Published in final edited form as:
Science. 2010 May 7; 328(5979): 723–725. doi:10.1126/science.1188046.

Timing of human protein evolution as revealed by massively parallel capture of Neandertal nuclear DNA sequences

Hernán A. Burbano[1,*], **Emily Hodges**[2,3,*], **Richard E. Green**[1,8], **Adrian W. Briggs**[1], **Johannes Krause**[1], **Matthias Meyer**[1], **Jeffrey M. Good**[1,9], **Tomislav Maricic**[1], **Philipp L.F. Johnson**[4], **Zhenyu Xuan**[2], **Michelle Rooks**[2,3], **Arindam Bhattacharjee**[5], **Leonardo Brizuela**[5], **Frank W. Albert**[1], **Marco de la Rasilla**[6], **Javier Fortea**[6,10], **Antonio Rosas**[7], **Michael Lachmann**[1], **Gregory J. Hannon**[2,3], and **Svante Pääbo**[1]

[1] Max Planck Institute for Evolutionary Anthropology, D-04103 Leipzig, Germany

[2] Watson School of Biological Sciences, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA

[3] Howard Hugues Medical Institute, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA

[4] Department of Biology, Emory University, Atlanta, GA

[5] Agilent Technologies Inc., Life Sciences Group, Santa Clara, California 95051, USA

[6] Área de Prehistoria, Departamento de Historia, Universidad de Oviedo, Oviedo, Spain

[7] Departamento de Paleobiología, Museo Nacional de Ciencias Naturales, Consejo Superior de Investigaciones Científicas, Madrid, Spain

Abstract

Whole genome shotgun sequencing is now possible for extinct organisms, as well as the targeted capture of specific regions. However, targeted resequencing of megabase sized parts of nuclear genomes has yet to be demonstrated for ancient DNA. Here we show that hybridization capture on microarrays can be used to generate large scale targeted data from Neandertal DNA even in the presence of ~99.8% microbial DNA. It is thus now possible to generate high quality data from large regions of the nuclear genome from Neandertals and other extinct organisms. Using this approach we have sequenced ~14,000 protein coding positions that have been inferred to have changed on the human lineage since the last common ancestor shared with chimpanzees. We identify 88 amino acid substitutions that have become fixed in all humans since the 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 cannot currently be determined 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, recent technological progress has allowed shotgun sequencing to ~1.3-fold coverage of the entire genome of the Neandertal, a human form whose ancestors split from modern human ancestors 200,000 to 330,000 years ago (1).

[8]Current address: Department of Biomolecular Engineering, University of California, Santa Cruz, USA.
[9]Current address: Division of Biological Sciences, University of Montana, Missoula, USA.
[10]Deceased.
[*]These authors contributed equally to this work.

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. While 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, due to the very high proportion (up to 99.9%) of microbial DNA in most ancient bones. Recently we developed a primer extension capture method to isolate specific DNA sequences from multiple Neandertal individuals (2). However, while useful for capture of small target regions such as mtDNA (2,3), 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.

Massively parallel hybridization capture on glass slide microarrays (4,5) might be a suitable way to capture large regions of the genome from Neandertal individuals, since microarrays can carry hundreds of thousands of probes. Here we demonstrate the power of array capture by isolating Neandertal DNA at thousands of genomic positions where nucleotide substitutions changing amino acids (non-synonymous substitutions) have occurred on the human lineage since its split from chimpanzee. For any such substitution that is fixed, i.e. occur 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: a) sites where Neandertal carries the derived state; this indicates that the *substitution* must have occurred before the population split of modern humans and Neandertals ~300,000 years ago (Fig 1a); and b) sites where Neandertal is ancestral; this indicates that *fixation* of a substitution in modern humans occurred after the population split with Neandertals (Fig 1b).

