Whole-Genome Shotgun Sequencing of Mitochondria from Ancient Hair Shafts

M. Thomas P. Gilbert, *et al.*

Science **317**, 1927 (2007);
DOI: 10.1126/science.1146971

Whole-Genome Shotgun Sequencing of Mitochondria from Ancient Hair Shafts

M. Thomas P. Gilbert,[1]* Lynn P. Tomsho,[2] Snjezana Rendulic,[2] Michael Packard,[2] Daniela I. Drautz,[2] Andrei Sher,[3] Alexei Tikhonov,[4] Love Dalén,[5] Tatyana Kuznetsova,[6] Pavel Kosintsev,[7] Paula F. Campos,[1] Thomas Higham,[8] Matthew J. Collins,[9] Andrew S. Wilson,[10] Fyodor Shidlovskiy,[11] Bernard Buigues,[12] Per G. P. Ericson,[13] Mietje Germonpré,[14] Anders Götherström,[15] Paola Iacumin,[16] Vladimir Nikolaev,[17] Malgosia Nowak-Kemp,[18] Eske Willerslev,[1] James R. Knight,[19] Gerard P. Irzyk,[19] Clotilde S. Perbost,[19] Karin M. Fredrikson,[20] Timothy T. Harkins,[20] Sharon Sheridan,[20] Webb Miller,[2]*† Stephan C. Schuster[2]*†

Although the application of sequencing-by-synthesis techniques to DNA extracted from bones has revolutionized the study of ancient DNA, it has been plagued by large fractions of contaminating environmental DNA. The genetic analyses of hair shafts could be a solution: We present 10 previously unexamined Siberian mammoth (*Mammuthus primigenius*) mitochondrial genomes, sequenced with up to 48-fold coverage. The observed levels of damage-derived sequencing errors were lower than those observed in previously published frozen bone samples, even though one of the specimens was >50,000 ^{14}C years old and another had been stored for 200 years at room temperature. The method therefore sets the stage for molecular-genetic analysis of museum collections.

Short fragments of mitochondrial DNA (mtDNA) have been the predominant genetic marker applied to phylogenetic and population-genetic studies of ancient samples (*1*–*3*). Although the use of complete mitochondrial genomes would provide greater analytical power, the degraded state of ancient DNA (aDNA) has prevented recovery and assembly of the full genome by conventional genetic methods. Although aDNA has been applied to phylogenetic questions for more than 20 years (*4*), only six complete mitochondrial genomes from ancient samples have been explicitly published: four from extinct moa species—*Emeus crassus* (two genomes), *Anomalopteryx didiformis*, and *Dinornis giganticus* (*5*, *6*)—and two from extinct woolly mammoth (*Mammuthus primigenius*) specimens (*7*, *8*).

Despite the field's slow start, recent developments in DNA amplification, sequencing, and analysis technologies have begun to revolutionize aDNA research, enabling the application of whole-genome shotgun sequencing approaches to a variety of aDNA sources. Recent applications of such approaches have demonstrated that nuclear DNA sequence (nuDNA), in addition to mtDNA, can be recovered and analyzed. For example, Noonan *et al*. (*9*) obtained more than 25,000 base pairs (bp) of nuDNA from a 40,000-^{14}C-year-old cave bear (*Ursus spelaeus*) bone. Using the recently developed sequencing-by-synthesis (SBS) technology (*10*), Poinar *et al*. (*11*) determined 13 million bp (Mbp) of nuDNA from a 28,000-^{14}C-year-old mammoth bone. The success of this study rapidly paved the way for application of SBS to extinct hominid samples and resulted in 1 Mbp of nuDNA from Neandertal bones (*12*, *13*). These reports have set the stage for a new era in aDNA research, but difficult challenges remain. For example, only one of these studies—the one that used exceptionally well-preserved frozen mammoth bone (*11*)—yielded sufficient quantities of endogenous DNA (i.e., DNA derived from the host and not bacterial, human, or other external contaminants) to make it economical to sequence entire nuclear genomes of extinct species.

