The American Journal of Human Genetics, Volume *93*

Supplemental Data

Pulling out the 1%: Whole-Genome Capture for the Targeted Enrichment of Ancient DNA Sequencing Libraries

Meredith L. Carpenter, Jason D. Buenrostro, Cristina Valdiosera, Hannes Schroeder, Morten E. Allentoft, Martin Sikora, Morten Rasmussen, Simon Gravel, Sonia Guillén, Georgi Nekhrizov, Krasimir Leshtakov, Diana Dimitrova, Nikola Theodossiev, Davide Pettener, Donata Luiselli, Karla Sandoval, Andrés Moreno-Estrada, Yingrui Li, Jun Wang, M. Thomas P. Gilbert, Eske Willerslev, William J. Greenleaf, and Carlos D. Bustamante

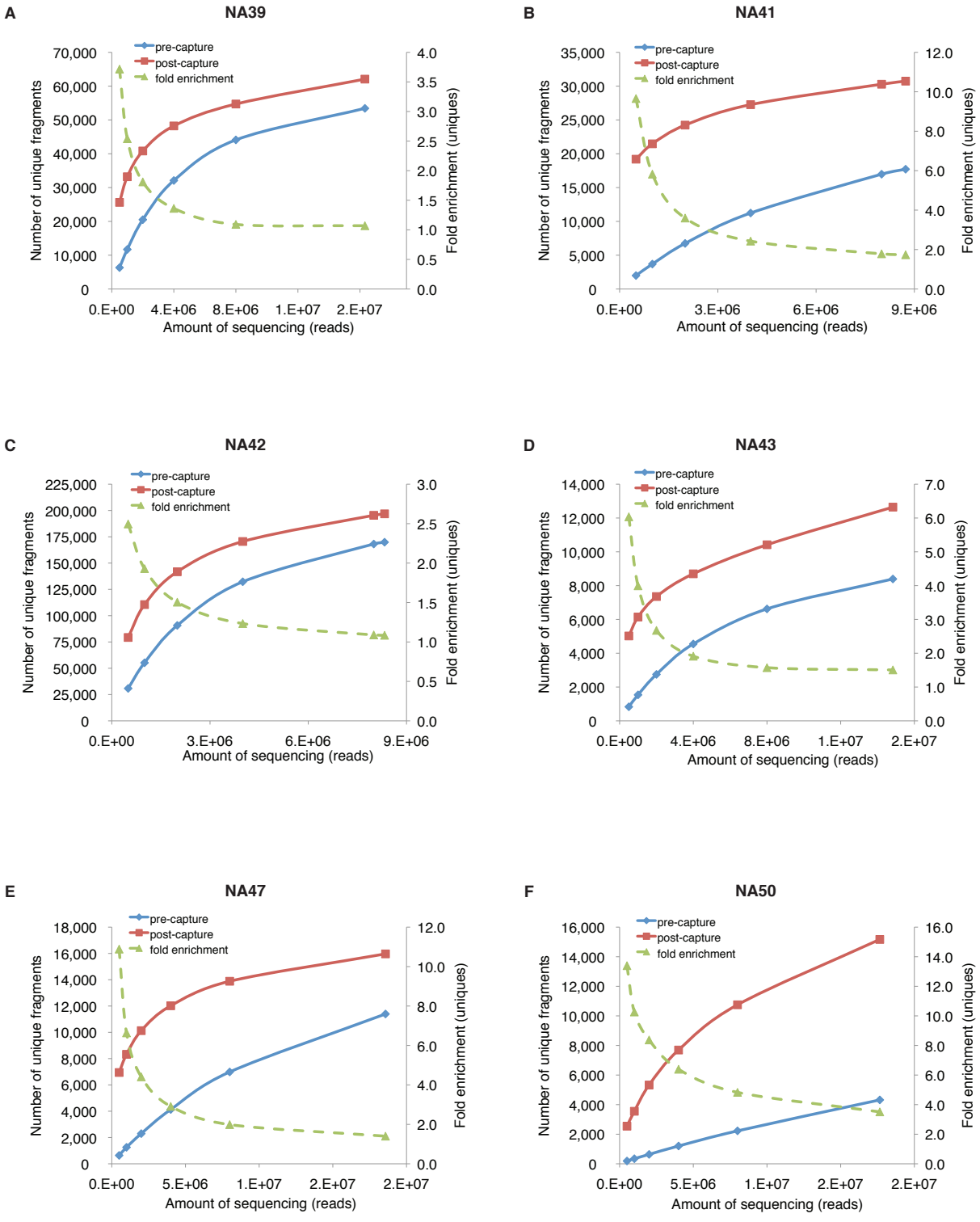
A
NA39
pre-capture
post-capture
fold enrichment
Number of unique fragments
Fold enrichment (uniques)
Amount of sequencing (reads)
B
NA41
C
NA42
D
NA43
E
NA47
F
NA50