To identify non-synonymous 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 (6) [Supporting online material (SOM)]. 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 to cover, at 3bp tiling, all the 13,841 substitutions that we inferred to have occurred on the human lineage using the reference human genome. The targeted regions amounted to 1.3Mb (SOM). We used this array to capture target sequences from DNA of a ~49,000-year-old Neandertal bone (Sidron 1253) from El Sidron cave, Spain (7). This bone contains a high amount of Neandertal DNA in absolute terms, but also a high proportion (99.8%) of microbial DNA (2), 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 (HGDP) using the same array design used for the Neandertal (Table S1). The DNA libraries from these individuals were barcoded, pooled and captured on a single array. All captured products were sequenced on the Illumina GAII platform and aligned to the human genome (SOM). Overall 37% of the sequence reads aligned to the target regions, representing ~ 190,000-fold target enrichment. We retrieved Neandertal sequence for 13,287 (96%) of the substitutions targeted on the array, with an average coverage of 4.8-fold after filtering for PCR duplicates (Fig. 1c–d). We considered a Neandertal position ancestral if all overlapping reads matched the chimpanzee state and derived if all reads carried the modern human state, disregarding reads that carried a third state or positions where Neandertal reads were found both in the ancestral and in the derived state. From each present-day individual we retrieved on average of 98% (97–99%) of positions with a targeted substitution with an average coverage of 10-fold (Fig. S1). We called genotypes for each individual and considered a position to be fixed derived if

it was homozygous and derived in all humans observed, and data was available for at least 25 individuals (50 chromosomes) (SOM).

We included several additional target regions on the array to assess levels of human DNA contamination, which can frequently affect ancient DNA experiments (8). One such region was the complete human mtDNA, which is known to differ between the Sidron 1253 Neandertal analyzed here and present-day humans at 130 positions (2). Despite the fact that 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 Sidron 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) (9) by shotgun sequencing while ~1000 present-day humans in the human diversity panel carry a derived state. The Sidron 1253 individual will obviously not match Vindija 33.16 at all of these sites. However, since Sidron 1253 is a male (10) 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 (1–12%) due to the small number of relevant positions (SOM).

Another way to estimate contamination across autosomes is to investigate patterns of allele counts. Since 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. Using this idea we produced a likelihood model that estimated contamination at the positions recovered from Sidron 1253, and calculated a 95% upper bound for contamination of 2% (SOM). From these results we conclude that the Sidron 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 10,592 non-synonymous 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, 9,525 (87%) are derived in Neandertal while 88 (0.8%) (Table S2) are ancestral (Fig. S2). In agreement with previous results generated by PCR (10), two substitutions that change amino acids in the gene *FOXP2*, involved in speech and language (11), are both derived in this Neandertal individual.

The 88 recently fixed substitutions occur in 83 genes (Table S2–S3). We asked if these genes cluster in any group of functionally related genes (12) (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 where the substitution occurred before the divergence from the Neandertal with respect to how evolutionarily conserved the positions in the encoded proteins are (13) (SOM) (Fig. 2). We find that the 88 recent substitutions tend to affect more amino acid positions that are more conserved than the older substitutions (Wilcoxon rank text; p-value: 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-value: 0.04). One possible explanation for these observations is that the effective population size of humans since their separation form the Neandertal lineage has been small, leading to a reduced efficiency of purifying selection, as seen *e.g*. in Europeans (14). We also looked for evidence that the recent substitutions may have been fixed by positive selection. One recent substitution lies in SCML1, a gene involved in spermatogenesis (15) that has been previously proposed as a target of positive selection in humans (16) as well as frequent positive selection in primates (17). However, we found no significant overrepresentation of the 83 genes containing recent substitutions among candidate genes from three genome wide scans for positive selection (18) (Table S4). Nevertheless, we believe that many or all of these amino acid substitutions may warrant functional studies in the future in order to elucidate if they have functional consequences *e.g*. (19).

