Supplementary Methods

Data cleaning

Data cleaning for HGDP45 was performed as in Conrad et al. (2006) (Figure S1). Genotyping of 48 SNPs was attempted; three SNPs did not pass the quality checks of Conrad et al. (2006) (rs10868335 and rs4877914 failed the assay and rs1034044 was monomorphic), resulting in 45 SNPs in HGDP45. Eight of the 45 SNPs for which the HGDP45 dataset was genotyped failed on the SNPlex^{TM} platform. After removing the 8 SNPs that failed on the SNPlex^{TM} platform, 61 samples were removed from AMAS40 because they were missing more than 20% of genotypes across the remaining 40 SNPs for which AMAS40 was genotyped. Re-genotyping for D9S1120 showed three samples with multiple peaks or inconsistencies in genotype; two of these samples were also excluded because <80% of the SNP data was present. Through re-genotyping for D9S1120, we also identified nine samples as potential dropouts and genotyped them at least one additional time. All of these samples were genotyped as homozygous for a shorter allele and heterozygous for the same short allele one or more times. For these nine samples, we treated the second longer allele as missing data when phasing. Aside from the two HGDP controls, 235 AMAS40 samples with a mean of 6.3% missing SNP data were used for phasing. Of these 235 samples, 80.9% had been WGAed at Geneservice Ltd., 1.7% had been WGAed with the GenomiPhi^{TM} kit, and 17.5% had not been WGAed.

Merging the two datasets

Loci for which one or both of the two HGDP samples that we had genotyped on the SNPlex^{TM} platform at the UCLA Core Facility were homozygous allowed us to determine whether there was a change in allele polarity or state between genotyping platforms for 31 of the 34 SNPs

shared between the two datasets. We also compared the minor allele frequencies for each SNP in the HGDP45 pooled Native American (excluding the Surui) and AMAS40 pooled Native American datasets. We changed the polarity or state of the genotypes for 12 SNPs in the AMAS40 dataset (Figure S1) based on discrepancies across platforms in homozygous genotypes for the same individual, as detailed in Table S3. There were three SNPs for which both HGDP controls were either heterozygous or missing data: rs7025722, rs7031647, and rs7863248. For all three SNPs, the difference in minor and major allele frequencies was 0.20 or greater and the frequencies did not suggest a change in polarity. For the other 19 SNPs, one or more of the HGDP controls was homozygous and no change in allele polarity or state was suggested.

Phasing

We used the software PHASE, version 2.1 (Stephens and Scheet 2001; Stephens, Smith, and Donnelly 2005), to estimate haplotypes (Figure S1). We estimated phase with the data in several distinct groupings: Worldwide (only the Worldwide grouping includes all samples in HGDP45 and AMAS40), East Asia together with Western Beringia, East Asia not including Western Beringia, the Americas together with Western Beringia, the Americas not including Western Beringia, and East Asia together with Western Beringia and the Americas. For each grouping, we estimated phase, masked 10% of alleles, re-estimated phase, and then computed error rates for imputation of the masked alleles (Table S4). The lowest error rates occurred when phase was estimated for East Asia without Western Beringia. It is likely that the error rates are relatively high because all groupings include samples from both HGDP45 and AMAS40 and, therefore, there is a substantial amount of missing data in all groupings (i.e., each dataset was not genotyped for several SNPs for which the other dataset was genotyped -- genotypes at a

minimum of 11 SNPs were imputed for each sample in AMAS40, and genotypes at a minimum of 6 SNPs were imputed for each sample in HGDP45). Because the error rate for the Worldwide grouping was not substantially higher than for any of the other groupings, we chose to use the Worldwide haplotypes for downstream analyses; some of our analyses used all samples in HGDP45 and AMAS40, and we preferred to have these analyses use haplotypes estimated with just one phasing strategy. We phased the data with the Worldwide strategy five times, verified that haplotype frequencies were consistent across replicate runs, and then blindly picked the output from one of the replicates for use in all downstream analyses.