Figure S1: Results of increased sequencing for samples (A) NA39, (B) NA41, (C) NA42, (D) NA43, (E) NA47, and (F) NA50. Shown is the yield of unique fragments pre-capture (blue) and post-capture (red) for each Peruvian bone sample with increasing amounts of sequencing. The fold enrichment in number of unique reads with increasing amounts of sequencing is plotted in green, with values on the secondary Y-axis.

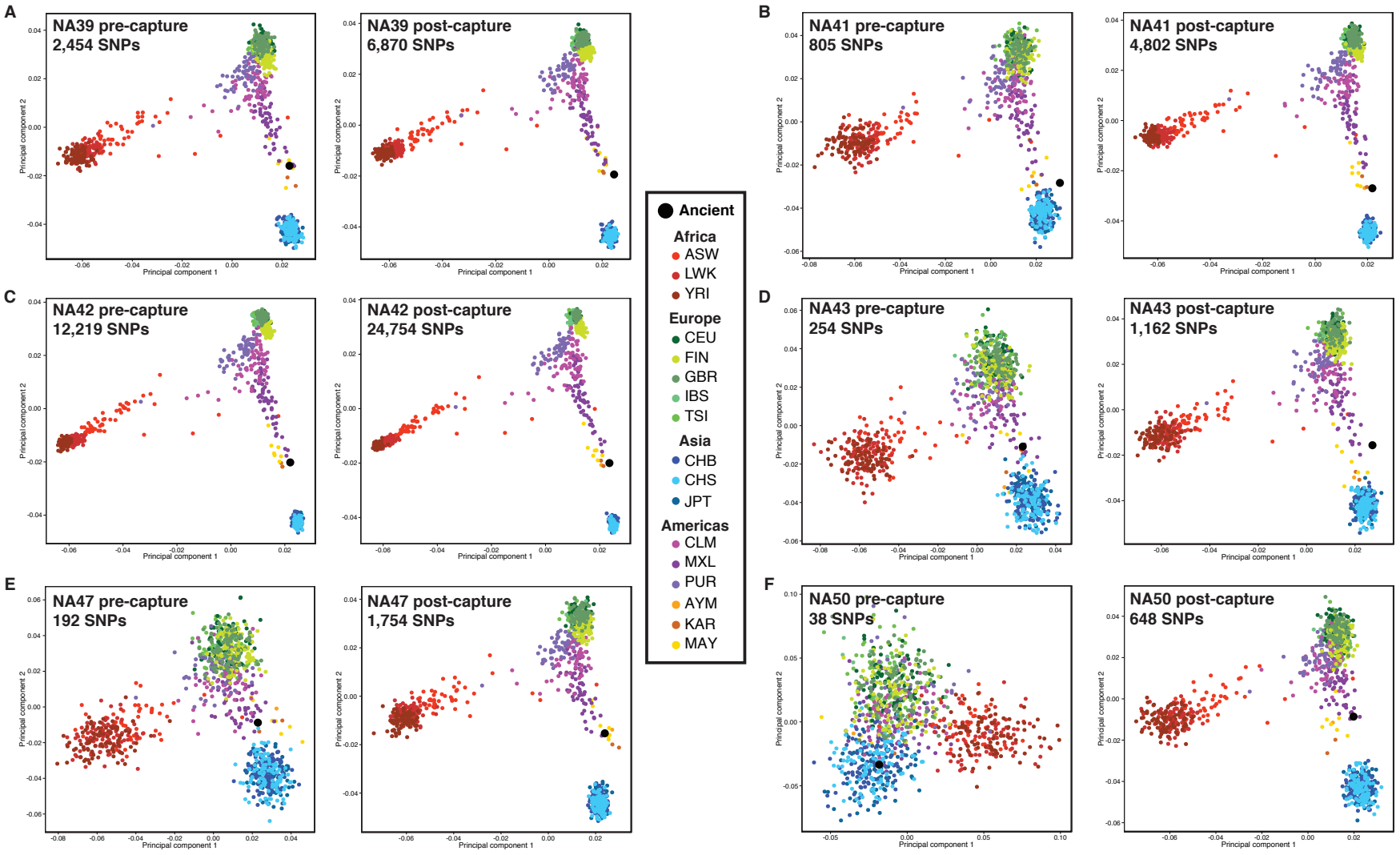
A
NA39 pre-capture
2,454 SNPs
NA39 post-capture
6,870 SNPs
B
NA41 pre-capture
805 SNPs
NA41 post-capture
4,802 SNPs
C
NA42 pre-capture
12,219 SNPs
NA42 post-capture
24,754 SNPs
D
NA43 pre-capture
254 SNPs
NA43 post-capture
1,162 SNPs
E
NA47 pre-capture
192 SNPs
NA47 post-capture
1,754 SNPs
F
NA50 pre-capture
38 SNPs
NA50 post-capture
648 SNPs
Principal component 1
Principal component 2
Ancient
Africa
ASW
LWK
YRI
Europe
CEU
FIN
GBR
IBS
TSI
Asia
CHB
CHS
JPT
Americas
CLM
MXL
PUR
AYM
KAR
MAY

