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.