Final datasets

Following phasing, we observed that 90.5% of chromosomes with the 9-repeat allele share a 76.26 kb haplotype we call the "American Modal Haplotype" (AMH). However, we noticed three distinct non-AMH haplotypes on three chromosomes (Northern Paiute 294, Apache 239, and Mixtec 31) with the 9-repeat allele that did not appear to result from recombination within the AMH. All three samples had been genotyped as homozygous for the 9-repeat allele at D9S1120 at UC Davis, WGAed by Geneservice with the amplified product designated as "Usable" by Geneservice, and genotyped on the SNPlex^{TM} platform at UCLA. All three haplotypes differed from the AMH at rs3849873, 1107 bp to the right of the D9S1120 amplicon, and two of the haplotypes, Apache 329 and Mixtec 31, also differed from the AMH at rs4877301, 513 bp to the right of the D9S1120 amplicon. The association of the 9-repeat allele with a non-AMH haplotype could result from any of the following events: 1) recurrent mutation at D9S1120, 2) multiple recombination events within the AMH, 3) SNP mutation, 4) SNP genotyping error, 5) genotyping error or allelic dropout at D9S1120, 6) error introduced by

whole genome amplification, or 7) phasing error. Our primary interest lay in determining whether there has been recurrent mutation at D9S1120. For all three samples we amplified, cloned, and sequenced two overlapping fragments, for a total of ~1405 bp, which included D9S1120 and the two closest SNPs to the right of D9S1120. Primer sets used are listed in Table S5. Seven clones were successfully sequenced for Apache 329 and Mixtec 31, and eight clones were successfully sequenced for Northern Paiute 294.

For Northern Paiute 294, seven of the eight clones consisted of the 9-repeat allele associated with the AMH alleles at both of the SNPs. The eighth clone was an 8-repeat allele at D9S1120 associated with the AMH. We believe that this resulted from an error introduced via cloning because of the low frequency of the clone and because we had not previously observed an eight-repeat allele for Northern Paiute 294 or any other sample. Because the SNPlex genotyping results were not replicated in any of the clones, it is likely that one or more of the SNP genotypes previously ascertained with SNPlex is the result of genotyping error.

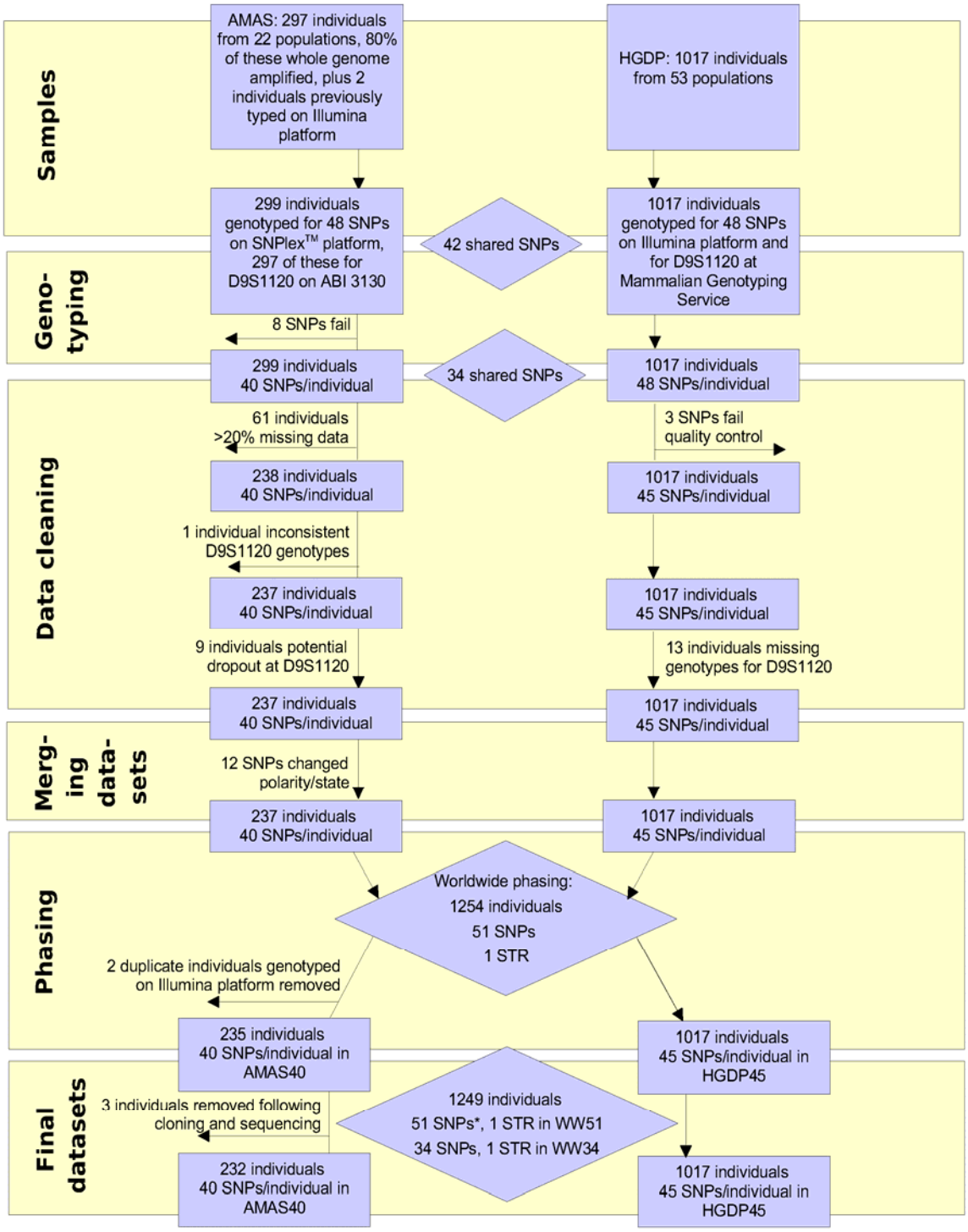
Four distinct cloned haplotypes were observed for Apache 329. The most common haplotype, observed in three of seven clones, was the 9-repeat allele with the AMH alleles at both SNPs, and the second most frequent haplotype, in two of seven clones, was a 16-repeat allele, also associated with the AMH. The two other low-frequency haplotypes were 1) the 9-repeat allele associated with the non-AMH allele at rs3849873 and 2) the 9-repeat allele associated with the non-AMH alleles at both rs3849873 and rs4877301. Four distinct cloned haplotypes were also observed for Mixtec 31; the most common haplotype, observed in three of seven of the clones, was a 17-repeat allele associated with the non-AMH alleles at both SNPs. Two of seven clones