Figure S2: Pre- and post-capture world PCAs for samples NA39, NA41, NA42, NA43, NA47, and NA50. Principal component analysis of SNPs overlapping between the 1000 Genomes reference panel, Native American individuals, and each ancient individual are shown. One million reads were sequenced for each pre- and post-capture library. The principal components were calculated using the modern individuals only, and the ancient individual was then projected onto the plot. Shown are a) NA39, b) NA41, c) NA42, d) NA43, e) NA47, and f) NA50. Population key: ASW, Americans of African ancestry in SW USA; AYM, Aymara from the Peruvian Andes; CEU, Utah residents (CEPH) with Northern and Western European ancestry; CHB, Han Chinese in Beijing, China; CHS, Southern Han Chinese; CLM, Colombians from Medellin, Columbia; FIN, Finnish in Finland; GBR, British in England and Scotland; IBS, Iberian population in Spain; JPT, Japanese in Tokyo, Japan; KAR, Karitiana from the Brazilian Amazon; LWK, Luhya in Webuye, Kenya; MAY, Mayan from Mexico; MXL, Mexican ancestry from Los Angeles, USA; PUR, Puerto Ricans from Puerto Rico; TSI, Toscani in Italy; YRI, Yoruba in Ibadan, Nigeria.

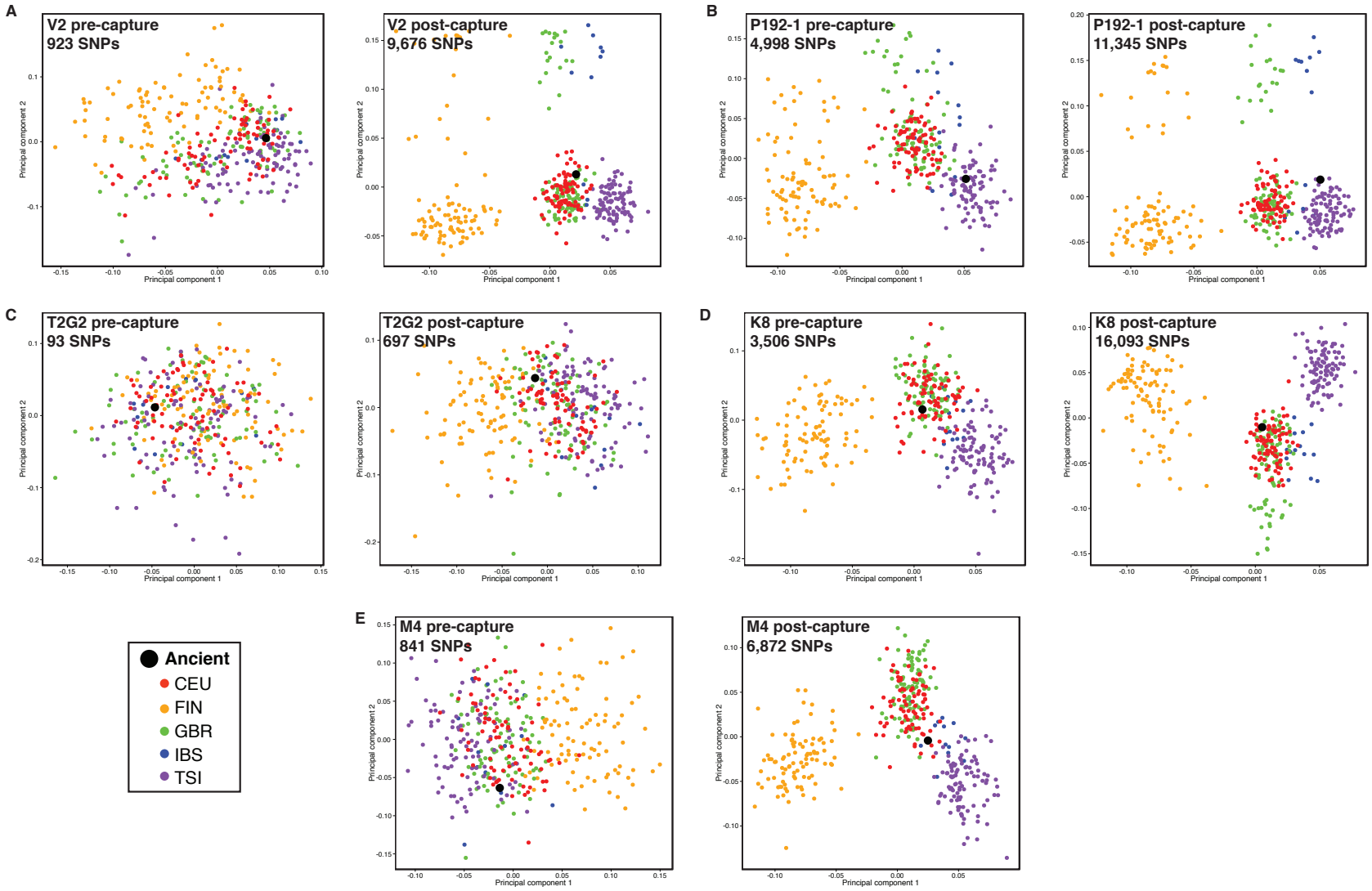
A
V2 pre-capture
923 SNPs
V2 post-capture
9,676 SNPs
B
P192-1 pre-capture
4,998 SNPs
P192-1 post-capture
11,345 SNPs
C
T2G2 pre-capture
93 SNPs
T2G2 post-capture
697 SNPs
D
K8 pre-capture
3,506 SNPs
K8 post-capture
16,093 SNPs
E
M4 pre-capture
841 SNPs
M4 post-capture
6,872 SNPs
Principal component 1
Principal component 2
Ancient
CEU
FIN
GBR
IBS
TSI

Figure S3: European-specific PCAs for Bulgarian samples and Danish hair sample. Principal component analysis plots using only the European populations from the 1000 Genomes reference panel and the ancient samples from Bulgaria (a-d) and the Danish hair sample (e). One million reads were sequenced for each pre- and post-capture library. The principal components were calculated using the modern individuals only, and the ancient individual was then projected onto the plot. Separation along PC2 by sequencing center has been observed previously for the 1000 Genomes European populations[1]. Population key: CEU, Utah residents (CEPH) with Northern and Western European ancestry; FIN, Finnish in Finland; GBR, British in England and Scotland; IBS, Iberian population in Spain; TSI, Toscani in Italy

