins in HomoloGene (*8*, *9*). Comparison of these three species allowed us to assign human/chimpanzee differences to their respective evolutionary lineages. We designed a 1 Million Agilent oligonucleotide array covering, at 3–base pair tiling, all 13,841 nonsynonymous substitutions

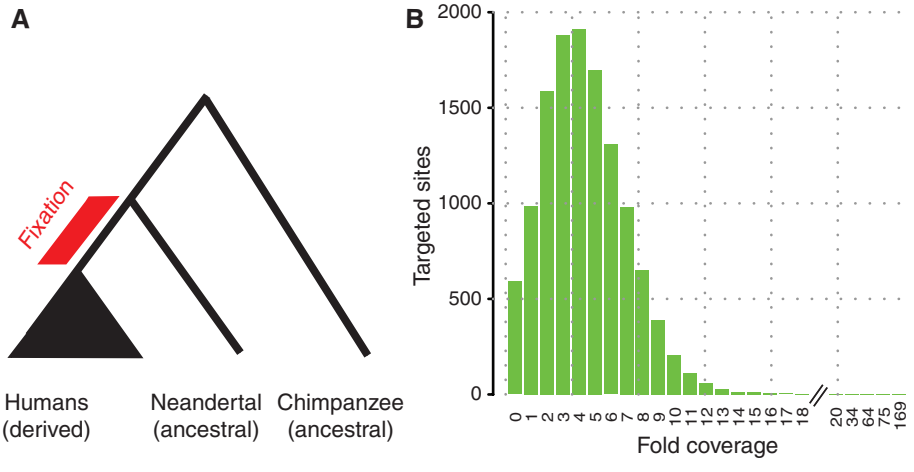


**Fig. 1.** (**A**) Identification of protein-coding changes that are likely to have become fixed recently (red bar) in modern humans after the population split from Neandertals. Such positions would be derived in all present-day humans but ancestral in the Neandertal. (**B**) Distribution of Neandertal coverage for ~14,000 amino acid substitution sites found in the human genome by comparison to primate outgroups. The same sites were also sequenced in 50 present-day humans. Of these, 88 were found to be fixed derived in present-day humans and ancestral in Neandertal, representing recently fixed protein-coding changes in the human genome.

[1]Max Planck Institute for Evolutionary Anthropology, D-04103 Leipzig, Germany. [2]Watson School of Biological Sciences, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA. [3]Howard Hughes Medical Institute, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA. [4]Division of Biological Sciences, University of Montana, Missoula, MT 59812, USA. [5]Department of Biology, Emory University, Atlanta, GA 30322, USA. [6]Agilent Technologies, Life Sciences Group, Santa Clara, CA 95051, USA. [7]Área de Prehistoria, Departamento de Historia, Universidad de Oviedo, Oviedo, Spain. [8]Departamento de Paleobiología, Museo Nacional de Ciencias Naturales, Consejo Superior de Investigaciones Científicas, Madrid, Spain.

*These authors contributed equally to this work.
†Present address: Department of Biomolecular Engineering, University of California, Santa Cruz, Santa Cruz, CA 95064, USA.
‡Present address: Department of Molecular and Cell Biology, Center for Systems Biology, University of Texas at Dallas, Richardson, TX 75080, USA.
§Deceased.

inferred to have occurred on the human lineage (*9*). We used this array to capture DNA from a ~49,000-year-old Neandertal bone (Sidrón 1253) from El Sidrón Cave, Spain (*10*, *11*). This bone contains a high amount of Neandertal DNA in absolute terms, but also a high proportion (99.8%) of microbial DNA (*4*), making it unsuitable for shotgun sequencing. To identify which of the 13,841 substitutions are fixed in present-day humans, we also collected data from 50 individuals from the Human Genome Diversity Panel (*12*) with the same array design as used for the El Sidrón Neandertal (table S1). The DNA libraries from these individuals were barcoded, pooled, and captured on a single array (*13*). All captured products were sequenced on the Illumina GAII platform and aligned to the human genome (*9*). Overall, 37% of the Neandertal sequence reads aligned to the target regions, representing ~190,000-fold target enrichment. We retrieved Neandertal sequence for 13,250 (96%) of the substitutions targeted on the array, with an average coverage of 4.8-fold after filtering for polymerase chain reaction (PCR) duplicates (Fig. 1B). We considered a Neandertal position ancestral if all overlapping reads matched the chimpanzee state and derived if all reads carried the modern human state or if we found a mixture of derived and third-state reads, disregarding positions that carried only a third state or positions where Neandertal reads were found both in the ancestral and in the derived state. From each present-day individual, a total of 25% (23 to 27%) of reads aligned to the target regions. In each individual, we retrieved on average 98% (97 to 99%) of targeted positions and had on average coverage of 10-fold (fig. S1). We estimated genotypes for each individual and considered a position to be fixed derived if it was homozygous and derived in all humans observed, and if data were available for at least 25 individuals (50 chromosomes) (*9*).