were the 9-repeat allele associated with the AMH. The other two haplotypes, observed in one clone each, were the 9-repeat allele associated with the non-AMH alleles at both SNPs and a 17-repeat allele associated with the AMH alleles at both SNPs. For the Apache 329 and Mixtec 31 clones, the 9-repeat allele was associated with 3 and 2 different haplotypes, respectively. In each case, however, the most common haplotype was the AMH with the 9-repeat allele. Hence, it is likely that the discrepancy between the SNPlex haplotypes and cloned haplotypes results from error introduced by cloning, WGA, or allelic dropout at D9S1120, and there is no further cause to suspect that there has been recurrent mutation to the 9-repeat allele at D9S1120. Northern Paiute 294, Apache 329, and Mixtec 31 were removed from all downstream analyses (Figure S1); hence, the final AMAS40 dataset consisted of 232 samples.

Literature Cited

Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, Rosenberg NA, Pritchard JK. 2006. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat Genet. 38:1251-1260.
Stephens M, Scheet P. 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet. 76:449-462.
Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet. 68:978-989.

Figure S1. Flow chart describing the genotyping, data cleaning, and phasing of the two sample sets in this study.



*Minimum of 11 imputed SNP genotypes for AMAS40 samples in WW51 and minimum of 6 imputed SNP genotypes for HGDP45 samples in WW34

Table S1. Populations sampled in this study.

Population	Symbol	Sample size	Region	Dataset
Aleut	34	17	Americas	AMAS40
Apache	39	19	Americas	AMAS40
Cherokee	49	10	Americas	AMAS40
Chippewa	45	11	Americas	AMAS40
Creek	47	4	Americas	AMAS40
Dogrib	37	11	Americas	AMAS40
Fox	48	2	Americas	AMAS40
Huichol	42	9	Americas	AMAS40
Inuit	53	3	Americas	AMAS40
Jemez	41	6	Americas	AMAS40
Mixtec	44	27	Americas	AMAS40
NorthernPaiute	36	8	Americas	AMAS40
Seri	38	9	Americas	AMAS40
Sioux	43	5	Americas	AMAS40
Washo	35	9	Americas	AMAS40
Karitiana	51	24	Americas	HGDP45
Maya	46	25	Americas	HGDP45
Piapoco	50	13	Americas	HGDP45
Pima	40	21	Americas	HGDP45
Surui	52	21	Americas	HGDP45
Chukchi	33	22	East-Central Asia	AMAS40
Koryaks	32	19	East-Central Asia	AMAS40
AltaiKazakh	13	14	East-Central Asia	AMAS40
Mongola-OuterMongolia	20	20	East-Central Asia	AMAS40
NorthernAltai	12	2	East-Central Asia	AMAS40
SouthernAltai	11	5	East-Central Asia	AMAS40
Cambodian	19	9	East-Central Asia	HGDP45
Dai	16	10	East-Central Asia	HGDP45
Daur	27	10	East-Central Asia	HGDP45
Han	24	34	East-Central Asia	HGDP45
Han-NorthernChina	23	10	East-Central Asia	HGDP45
Hezhen	30	9	East-Central Asia	HGDP45
Japanese	31	29	East-Central Asia	HGDP45
Lahu	15	10	East-Central Asia	HGDP45
Miao	22	10	East-Central Asia	HGDP45
Mongola-InnerMongolia	25	10	East-Central Asia	HGDP45
Naxi	14	9	East-Central Asia	HGDP45
Oroqen	28	10	East-Central Asia	HGDP45
She	26	10	East-Central Asia	HGDP45
Tu	17	10	East-Central Asia	HGDP45
Tujia	21	10	East-Central Asia	HGDP45
Uygur	9	10	East-Central Asia	HGDP45
Xibo	10	9	East-Central Asia	HGDP45
Yakut	29	25	East-Central Asia	HGDP45
Yi	18	10	East-Central Asia	HGDP45
Balochi	2	24	South Asia	HGDP45