Sample ID	Diagnostic mutations	Tentative mtDNA haplogroup assignment
V2	insufficient coverage	n/a
P192-1	**CR:** 750G 1811G 2706G 4640A 4769G 7028T 8860G 9656C 11719A 12308G 12372A 13743C 14139G 15326G 15454C	U3b
T2G2	**HVR2:** 263G **CR:** 750G 1438G 2706G 4769G 7028T 8860G 15326G **HVR1:** 16311C	HV(16311)
K8	insufficient coverage	n/a
M4	insufficient coverage	n/a
NA39	**HVR2:** 73G 263G 499A **CR:** 750G 827G 1438G 2706G 3547G 4769G 4820A 4977C 6473T 7028T 8860G 9950C 11177T 11719A 13590A 14766T 15326G 15535T **HVR1:** 16189C 16217C	B2
NA40	**HVR2:** 73G 263G 489C **CR:** 750G 1438G 2706G 4769G 7028T 8701G 8860G 9540C 10398G 10400T *10873T* 11719A *12705C* 14766T 14783C 15043A 15301A 15326G **HVR1:** *16223C*	M
NA41	insufficient coverage	n/a
NA42	**HVR2:** 73G 263G 489C **CR:** 750G 1438G 2092T 2706G 3010A 4769G 4883T 5178A 7028T 8414T 8701G 8860G 9540C 10398G 10400T 10873C 11719A 12705T 14668T 14766T 14783C 15043A 15301A 15326G **HVR1:** 16223T 16325C 16362C	D1
NA43	insufficient coverage	n/a
NA47	insufficient coverage	n/a
NA50	insufficient coverage	n/a

Table S1: Mitochondrial haplogroups for samples with >1X coverage of the mtDNA. Mutations are given relative to the revised Cambridge reference sequence (rCRS). Mismatches are indicated in italics; markers not shown were not covered by any reads. CR: Control region; HVR1: Hypervariable region 1; HVR2: Hypervariable region 2.

Sample ID	Pre- or post-capture	ChrX:All	Chr7:All	Tentative sex of individual
V2	PRE	0.051	0.058	Archaeological evidence: M DNA: possible F
	POST	0.045	0.055	
P192-1	PRE	0.025	0.055	Archaeological evidence: M DNA: M
	POST	0.026	0.059	
T2G2	PRE	0.029	0.057	Archaeological evidence: not available DNA: possible F
	POST	0.023	0.035	
K8	PRE	0.054	0.054	Archaeological evidence: 'prob M' DNA: F
	POST	0.049	0.056	
M4	PRE	0.038	0.055	Archaeological evidence: M DNA: not assigned (F contamination?)
	POST	0.034	0.054	
NA39	PRE	0.025	0.051	Archaeological evidence: not available DNA: M
	POST	0.024	0.051	
NA40	PRE	0.044	0.054	Archaeological evidence: not available DNA: F
	POST	0.042	0.053	
NA41	PRE	0.047	0.056	Archaeological evidence: not available DNA: possible F
	POST	0.042	0.055	
NA42	PRE	0.024	0.054	Archaeological evidence: not available DNA: M
	POST	0.024	0.055	
NA43	PRE	0.045	0.051	Archaeological evidence: not available DNA: possible F
	POST	0.040	0.050	
NA47	PRE	0.043	0.054	Archaeological evidence: not available DNA: possible F
	POST	0.043	0.052	
NA50	PRE	0.042	0.061	Archaeological evidence: not available DNA: possible F
	POST	0.046	0.054	

Table S2: X chromosome capture ratios in pre- and post-capture libraries. The proportion of reads mapping to the X chromosome out of all unique reads was determined before and after capture. As a control, the same calculation was performed for chromosome 7, which is approximately the same size as the X chromosome. The tentative sex of each individual based on available archaeological evidence and DNA evidence (determined using a recently reported karyotyping script for aDNA sequencing data[2]) is reported.

Sample ID	Number of SNPs that change between pre- and post-capture	Number of SNPs that match an allele in NA21732 after capture but not before capture
M4	5	3
NA39	1	1
NA40	6	3
NA41	1	0
NA42	3	2
NA43	0	0
NA47	1	1
NA50	0	0
TOTAL	17	10

Table S3: Testing for bias towards SNPs found in probe individual NA21732 after capture. For the eight libraries that were sequenced to higher coverage, SNPs that were called differently before and after capture were compared to the alleles found in the reference genome used to make the capture probes (NA21732).