Hair shafts are a promising source of aDNA. Long-term hair survival occurs in a variety of natural environments, and large quantities are present in taxonomic collections representing most extant, and many recently extinct, mammalian taxa. Most hair-based genetic studies have used roots, instead of shafts, as a DNA source (*14*), primarily because hair shafts comprise dead keratinized cells that contain relatively low levels of DNA. However, several studies have reported shafts as a viable source of modern (*15*) and ancient (*16*) mtDNA. Furthermore, several properties of shafts suggest that they constitute an attractive DNA source for SBS. First, their relative abundance (when present) renders them preferable to bones, because the destructive nature of sampling can lead to the loss of important morphological information. Second, turnover of keratinocytes in the hair bulb is exceedingly high, second only to that of the cells of the gut epithelium (*17*). Therefore, baseline mitochondrial levels in these cells (and thus the precortical cells that develop into the bulk of the shaft) may be higher than those in other tissues commonly used for aDNA analyses. Third, even when degraded, shafts are resistant to contamination from exogenous DNA such as bacteria, blood, and skin cells (*16*, *18*). We demonstrate here that hair shafts surpass comparably stored bone as an aDNA source for use in SBS approaches, in regard to preservation and concentration of mtDNA.

We successfully extracted sufficient DNA for SBS from 10 samples of mammoth coat-hair shafts, collected from permafrost deposits spanning northern Siberia [Table 1, Fig. 1, and supporting online material (SOM) text]. Due to the pilot nature of this study, we used as much hair as was readily available (0.2 to 5.2 g per extraction). The degradation of aDNA correlates exponentially with temperature (*19*), thus DNA survival depends on sample age and the storage history (including the time and temperature at which it has been stored pre- and postcollection). Surprisingly, we successfully extracted DNA from the sample (M13) that had been at room temperature for the longest period and that had the lowest amount of material available [0.2 g, in comparison to 0.75 and 1 g bone (*7*, *11*) and up to 0.4 g frozen muscle (*8*) used in the previous studies]. Although hair morphology varies significantly both between species (*20*), and among hair types on individuals, and thus the general applicability of this method remains to be shown, previous studies have demonstrated successful recovery of DNA from a variety of modern hair types and species (SOM text). Thus, this method will likely be widely applicable.

The combined use of hair shafts and SBS resulted in 10 full mitochondrial genome sequences, with 7.3- to 48.0-fold coverage (Table 1). The sequences are complete, except that we have not tried to assemble the variable number of tandem repeats (VNTR), which is

[1]Centre for Ancient Genetics, University of Copenhagen, Universitetsparken 15, DK-2100 Copenhagen, Denmark. [2]Center for Comparative Genomics and Bioinformatics, Pennsylvania State University, 310 Wartik Building, University Park, PA 16802, USA. [3]Severtsov Institute of Ecology and Evolution, Russian Academy of Sciences, 33 Leninsky Prospect, Moscow 119071, Russia. [4]Zoological Institute, Russian Academy of Sciences, Universitetskaya Naberezhnaya, St. Petersburg 199034, Russia. [5]Centro UCM-ISCIII de Evolución y Comportamiento Humanos, c/Sinesio Delgado 4, 28029 Madrid, Spain. [6]Department of Paleontology, Faculty of Geology, Lomonosov Moscow State University, Leninskiye Gory, Moscow 119992, Russia. [7]Institute of Plant and Animal Ecology, The Urals Branch of the Russian Academy of Sciences, 202 8th of March Street, Ekaterinburg 620144, Russia. [8]Research Laboratory for Archaeology and the History of Art, Dyson Perrins Building, South Parks Road, Oxford OX1 3QY, UK. [9]Departments of Biology and Archaeology, BioArch, University of York, York YO10 5YW, UK. [10]Department of Archaeological Sciences, University of Bradford, Bradford BD7 1DP, UK. [11]The Ice Age Museum, All-Russia Exhibition Centre, pavilion 71, Moscow 129223, Russia. [12]2 Avenue de la Pelouse, F-94160 Saint Mandé, France. [13]Department of Vertebrate Zoology, Swedish Museum of Natural History, Post Office Box 50007, S-10405, Stockholm, Sweden. [14]Department of Palaeontology, Royal Belgian Institute of Natural Sciences, Vautierstraat 29, 1000 Brussels, Belgium. [15]Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, SE-752 36 Uppsala, Sweden. [16]Department of Earth Sciences, University of Parma, Parco Area delle Scienze 157/A, 43100 Parma, Italy. [17]Department of Glaciology, Institute of Geography, Russian Academy of Science, 29 Staromonetny Pereulok, Moscow 109017, Russia. [18]Oxford University Museum of Natural History, Parks Road, Oxford OX1 3PW, UK. [19]454 Life Sciences, 20 Commercial Street, Branford, CT 06405, USA. [20]Roche Diagnostics Corporation, 9115 Hague Road, Indianapolis, IN 46250–0414, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: webb@bx.psu.edu (W.M.); scs@bx.psu.edu (S.C.S.)