Brahui	3	24	South Asia	HGDP45
Burusho	8	23	South Asia	HGDP45
Hazara	5	24	South Asia	HGDP45
Kalash	7	25	South Asia	HGDP45
Makrani	1	25	South Asia	HGDP45
Pathan	6	24	South Asia	HGDP45
Sindhi	4	25	South Asia	HGDP45
Bedouin	NA	47	Middle East	HGDP45
Druze	NA	45	Middle East	HGDP45
Mozabite	NA	29	Middle East	HGDP45
Palestinian	NA	48	Middle East	HGDP45
Adygei	NA	15	Europe	HGDP45
Basque	NA	24	Europe	HGDP45
French	NA	29	Europe	HGDP45
Italian	NA	12	Europe	HGDP45
Orcadian	NA	14	Europe	HGDP45
Russian	NA	24	Europe	HGDP45
Sardinian	NA	27	Europe	HGDP45
Tuscan	NA	6	Europe	HGDP45
Melanesian	NA	19	Oceania	HGDP45
Papuan	NA	16	Oceania	HGDP45
BantuKenya	NA	12	Africa	HGDP45
BantuSouthernAfrica	NA	8	Africa	HGDP45
BiakaPygmy	NA	31	Africa	HGDP45
Mandenka	NA	23	Africa	HGDP45
MbutiPygmy	NA	14	Africa	HGDP45
San	NA	7	Africa	HGDP45
Yoruba	NA	25	Africa	HGDP45

Table S2. SNPs for which the samples in this study were genotyped.

Locus	Position (NCBI Build 35)	Dataset
rs17088334	85053298	HGDP45
rs17337860	85084053	HGDP45
rs17088374	85096450	WW34
rs6559853	85110900	HGDP45
rs2841445	85126731	WW34
rs17088420	85141513	HGDP45
rs2841486	85155600	WW34
rs4877918	85181132	WW34
rs7032592	85193290	WW34
rs7862916	85206478	WW34
rs2814734	85222343	WW34
rs2841443	85225848	AMAS40
rs2592991	85230060	AMAS40
rs13291799	85237265	WW34
rs2593017	85240224	AMAS40
rs2841453	85244429	AMAS40
rs12685505	85247160	AMAS40
rs12555508	85250750	AMAS40
rs10512167	85253684	WW34
rs7048862	85269825	WW34
rs3849872	85284382	WW34
rs11140976	85300443	WW34
rs11140984	85314284	WW34
rs17426617	85314387	WW34
rs6559867	85316239	WW34
D9S1120	85321417	-
rs4877301	85321930	WW34
rs3849873	85322587	WW34
rs1992812	85329420	HGDP45
rs1447026	85329946	WW34
rs7042753	85331227	WW34
rs4877940	85332664	HGDP45
rs1374500	85339567	WW34
rs4082114	85339786	HGDP45
rs10512168	85345810	WW34
rs1014690	85351123	HGDP45
rs1834264	85363552	WW34
rs2278111	85378655	WW34
rs1030303	85395220	HGDP45
rs10481762	85407691	WW34
rs1863120	85420237	HGDP45
rs1469202	85435373	WW34
rs11141033	85448369	HGDP45
rs2814724	85463213	WW34
rs7025722	85480379	WW34
rs7031647	85506670	WW34
rs10117745	85525431	WW34
rs7863248	85537681	WW34
rs4877949	85551964	WW34
rs4877950	85570008	WW34
rs3916184	85582785	WW34
rs12379804	85595034	WW34

Table S3. SNPs for which we changed the minor allele or allele state in AMAS40 due to different genotyping platforms (Illumina and SNPlex^{Tm}) used for HGDP45 and AMAS40. Two samples from HGDP45 were genotyped on the new platform (SNPlex^{Tm}) with AMAS40 to assist in determining changes in minor alleles or allele states.