	Frequency of C->T transitions at given base in read					
	Pre-capture base 1	Pre-capture base 25	Difference (base 1 - base 25)	Post-capture base 1	Post-capture base 25	Difference (base 1 - base 25)
All reads						
V2	0.002	0.000	**0.002**	0.002	0.002	**0.000**
P192-1	0.083	0.016	**0.067**	0.080	0.018	**0.062**
T2G2	0.083	0.019	**0.064**	0.198	0.034	**0.164**
K8	0.000	0.002	**-0.001**	0.001	0.001	**0.000**
M4	0.004	0.001	**0.003**	0.009	0.002	**0.007**
NA39	0.066	0.005	**0.061**	0.062	0.005	**0.058**
NA40	0.024	0.003	**0.021**	0.028	0.004	**0.024**
NA41	0.025	0.002	**0.023**	0.031	0.004	**0.027**
NA42	0.070	0.005	**0.065**	0.071	0.006	**0.065**
NA43	0.067	0.005	**0.062**	0.056	0.005	**0.051**
NA47	0.054	0.004	**0.051**	0.052	0.005	**0.047**
NA50	0.083	0.007	**0.076**	0.049	0.011	**0.038**
Reads <70 bp						
V2	0.000	0.000	**0.000**	0.000	0.000	**0.000**
P192-1	0.067	0.009	**0.058**	0.065	0.009	**0.056**
T2G2	0.051	0.008	**0.042**	0.240	0.033	**0.207**
K8	0.000	0.000	**0.000**	0.000	0.000	**0.000**
M4	0.004	0.000	**0.004**	0.002	0.000	**0.002**
NA39	0.070	0.002	**0.068**	0.068	0.000	**0.068**
NA40	0.023	0.002	**0.021**	0.025	0.002	**0.024**
NA41	0.081	0.000	**0.081**	0.107	0.000	**0.107**
NA42	0.079	0.002	**0.077**	0.080	0.002	**0.078**
NA43	0.065	0.001	**0.064**	0.068	0.002	**0.067**
NA47	0.056	0.003	**0.054**	0.058	0.003	**0.056**
NA50	0.129	0.003	**0.126**	0.120	0.000	**0.120**
Reads >70 bp						
V2	0.002	0.000	**0.002**	0.002	0.002	**0.000**
P192-1	0.089	0.019	**0.070**	0.065	0.020	**0.045**
T2G2	0.176	0.051	**0.125**	0.183	0.035	**0.149**
K8	0.001	0.002	**-0.001**	0.001	0.001	**0.000**
M4	0.004	0.001	**0.003**	0.010	0.002	**0.008**
NA39	0.065	0.005	**0.060**	0.062	0.005	**0.057**
NA40	0.025	0.004	**0.021**	0.028	0.004	**0.024**
NA41	0.024	0.002	**0.022**	0.031	0.004	**0.027**
NA42	0.069	0.006	**0.063**	0.070	0.006	**0.064**
NA43	0.067	0.008	**0.060**	0.053	0.005	**0.048**
NA47	0.053	0.004	**0.049**	0.050	0.006	**0.045**
NA50	0.053	0.010	**0.043**	0.041	0.012	**0.028**
Shading key:	**<0.02 (Low)**	**0.02-0.08 (Med)**	**>0.08 (High)**			

Table S4: DNA damage patterns in pre- and post-capture libraries. The frequency of C-to-T transitions at the given base in a read is listed for bases 1 and 25 in the pre- and post-capture libraries. The difference between the frequencies is also shown, with high numbers (shaded in dark red) indicating relatively higher levels of damage, and low numbers (shaded in blue) indicating lower levels of damage.

REFERENCES

1. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56-65.
2. Skoglund, P., Stora, J., Gotherstrom, A., and Jakobsson, M. (2013). Accurate sex identification of ancient human remains using DNA shotgun sequencing. Journal of Archaeological Science 40, 4477-4482.