Table 1. Description of mammoth mitochondrial sequences, including the year that the sample was discovered, where known; the ^{14}C reference of specimens dated in this study; the percentage of mitochondrial sequences among SBS sequences; the number of contigs assembled out of mitochondrial sequences; the average read length before trimming, based on Krause (*7*) sequence; the average percentage identity with respect to assembly after automatic computational quality processing (i.e., the final read used in alignment); and the percentage difference from M1 sequence. nd, not determined.

Sample	Tissue	^{14}C date	Year collected	OxA ^{14}C reference	Sequencing technology	% Mitochondrial	Contigs	Fold coverage	Average untrimmed read length	% C→T damage	% Trimmed read identity	% Diff vs. M1
M1	Hair	nd	nd		454	1.99	1	48.0	119.0	0.243	99.86	0.00
M2 (Jarkov)	Hair	20,380 ± 140	1997		454	0.76	1	13.2	99.5	0.427	99.79	0.09
M3 (Fishhook)	Hair	20,620 ± 70	1990		454	1.21	1	20.3	128.1	0.347	99.82	0.03
M4	Hair	18,545 ± 70	nd	17098	454	0.65	1	7.8	75.9	0.314	99.85	0.16
M5	Hair	nd	nd		454	1.30	1	19.8	112.6	0.537	99.76	0.15
M8 (Dima)	Hair	46,900 ± 700	1977	17102	454	1.43	1	27.7	71.1	0.899	99.60	0.14
M13 (Adams)	Hair	35,800 ± 1200	1799/1806		454	0.76	1	19.1	60.5	0.713	99.73	0.09
M18	Hair	17,125 ± 70	nd	17116	454	1.50	1	24.6	129.8	0.388	99.83	0.05
M22	Hair	50,200 ± 900	2000	17111	454	2.09	1	17.0	96.8	0.556	99.72	0.17
M26	Hair	24,740 ± 110	2003	17114	454	0.46	1	7.3	91.2	0.253	99.84	0.22
Poinar	Bone	27,740 ± 220	2005		454	0.08	2	7.7	101.1*	1.699	99.51	0.09
Rogaev	Muscle	33,750 to 31,950			Sanger	nd	nd	nd	nd	nd	nd	0.21
Krause	Bone	12,170 ± 50			Sanger	nd	nd	nd	nd	nd	nd	0.13

*The sequencing technology (Roche GS 20) used for the generation of the Poinar (*11*) sequence precluded obtaining longer average read length.