We included several additional target regions on the array to assess levels of human DNA contamination, which can frequently affect ancient DNA experiments (*14*). One such region was the complete human mtDNA, which is known to differ between the Sidrón 1253 Neandertal analyzed here and almost all (99%) present-day humans at 130 positions (*4*). Even though the array probes were designed to match present-day human mtDNA, 253,549 of the 254,296 (99.71%) fragments that overlapped these 130 positions matched the Neandertal state. We therefore conclude that the vast majority of mtDNA in the Sidrón 1253 library is of Neandertal origin.

For a more direct estimate of contamination in the nuclear DNA, we used 46 nucleotide sites on the X chromosome that differ between present-day humans and chimpanzees and that were found to be ancestral in a Neandertal from Croatia (Vindija 33.16) by shotgun sequencing (*1*), whereas ~1000 present-day humans in the human diversity panel carry a derived state. The Sidrón 1253 individual will obviously not match Vindija 33.16 at all of these sites. However, because Sidrón 1253 is a male (*15*) and thus carried a single X chromosome, at sites where he does match Vindija 33.16, all reads should carry the ancestral base while apparent heterozygosity will indicate human DNA contamination. By analyzing the consistency of reads overlapping these sites on the X chromosome, we calculated a maximum likelihood estimator of X-chromosomal contamination of 4%, although confidence intervals are large (1 to 12%) due to the small number of relevant positions (*9*).

Another way to estimate contamination across autosomes is to investigate patterns of allele counts. Because at every site an individual is either homozygous derived, homozygous ancestral, or heterozygous, DNA from a single individual will yield at each site either only derived alleles, only ancestral alleles, or a draw with equal chance for either. Contamination from other individuals would cause systematic deviation from these patterns. We thus produced a likelihood model that estimated contamination at the positions recovered from Sidrón 1253, and calculated a 95% upper bound for contamination of 2% (*9*). From these results we conclude that the Sidrón 1253 data are not substantially affected by human DNA contamination.

In total, we determined with high confidence the Neandertal and present-day human state for 10,952 nonsynonymous substitutions. In 10,015 (91.5%) of all cases the Neandertal carries the derived state, whereas in 937 (8.5%) cases the ancestral state was found (fig. S2). Of the positions that are fixed in the derived state in present-day humans, 9525 (87%) are derived in Neandertal, whereas 88 (0.8%) (table S2) are ancestral (fig. S2). In agreement with previous results generated by PCR (*15*), two substitutions that change amino acids in the gene *FOXP2* (*16*), involved in speech and language (*17*), are both derived in this Neandertal individual.

The 88 recently fixed substitutions occur in 83 genes (tables S2 and S3). We asked if these genes cluster in any group of functionally related genes relative to the genes that were targeted in the capture array (*18*) (as defined in the Gene Ontology) but found no such groups. We furthermore asked if the 88 substitutions that recently became fixed in humans differ from those that occurred before the divergence from the Neandertal with respect to how evolutionarily conserved the positions in the encoded proteins are (*9*, *19*) (Fig. 2). We found that the 88 recent substitutions tend to affect amino acid positions that are more conserved than the older substitutions (Wilcoxon rank text; $P = 0.014$). Similarly, the recently fixed substitutions caused more radical amino acid changes with respect to the chemical properties of the amino acids (Wilcoxon rank test; $P = 0.04$). One possible explanation for these observations is that the effective population size of humans since their separation from the Neandertal lineage has been small, leading to a reduced efficiency of purifying selection, as seen, e.g., in Europeans (*20*). We also looked for evidence that the recent substitutions may have been fixed by positive selection. One recent substitution occurred in *SCML1*, a gene involved in spermatogenesis (*21*) that has been previously proposed as a target of positive selection in humans (*22*) as well as frequent positive selection in primates (*23*). However, we found no significant overrepresentation of the 83 genes among candidate genes in three genome-wide scans for positive selection (*24*) (table S4). Nevertheless, we believe that all of these amino acid substitutions warrant functional studies.