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, especially in conjunction with the enzymatic removal of uracil residues that are frequent in ancient DNA (20). Since the Sidron 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Claudia S. Burbano, Cesare de Filippo, Janet Kelso, and David Reich for helpful comments; Martin Kircher and Udo Stenzel for expert technical support; Carlos D. Bustamante and Kirk E. Lohmueller for access to human resequencing databases; David L. Goode and Arendt Sidow for providing conservation scores; Joshua M. Akey for providing coordinates of genome wide scans for selection; Ivo Gut for human genotyping; Emily Leproust and Maethreyan Srinivasan for providing early access to the 1 Million feature Agilent microarrays. The government of the Principado de Asturias funded excavations at the Sidron site. JMG was supported during this research by an NSF international postdoctoral fellowship (OISE-0754461). EH is supported by a postdoctoral training grant from the NIH and by a gift from the Stanley foundation. GJH 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.

References

1. Green RE, et al. 2010 Submited Manuscript.
2. Briggs AW, et al. Science. Jul 17.2009 325:318. [PubMed: 19608918]
3. Krause J, et al. Curr Biol. Dec 30.2009
4. Hodges E, et al. Nat Protoc. 2009; 4:960. [PubMed: 19478811]
5. Hodges E, et al. Nat Genet. Dec.2007 39:1522. [PubMed: 17982454]
6. Sayers EW, et al. Nucleic Acids Res. Jan.2010 38:D5. [PubMed: 19910364]
7. Rosas A, et al. Proc Natl Acad Sci U S A. Dec 19.2006 103:19266. [PubMed: 17164326]
8. Green RE, et al. EMBO J. Sep 2.2009 28:2494. [PubMed: 19661919]
9. Malez M, Ullrich H. Palaeontologia Jugoslavia. 1982; 29:1.
10. Krause J, et al. Curr Biol. Nov 6.2007 17:1908. [PubMed: 17949978]
11. Vargha-Khadem F, Gadian DG, Copp A, Mishkin M. Nat Rev Neurosci. Feb.2005 6:131. [PubMed: 15685218]

12. Prufer K, et al. BMC Bioinformatics. 2007; 8:41. [PubMed: 17284313]
13. Cooper GM, et al. Genome Res. Jul.2005 15:901. [PubMed: 15965027]
14. Lohmueller KE, et al. Nature. Feb 21.2008 451:994. [PubMed: 18288194]
15. Boeckmann B, et al. Nucleic Acids Res. Jan 1.2003 31:365. [PubMed: 12520024]
16. Bustamante CD, et al. Nature. Oct 20.2005 437:1153. [PubMed: 16237444]
17. Wu HH, Su B. BMC Evol Biol. 2008; 8:192. [PubMed: 18601738]
18. Akey JM. Genome Res. May.2009 19:711. [PubMed: 19411596]
19. Gralle M, Paabo S. 2010 Submited Manuscript.
20. Briggs AW, et al. Nucleic Acids Res. Dec 22.2009