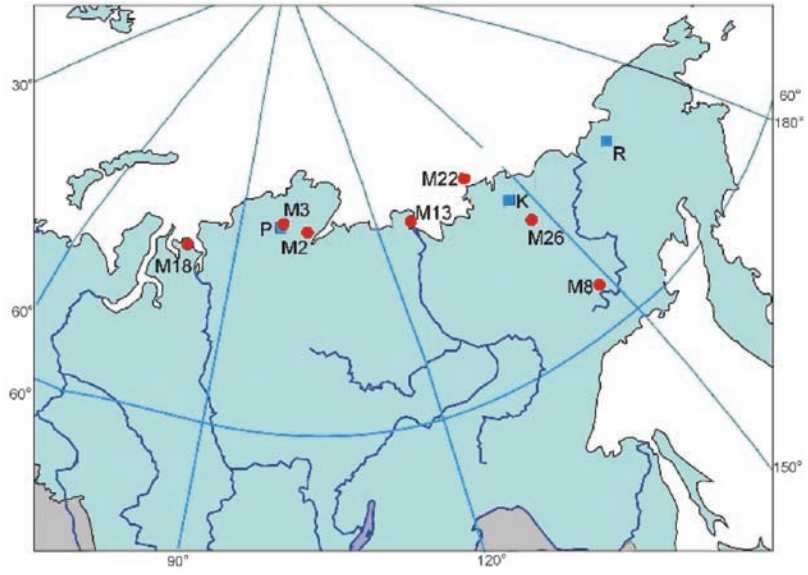

Fig. 1. Sites of recovery of the mammoth hair specimens whose mitochondrial genome sequences are reported here. The locations of M1, M4, and M5 are not known, but most probably originate from Northern Yakutia (about 66° to 76°N, 106° to 160°E). Recovery sites for other mitochondrial genomes used in this study—Krause (*7*), Rogaev (*8*), and Poinar (*11*, *21*)—are indicated as blue squares labeled K, R, and P, respectively.

difficult to sequence [even with polymerase chain reaction (PCR) and sequencing (*8*)] or to align with any certainty. For example, in the sequence of Krause *et al.* (*7*), this region of the mammoth mitochondrial genome is 320 bp, whereas it is 393 bp in the sequence of Rogaev *et al.* (*8*), so comparison of these regions is essentially uninformative. Overall, the yield of mtDNA sequence was 5.75 to 26 times as high as that from the permafrost-preserved bone reported previously (*11*, *21*), supporting previous hypotheses that in comparison to bone, the ratio of mtDNA to nuDNA in the hair shaft is elevated (*16*, *22*).

Three widely recognized difficulties are associated with sequencing aDNA: DNA damage, sequencing errors, and numts. Numts are mitochondrial sequences that were inserted into the nuclear genome during genome evolution after duplication and may cause artifacts in PCR-based studies or shotgun assemblies with low coverage. Our approach solves all of these problems through the high redundancy of our sequencing and the fact that SBS targets unique, individual DNA template molecules.

We assessed the state of DNA preservation through two parameters—untrimmed read length and DNA damage [cytosine-to-thymine (C→T) miscoding lesions, derived from the hydrolytic deamination of cytosine to uracil, observed in the pyrosequencing data] (*21*, *23*). The sizes of unbroken aDNA fragments could be measured because the study was conducted on a SBS instrument (Roche GS FLX) that can generate reads up to a length of 250 bp. We observed an average sample-dependent mitochondrial read length between 60.5 and 128.1 bp. The previously described average read length of 101 bp from a bone sample (*11*) was limited by the instrument read length (Roche GS20), leaving open the possibility that the bone sample retained longer fragments of mtDNA than those that we observed. However, comparing the individual reads versus locations in the assembly consensus sequence containing C, the hair-generated data show a substantially improved (i.e., lower) C→T DNA damage rate of 0.24 to 0.9% versus 1.7% in bone. In contrast to the bone, which was kept frozen for the entire period postexcavation from the permafrost, most of the hair samples have been at room temperature for a number of years (Table 1).

To investigate what effect this might have on the DNA preservation of the samples, we calculated approximate thermal ages (*19*) of those specimens for which we knew or could estimate sufficient information for the calculation [including for comparison the Poinar mammoth (*11*)]. The model incorporated temperature data from weather stations local to the respective sites, with altitude correction (lapse rate of +6.5°C km^{-1}) that used elevations estimated from the sample coordinates (with the use of GoogleEarth v.4.1). Furthermore, to control for differences in sample burial depth (and thus temperature of the burial site), the model incorporated two depths of burial for each sample—shallow (where the sample temperature could be expected to fluctuate during the year) and deep (where the sample would experience a con-