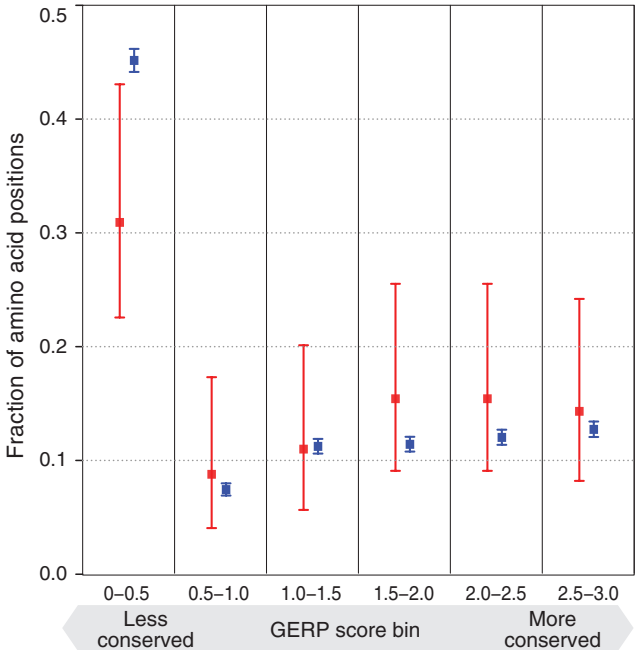


**Fig. 2.** Evolutionary conservation at positions affected by substitutions that are fixed in present-day humans. For each bin of conservation GERP (Genomic Evolutionary Rate Profiling) scores, the fractions of derived and ancestral alleles of all positions where the Neandertal carries derived (blue) and ancestral alleles (red), respectively, are given. Error bars are 95% binomial confidence intervals.

Our results demonstrate that hybridization capture arrays can generate data from genomic target regions of megabase size from ancient DNA samples, even when only ~0.2% of the DNA in a sample stems from the endogenous genome. By generating an average coverage of 4- to 5-fold, errors from sequencing and small amounts of human DNA contamination can be minimized. A further approximately 5-fold reduction of errors was achieved here by the enzymatic removal of uracil residues that are frequent in ancient DNA (*25*). Because the Sidrón 1253 Neandertal library used for this study has been amplified and effectively immortalized, the same library should be able to provide similar-quality data for any other genomic target region, or even the entire single-copy fraction of the Neandertal genome.