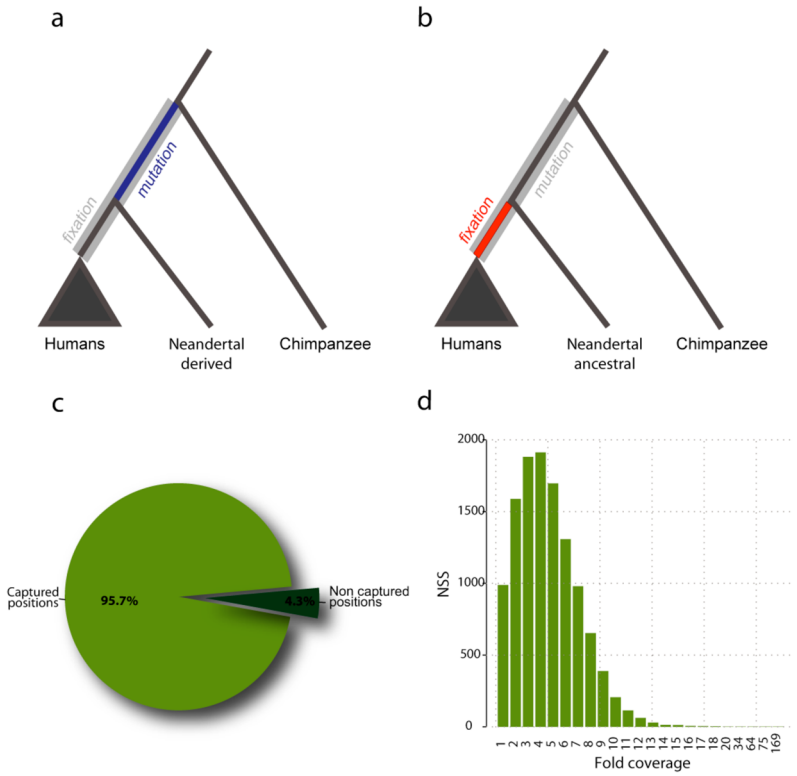

Figure 1.

(a) For the 9,525 substitutions where the Neandertal carries the derived human allele, the substitution has occurred prior the Neandertal-modern human split (blue part of the hominin lineage), whereas fixation could have happen any time later (grey part of the hominin lineage). **(b)** For the 88 substitutions where the Neandertal carries the ancestral, chimpanzee-like base, the fixation occurred after the Neandertal-modern human split (red part of the hominin lineage), whereas the substitution could have happened at any point along the hominin lineage (grey). **(c)** Percentage of the 13,841 positions captured. **(d)** Distribution of coverage for the positions that encode amino acid substitutions along the hominin lineage. The average coverage is 4.8-fold.

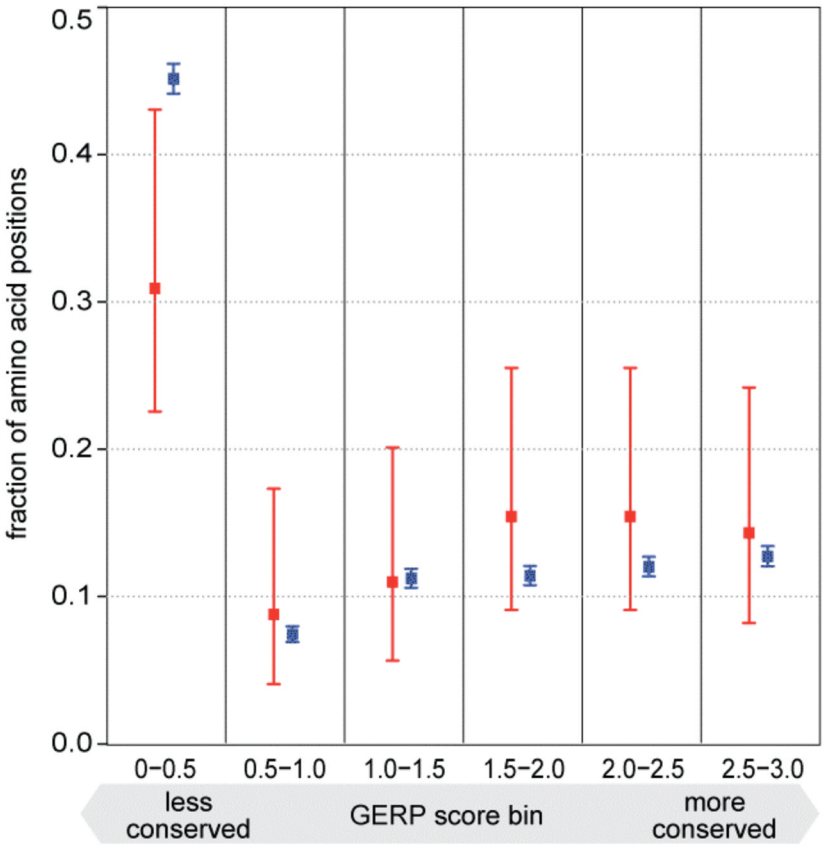

Figure 2.
Evolutionary protein conservation at positions affected by the substitutions studied. For each bin of conservation GERP scores, the fractions of amino acid substitutions where the Neandertal carry the derived allele (blue) and the ancestral allele (red) are shown.