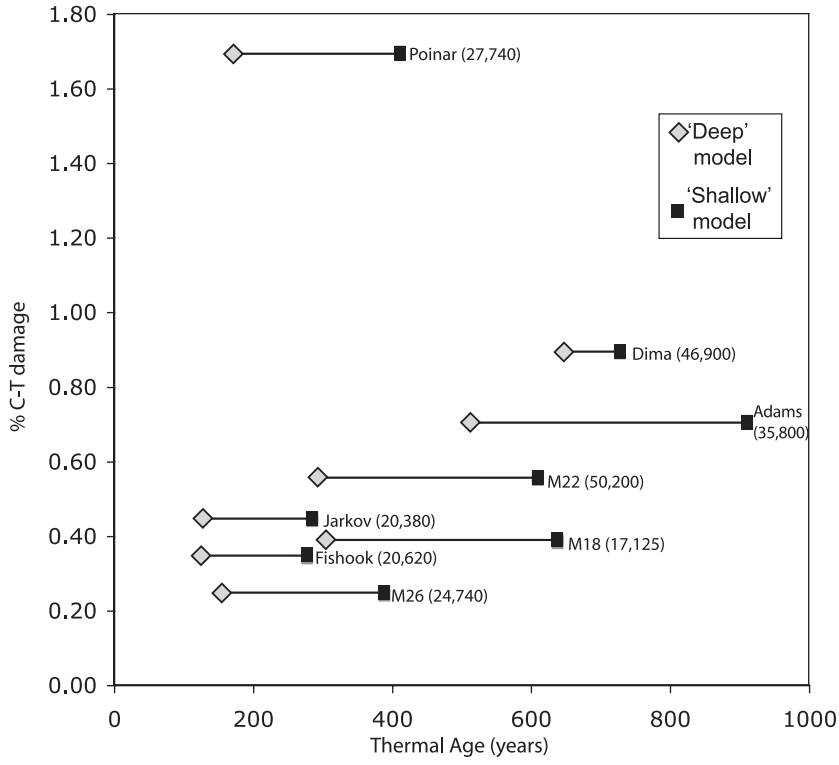

Fig. 2. Comparison of estimated thermal age of samples against percentage C-T damage with the use of alternative temperature models for Siberia (reflecting the range of published estimates). Approximate thermal ages were calculated according to the methods of Smith *et al.* (*19*) for mammoths for which sufficient information was known, with the use of two alternative burial models. The mean ^{14}C age for each sample is also shown.

stant temperature)—and factored in time and temperature (at a conservative assumption of 10°C) since collection (*19*) (Fig. 2 and SOM text).

The data indicates that although the approximate thermal ages of several of the samples are older than the Poinar mammoth, their numbers of damaged derived miscoding lesions were lower (Fig. 2). The explanation for this remains unclear. It is possible that as hypothesized previously, hair cell keratinization protects the DNA within hair shafts from contact with free water, a requisite of the hydrolytic deamination underlying C→T damage (*16*). DNA may also be conserved because the hair, in contrast to porous bone, prevents access of bacteria to the site of DNA storage, thereby restricting the breakdown of biopolymers. Alternatively, the observation may be explained by other as-yet untested hypotheses. For example, special properties of hair shaft keratinocytes may confer advantages, such as an absence of postmortem cell autolysis; other molecules within the hair shaft (e.g., melanin) may provide protection; or the relatively unique preservation conditions that hair preservation in the archaeological record requires may in turn limit DNA degradation. Whatever the explanation, DNA degradation within the hair shafts does not appear to conform to current hypotheses about DNA degradation, and by inference the limits within which usable levels of DNA can be recovered from ancient samples may be greater than conventionally believed. This is in many ways unsurprising, given that many models of DNA degradation are based on theoretical degradation rates that were initially calculated to apply to DNA in free solution (*19*), and therefore it is plausible that their general applicability across biological tissues may not be straightforward.

Sequencing error—i.e., the difference between the (possibly damaged) molecule and the machine output—was also lower with the GS FLX. In all cases, the sum of damage plus sequencing error, as measured by the difference between the consensus sequence and the individual reads, was between 0.14 and 0.4%. Note that a C→T damage rate of 0.8% creates roughly a 0.2% component of the overall error rate, because only about one-quarter of nucleotides are C. Furthermore, although numts have been known to cause complications in mtDNA extracted from various mammalian tissues (including hair from some elephants) (*24*), a careful analysis (see SOM text for details) showed that contamination of our assemblies by numts was negligible.