**References and Notes**

1. R. E. Green *et al.*, *Science* **328**, 710 (2010).
2. H. N. Poinar *et al.*, *Science* **311**, 392 (2006).
3. W. Miller *et al.*, *Nature* **456**, 387 (2008).
4. A. W. Briggs *et al.*, *Science* **325**, 318 (2009).
5. J. Krause *et al.*, *Curr. Biol.* **20**, 231 (2010).
6. E. Hodges *et al.*, *Nat. Protoc.* **4**, 960 (2009).
7. E. Hodges *et al.*, *Nat. Genet.* **39**, 1522 (2007).
8. E. W. Sayers *et al.*, *Nucleic Acids Res.* **38** (Database issue), D5 (2010).
9. Materials and methods are available as supporting material on *Science* Online.
10. A. Rosas *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 19266 (2006).
11. T. De Torres *et al.*, *Archaeometry* 10.1111/j.1475-4754.2009.00491.x (2009).
12. H. M. Cann *et al.*, *Science* **296**, 261 (2002).
13. M. Meyer, M. Kircher, *Cold Spring Harb. Protoc.* 10.1101/pdb.prot5448 (2010).
14. R. E. Green *et al.*, *EMBO J.* **28**, 2494 (2009).
15. J. Krause *et al.*, *Curr. Biol.* **17**, 1908 (2007).
16. W. Enard *et al.*, *Nature* **418**, 869 (2002).
17. F. Vargha-Khadem, D. G. Gadian, A. Copp, M. Mishkin, *Nat. Rev. Neurosci.* **6**, 131 (2005).
18. K. Prüfer *et al.*, *BMC Bioinformatics* **8**, 41 (2007).
19. G. M. Cooper *et al.*, *Genome Res.* **15**, 901 (2005).
20. K. E. Lohmueller *et al.*, *Nature* **451**, 994 (2008).
21. B. Boeckmann *et al.*, *Nucleic Acids Res.* **31**, 365 (2003).
22. C. D. Bustamante *et al.*, *Nature* **437**, 1153 (2005).
23. H. H. Wu, B. Su, *BMC Evol. Biol.* **8**, 192 (2008).
24. J. M. Akey, *Genome Res.* **19**, 711 (2009).
25. A. W. Briggs *et al.*, *Nucleic Acids Res.* **38**, e87 (2010).


26. We thank C. S. Burbano, C. de Filippo, J. Kelso, and D. Reich for helpful comments; M. Kircher, K. Pruefer, and U. Stenzel for technical support; C. D. Bustamante and K. E. Lohmueller for access to human resequencing databases; D. L. Goode and A. Sidow for providing conservation scores; J. M. Akey for providing coordinates of genome-wide scans for selection; I. Gut for human genotyping; E. Leproust and M. Srinivasan for providing early access to the 1 Million feature Agilent microarrays; and the Genome Center at Washington University for pre-publication use of the orangutan genome assembly (http://genome.wustl.edu/genomes/view/pongo_abelii/). The government of the Principado de Asturias funded excavations at the Sidrón site. J.M.G. was supported by an NSF international postdoctoral fellowship (OISE-0754461) and E.H. by a postdoctoral training grant from the NIH and by a gift from the Stanley Foundation. G.J.H. is an investigator of the Howard Hughes Medical Institute, which together with the Presidential Innovation Fund of the Max Planck Society provided generous financial support. DNA sequences are deposited in the European Bioinformatics Institute short read archive, with accession number ERP000125. The array capture technologies used in this study are the subject of pending patent filings U.S. 60/478, 382 (filed 2003) and U.S. 61/205, 834 (filed 2009), on which G.J.H. and E.H. are listed as inventors.


**Supporting Online Material**

www.sciencemag.org/cgi/content/full/328/5979/723/DC1
Materials and Methods
Figs. S1 to S4
Tables S1 to S5


8 February 2010; accepted 1 April 2010
10.1126/science.1188046


# Fermi Gamma-Ray Imaging of a Radio Galaxy


**The Fermi-LAT Collaboration*†**



**The Fermi Gamma-ray Space Telescope has detected the γ-ray glow emanating from the giant radio lobes of the radio galaxy Centaurus A. The resolved γ-ray image shows the lobes clearly separated from the central active source. In contrast to all other active galaxies detected so far in high-energy γ-rays, the lobe flux constitutes a considerable portion (greater than one-half) of the total source emission. The γ-ray emission from the lobes is interpreted as inverse Compton–scattered relic radiation from the cosmic microwave background, with additional contribution at higher energies from the infrared-to-optical extragalactic background light. These measurements provide γ-ray constraints on the magnetic field and particle energy content in radio galaxy lobes, as well as a promising method to probe the cosmic relic photon fields.**


Centaurus A (Cen A) is one of the brightest radio sources in the sky and was among the first identified with a galaxy (NGC 5128) outside of our Milky Way (*1*). Straddling the bright central source is a pair of extended radio lobes with a total angular extent of ~10° (*2*, *3*), which makes Cen A the largest discrete nonthermal extragalactic radio source visible from Earth. At a distance of 3.7 Mpc (*4*), it is the nearest radio galaxy to Earth, and the implied physical source size is ~600 kpc. Such double-lobed radio structures associated with otherwise apparently normal giant elliptical galaxies have become the defining feature of radio galaxies in general. The consensus explanation for this phenomenon is that the lobes are fueled by relativistic jets produced by accretion activity in a super-massive black hole residing at the galaxy's center.