Locus	Position	Change in minor allele?	Change in allele states?	HGDP45	AMAS40	Polarity apparent from homozygous HGDP controls?	Regional allele frequency consistent with controls?
rs7042753	85331227	-	Y	A/G	T/C	Y	Y
rs11140976	85300443	-	Y	A/G	T/C	Y	Ambiguous
rs4877949	85551964	-	Y	A/G	T/C	Y	Ambiguous
rs12379804	85595034	Y	-	C/G	G/C	Y	Y
rs10481762	85407691	-	Y	G/T	C/A	Y	Y
rs2841445	85126731	-	Y	G/T	C/A	Y	Y
rs2278111	85378655	-	Y	T/C	A/G	Y	Y
rs2841486	85155600	-	Y	T/C	A/G	Y	Y
rs10117745	85525431	-	Y	T/C	A/G	N	Y
rs3849872	85284382	-	Y	T/C	A/G	Y	Ambiguous
rs2814724	85463213	-	Y	T/G	A/C	Y	Ambiguous

Table S4. Phase error rates for different geographic groupings of the populations.

Grouping	**Error rate (10% of genotypes masked)**
Worldwide (no grouping)	0.1022
East Asia plus Americas	0.0923
East Asia	0.1161
Americas	0.0904
East Asia minus Western Beringia	0.0847
Americas plus Western Beringia	0.0984

Table S5. PCR and sequencing primers used to determine short haplotypes for three chromosomes with the 9-repeat allele and with haplotypes very divergent from the AMH.

Primer	Primer sequence (5' to 3')	Use
D9-1440R[2]	GGGTATGCTGAGGATTAAATGAG	PCR
D9-35F[2]	AGGATTTGAGACAAATGAAAGCA	PCR
D9S1120R[1]	TTAGCTGCTTCTGGGAAAGA	PCR and sequencing
D9S1120F[1]	TAGGATTTGAGACAAATGAAAGC	PCR and sequencing
M13-20-F	GTAAAACGACGGCCAGT	sequencing
M13-R	AACAGCTATGACCATG	sequencing
D9-630F	ATTTCCCAAATATAGTTGATCGT	sequencing
D9-1408R	ATAAAGCACTTGGTATGTCACAG	sequencing

[1] Short fragment [2] Long fragment

Table S6. Chromosomes with the 9-repeat allele and non-AMH haplotypes in WW34. The position of each SNP, in Mb, is for NCBI Build 35. From D9S1120 in either direction, cells were colored red until an allele that differed from the AMH. Recombinant haplotypes are labeled R1 through R7.

Sample	Haplotype	Position	85.254	85.270	85.284	85.300	85.314	85.314	85.316	85.321	85.322	85.323	85.330
		rs SNP ID/ microsatellite	10512167	7048862	3849872	11140976	11140984	17426617	6559867	D9S1120	4877301	3849873	1447026
-	AMH		G	T	C	A	A	T	A	9	G	C	A
Inuit_99	R1		G	T	C	A	A	T	A	9	C	C	A
Dogrib_09	R2		G	C	C	A	A	T	A	9	G	C	A
Washo_101	R3		G	T	C	A	A	T	A	9	G	C	G
N.Paiute_046	R4		A	T	C	A	A	T	A	9	G	C	G
Koryak_114	R5		A	T	C	A	A	T	A	9	G	C	A
Koryak_66	R5		A	T	C	A	A	T	A	9	G	C	A
Apache_492	R5		A	T	C	A	A	T	A	9	G	C	A
Apache_398	R5		A	T	C	A	A	T	A	9	G	C	A
Chippewa_15	R5		A	T	C	A	A	T	A	9	G	C	A
Chippewa_03	R5		A	T	C	A	A	T	A	9	G	C	A
Washo_101	R5		A	T	C	A	A	T	A	9	G	C	A
Washo_104	R5		A	T	C	A	A	T	A	9	G	C	A
N.Paiute_078	R5		A	T	C	A	A	T	A	9	G	C	A
Cherokee_55	R5		A	T	C	A	A	T	A	9	G	C	A
Cherokee_53	R5		A	T	C	A	A	T	A	9	G	C	A
Inuit_52	R6		G	C	C	G	G	A	A	9	G	C	A
Jemez_24	R7		A	T	C	A	A	A	G	9	G	C	A