Our findings have profound implications on the scope of future studies. Included in our data set are recently discovered mammoth permafrost specimens, including the Jarkov (M2), the Fishhook (M3), and the baby Dima (M8). Perhaps the most well-known sample among those we analyzed is M13, known colloquially as the Adams mammoth. This was the first mammoth to be scientifically studied, and the resulting documentation showed beyond reasonable doubt that an animal species can go extinct. The almost perfectly preserved permafrost mummy was found in 1799 by a hunter of the Tungus tribe, who collected its tusks in the summer of 1804 and eventually helped the Russian botanist Michael Adams to collect the remainder of the specimen in 1806. To this date, the Adams skeleton is one of the most complete, and it has been continuously on display at the Zoological Museum in St. Petersburg (*25*). In the process of recovering the entire skeleton, large amounts of hair, a total of 36 pounds (16.4 kg), were taken to St. Petersburg and distributed to other institutions around the world for investigation. The hair specimens have been stored for the past 200 years at room temperature, similar to most other samples that might be available for future analysis. Notably, even though these storage conditions are not optimal for DNA preservation (*19*), we were able to obtain a complete mitochondrial sequence from this specimen with the use of our whole-genome shotgun method, on no more than 0.2 g of hair shaft. The finding that aDNA can be extracted from a specimen kept at room temperature for two centuries puts a large number of collections stored in natural history museums within reach of molecular genomic analysis and may allow us to add molecular-genetic data to the collections of Charles Darwin, Alexander von Humboldt, and Carl von Linné.

References and Notes

1. D. M. Lambert *et al.*, *Science* **295**, 2270 (2002).
2. B. Shapiro *et al.*, *Science* **306**, 1561 (2004).
3. I. Barnes *et al.*, *Curr. Biol.* **17**, 1072 (2007).
4. R. Higuchi, B. Bowman, M. Freiberger, O. A. Ryder, A. C. Wilson, *Nature* **312**, 282 (1984).
5. A. Cooper *et al.*, *Nature* **409**, 704 (2001).
6. O. Haddrath, A. J. Baker, *Proc. R. Soc. London Ser. B* **268**, 939 (2001).
7. J. Krause *et al.*, *Nature* **439**, 724 (2006).
8. E. I. Rogaev *et al.*, *PLoS Biol.* **4**, e73 (2006).
9. J. P. Noonan *et al.*, *Science* **309**, 597 (2005).
10. M. Margulies *et al.*, *Nature* **437**, 376 (2005).
11. H. N. Poinar *et al.*, *Science* **311**, 392 (2006).
12. R. E. Green *et al.*, *Nature* **444**, 330 (2006).
13. J. P. Noonan *et al.*, *Science* **314**, 1113 (2006).
14. S. Pilkington, C. Summers, S. L. Thein, N. T. O'Connor, J. S. Wainscoat, *Lancet* **329**, 112 (1987).
15. R. Higuchi, C. H. von Beroldingen, G. F. Sensabaugh, H. A. Erlich, *Nature* **332**, 543 (1988).
16. M. T. P. Gilbert *et al.*, *Curr. Biol.* **14**, R463 (2004).
17. E. J. van Scott, T. M. Ekel, R. Auerbach, *J. Invest. Dermatol.* **41**, 269 (1963).
18. M. T. P. Gilbert *et al.*, *Forensic Sci. Int.* **156**, 208 (2006).
19. C. I. Smith, A. T. Chamberlain, M. S. Riley, C. Stringer, M. J. Collins, *J. Hum. Evol.* **45**, 203 (2003).
20. P. Pineau. *Pour La Science, Dossier Hors-serie* **43**, 35 (April–June, 2004).
21. M. T. P. Gilbert *et al.*, *Nucleic Acids Res.* **35**, 1 (2007).
22. C. A. Linch, D. A. Whiting, M. M. Holland, *J. Forensic Sci.* **46**, 844 (2001).
23. M. Stiller *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 13578 (2006).
24. A. D. Greenwood, S. Pääbo, *Mol. Ecol.* **8**, 133 (1999).
25. I. P. Tolmachoff, *Siberia Trans. Am. Philos. Soc.* **23**, 1 (1929).