With its unprecedented sensitivity and imaging capability (per-photon resolution: $\theta_{68} \simeq 0^{\circ}.8 E_{\mathrm{GeV}}^{-0.8}$), the Fermi Large Area Telescope (LAT) (*5*) has detected and imaged the radio lobes of Cen A in high-energy γ-rays. The LAT image resulting from ~10 months of all-sky survey data (Fig. 1) clearly shows the γ-ray peak coincident with the active galactic nucleus detected by the Compton/EGRET instrument (*6*) and extended emission from the southern giant lobe. Because the northern lobe is characterized by lower surface-brightness emission (in radio), it is not immediately apparent from a naked-eye inspection of the γ-ray counts map. Nevertheless, from a counts profile extracted along the north-south axis of the source (Fig. 2), γ-ray excesses from both lobes are clearly visible.

Spectra for each of the lobes together with the central source (hereafter referred to as the "core") were determined with a binned maximum likelihood analysis implemented in GTLIKE (*7*) using events from 0.2 to 30 GeV in equal logarithmically spaced energy bins. We modeled background emission by including the Galactic diffuse component, an isotropic component, and nearby γ-ray point sources [see the supporting online material (SOM)]. We fit the core as a point source at the known radio position and modeled the lobe emission with a 22-GHz Wilkinson Microwave Anisotropy Probe (WMAP) image (Fig. 1) (*8*) with the core region within a 1° radius excluded as a spatial template. The modeled lobe region roughly corresponds to the regions 1 and 2 (north) and 4 and 5 (south) defined in (*9*), where region 3 is the core (Fig. 2). Assuming a power law for the γ-ray spectra, we find a large fraction (>1/2) of the total >100-MeV emission from Cen A to originate from the lobes with the flux in each of the northern $\{[0.77(+0.23/-0.19)_{\mathrm{stat.}}(\pm 0.39)_{\mathrm{syst.}}] \times 10^{-7}$ ph cm$^{-2}$ s$^{-1}\}$ and southern $\{[1.09(+0.24/-0.21)_{\mathrm{stat.}}(\pm 0.32)_{\mathrm{syst.}}] \times 10^{-7}$ ph cm$^{-2}$ s$^{-1}\}$ lobes smaller than the core flux $\{[1.50(+0.25/-0.22)_{\mathrm{stat.}}(\pm 0.37)_{\mathrm{syst.}}] \times 10^{-7}$ ph cm$^{-2}$ s$^{-1}\}$ (stat., statistical; syst., systematic). Uncertainties in the LAT effective area, the Galactic diffuse model used, and the core exclusion region were considered to be sources of systematic error (SOM). The resultant test statistic (*10*) for the northern and southern giant lobes are 29 and 69, which correspond to detection significances of 5.0σ and 8.0σ, respectively. The lobe spectra are steep, with photon indices $\Gamma = 2.52(+0.16/-0.19)_{\mathrm{stat.}}(\pm 0.25)_{\mathrm{syst.}}$ (north) and $2.60(+0.14/-0.15)_{\mathrm{stat.}}(\pm 0.20)_{\mathrm{syst.}}$ (south) in which photons up to ~2 to 3 GeV are currently detected. These values are consistent with that of the core $[\Gamma = 2.67(\pm 0.10)_{\mathrm{stat.}}(\pm 0.08)_{\mathrm{syst.}}]$, which


*All authors with their affiliations appear at the end of this paper.
†To whom correspondence should be addressed. E-mail: Teddy.Cheung.ctr@nrl.navy.mil (C.C.C.); fukazawa@hep01.hepl.hiroshima-u.ac.jp (Y.F.); jurgen.knodlseder@cesr.fr (J.K.); stawarz@slac.stanford.edu (Ł.S.)

# Targeted Investigation of the Neandertal Genome by Array-Based Sequence Capture

Hernán A. Burbano *et al.*

*Science* **328**, 723 (2010);
DOI: 10.1126/science.1188046