26. This sequencing-by-synthesis study was made possible through generous funding from the Pennsylvania State University, Roche Applied Sciences, and a private sponsor (AvB). W.M. was supported by grant HG002238 from the National Human Genome Research Institute. M.T.P.G. acknowledges funding from the Marie Curie actions FP6 025002 Formaplex grant. L.D. acknowledges funding from the Marie Curie actions FP6 041545 Pleistocene Genetics grant. A.S. thanks Russian Foundation for Basic Research grants 04-04-48770 and 07-04-01612. T.K. thanks the Association Internationale pour la Promotion de la Cooperation avec les Scientifiques des Nouveaux Etats Independants de l'Ancienne Union Sovietique grants 03-51-668. This project is funded, in part, under a grant with the Pennsylvania Department of Health with the use of Tobacco Settlement Funds appropriated by the legislature. The Department specifically disclaims responsibility for any analyses, interpretations, or conclusions. The Genbank entries for the sequences in this manuscript are: specimen M1, EU153444; specimen M13, EU153445; specimen M18, EU153447; specimen M2, EU153449; specimen M22, EU153452; specimen M26, EU153454; specimen M3, EU153455; specimen M4, EU153456; specimen M5, EU153457; specimen M8, EU153458; and specimen Poinar, EU155210.

Supporting Online Material
www.sciencemag.org/cgi/content/full/317/5846/1927/DC1
SOM Text
Figs. S1 and S2
Tables S1 and S2
References

25 June 2007; accepted 13 August 2007
10.1126/science.1146971

Structures of the CCR5 N Terminus and of a Tyrosine-Sulfated Antibody with HIV-1 gp120 and CD4

Chih-chin Huang,[1]* Son N. Lam,[2]* Priyamvada Acharya,[1] Min Tang,[1] Shi-Hua Xiang,[3] Syed Shahzad-ul Hussan,[2] Robyn L. Stanfield,[4] James Robinson,[5] Joseph Sodroski,[3] Ian A. Wilson,[4] Richard Wyatt,[1] Carole A. Bewley,[2]† Peter D. Kwong[1]†

The CCR5 co-receptor binds to the HIV-1 gp120 envelope glycoprotein and facilitates HIV-1 entry into cells. Its N terminus is tyrosine-sulfated, as are many antibodies that react with the co-receptor binding site on gp120. We applied nuclear magnetic resonance and crystallographic techniques to analyze the structure of the CCR5 N terminus and that of the tyrosine-sulfated antibody 412d in complex with gp120 and CD4. The conformations of tyrosine-sulfated regions of CCR5 (α-helix) and 412d (extended-loop) are surprisingly different. Nonetheless, a critical sulfotyrosine on CCR5 and on 412d induces similar structural rearrangements in gp120. These results now provide a framework for understanding HIV-1 interactions with the CCR5 N terminus during viral entry and define a conserved site on gp120, whose recognition of sulfotyrosine engenders posttranslational mimicry by the immune system.

Entry of human immunodeficiency virus type 1 (HIV-1) into host cells requires its gp120 envelope glycoprotein to bind to two cell-surface receptors, CD4 and a co-receptor, either CCR5 or CXCR4 [reviewed in (*1*, *2*)]. CCR5 and CXCR4 are members of a family of chemokine receptors that are G protein–coupled receptors (*3*) characterized by seven transmembrane helices, an extracellular N terminus, which is variable in length, and three extracellular loops (ECLs) (Fig. 1A). The structure of the co-receptor has not been determined, but some insight has come from the crystal structures of other family members (*4*).

Elements critical to interactions with HIV-1 are located in the co-receptor N terminus and around its second extracellular loop (ECL2) (*5*–*8*). The co-receptor N terminus interacts with a highly conserved 4-stranded bridging sheet in gp120, which assembles upon CD4 binding, whereas the ECL2 region of the co-receptor interacts with the tip of the immunodominant V3 loop in gp120. Considerable distance separates these two interactive regions, which suggests that they are independent (*9*–*12*).

The N-terminal interaction of co-receptor with HIV-1 requires an unusual posttranslational modification, *O*-sulfation of tyrosine (*13*). On CCR5, tyrosines at residues 3, 10, 14, and 15 may be *O*-sulfated, but sulfations at residues 10 and 14 are sufficient to facilitate interaction with HIV-1 (*14*). Interestingly, many CD4-induced antibodies that react with the bridging sheet region are also modified by *O*-sulfation (*15*). To define structurally the interaction of HIV-1 with the N terminus of CCR5 and to understand the molecular details of the mimicry of this interaction by CD4-induced antibodies, we used a combination of nuclear magnetic resonance (NMR) and x-ray crystallography to determine the structures of the N terminus of CCR5 and of a functionally sulfated antibody, 412d, in complex with HIV-1 gp120. Analysis of these structures, combined with molecular docking and saturation transfer difference NMR, identified a conserved site on gp120, which recognizes sulfotyrosine with high selectivity.

We used NMR techniques that exploit the transfer of information from bound to ligand-free states (*16*, *17*) to analyze the interactions of a 14-residue peptide ($CCR5^{2\text{-}15}$), which consisted of residues 2 to 15 of CCR5 with sulfotyrosine (Tys) at positions 10 and 14 (Fig. 1) (*18*). We collected two-dimensional (2D) nuclear Overhauser enhancement spectroscopy (NOESY) spectra of solutions containing $CCR5^{2\text{-}15}$ either free or in the presence of gp120, CD4, or a gp120-CD4 complex (peptide:protein ratio of 40:1). Whereas spectra containing free $CCR5^{2\text{-}15}$ or $CCR5^{2\text{-}15}$ with either gp120 or CD4 contained few cross peaks, $CCR5^{2\text{-}15}$ in the presence of the gp120-CD4 complex gave rise to high-quality spectra containing numerous NOEs (Fig. 1B and fig. S1). Complete ^{1}H, ^{13}C, and ^{15}N assignments of $CCR5^{2\text{-}15}$ (table S1) were made on the basis of standard 2D homonuclear and heteronuclear NMR experiments that measure scalar and dipolar couplings.

The NOESY data of $CCR5^{2\text{-}15}$ in the presence of gp120-CD4 (Fig. 1B) were sufficient for calculating a high quality ensemble of NMR structures (Fig. 1C). Structure calculations were carried out on the ordered region comprising residues 7 to 15. A total of 70 distance restraints (corresponding to 35 intraresidue and 35 interresidue NOEs), and 56 dihedral angle restraints were included in the final round of structure calculations, which gave rise to an ensemble of 40 structures with a backbone root-mean-square deviation (rmsd) of 0.46 Å and an rmsd of 1.39 Å for all atoms in the ordered region (residues 9 to 14) (table S2). Superpositions of the final ensemble defined a helical conformation for residues 9 to 15, which deviated from the ideal by a backbone rmsd of only 0.26 Å (Fig. 1D). Sulfotyrosines 10 and 14 extended from the same face of the helix, with sulfate moieties separated by ~10 Å and an ~90° rotation around the helix axis.

We were unable to obtain crystals of $CCR5^{2\text{-}15}$ in complex with HIV-1 gp120-CD4, and the size and glycosylation of the ternary complex hindered direct determination by NMR. We were, however, able to obtain ~3.5 Å diffraction from crystals of the antigen-binding fragment (Fab) of the 412d antibody, in complex with gp120 (core with V3, CCR5-dependent isolate YU2) and CD4. The 412d antibody is functionally tyrosine-sulfated, binds to a CD4-induced epitope that overlaps the site of co-receptor binding on HIV-1 gp120, and recognizes preferentially CCR5-dependent strains of HIV-1 gp120 (*15*). Moreover, the tyrosine-sulfated region of 412d can be sub-

[1]Vaccine Research Center, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20892, USA. [2]Laboratory of Bioorganic Chemistry, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD 20892, USA. [3]Department of Cancer Immunology and AIDS, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02115, USA. [4]Department of Molecular Biology and Skaggs Institute for Chemical Biology, The Scripps Research Institute, La Jolla, CA 92037, USA. [5]Department of Pediatrics, Tulane University Medical Center, New Orleans, LA 70112, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: caroleb@mail.nih.gov (C.A.B.); pdkwong@nih.gov (P.D.K.)