File S1

SUPPORTING METHODS AND RESULTS

MCMC algorithm We developed a Metropolis-Hastings MCMC algorithm (Gamerman and Hedibert 2006) to obtain samples from the joint posterior probability distribution for all model parameters. Haplotype frequencies were estimated using independence or random-walk chains. When independence chains were used, proposal values for haplotype frequencies (a vector p_{ij} containing values for each locus and population) were sampled from Dirichlet distributions that were independent of $\mathbf{p}$ from the previous time-step and similar in form to the expected posterior distribution for these parameters. This proposal distribution is very efficient when dealing with few haplotypes and intermediate haplotype frequencies. Random-walk chains were used when these criteria were not met, which involved sampling haplotype frequencies from Dirichlet distributions that were proportional to the vector p_{ij} from the previous MCMC step. At least one of these two proposal algorithms generally worked well with each data set, however, more complicated, alternative proposal distributions might be considered when a very large number of haplotypes are analyzed. The α and β parameters associated with the conditional prior on haplotype frequencies were estimated using random-walk chains. Specifically, new values for each α and β pair were proposed from bivariate Gaussian distributions centered on the previous parameter values with user adjusted variance and covariance. Specification of a high covariance between proposal values of α and β was imposed to increase chain mixing. The MCMC algorithm was written in C++ using the GNU Scientific Library (Galassi *et al.* 2009) and is available from the authors at `http://www.uwyo.edu/buerkle/software/` as the stand-alone software *bamova*.

Simulations: estimation of ϕ statistics We conducted a series of simulations to determine whether the proposed model provided reasonable estimates of genome-level ϕ-statistics. For these simulations we were solely concerned with genetic differentiation among popula-

tions (rather than also considering differentiation among groups of populations). For each of our three likelihood models we simulated sequence data using an infinite sites coalescent model (using R. Hudson's software *ms*; HUDSON 2002). One group of data consisted of sequences from 25 genetic regions, whereas the second group consisted of sequences from 500 genetic regions. All simulations assumed five populations split from a common ancestor τ generations in the past, where τ has units of $4N_e$. We varied τ from 0 to 1 in steps of 0.05 to produce 21 data sets each for 25 and 500 loci. The ancestral population and all five descendant populations were assigned population mutation rates $\theta = 4N_e\mu$ of 0.5, where μ is the per locus mutation rate. We assumed no migration following population subdivision. Forty gene copies were sampled from each of the five populations. For the *known haplotype model* analyses we treated the simulated sequences directly as the sampled data. For *NGS–individual model* and *NGS-population model* analyses we re-sampled the simulated sequence data sets such that coverage for each sequence was Poisson distributed ($\lambda = 2$). For the *NGS–individual model* analyses we retained information on which individual each sequence came from, whereas we only retained population identification for *NGS-population model* analyses. Each data set was analyzed using our *bamova* software, with MCMC details as described in the main document.

MCMC implementation of the proposed Bayesian models accurately quantified genetic structure among five simulated populations with sequence data from 25 or 500 genetic regions (Figure S1). In general, estimates of mean genome-level ϕ_{ST} (μ_{ST}) increased with the time since divergence of the five populations (τ). Credible intervals for genome-level parameters were relatively narrow, particularly when estimates were based on 500 genetic regions (Figure S1, S2). Moreover, credible intervals, and thus the uncertainty in genome-level parameters, were similar for all three first-level likelihood models (*known haplotype model*, *NGS-individual model*, and *NGS-population model*). We detected considerable variation in the extent of population structure among genetic regions (and hence non-zero σ_{ST} for genome-level ϕ_{ST}), except when the population divergence time was very low (Fig. S2). Pos-

terior probability estimates for μ_{ST} were similar to the empirical mean of the locus-specific ϕ statistics calculated directly from the raw data; however, the estimates of σ_{ST} were generally lower than the empirical standard deviation of ϕ_{ST} from the raw data.

In the analyses of simulated data sets, ϕ_{ST} increased reliably and as expected with time since population divergence. Moreover, estimates of genome-level ϕ_{ST} using the *known haplotypes model* were very similar to non-Bayesian point estimates of mean ϕ_{ST} (Figure S1). Additionally the estimates of genome-level ϕ_{ST} for the *known haplotypes model*, the *NGS–individual model*, and the *NGS–population model* were similar. This similarity in results among models suggest that high-coverage NGS data can provide parameter estimates with precision and accuracy equivalent to Sanger sequencing. Furthermore, the estimates of genome-level ϕ_{ST} for the SeattleSNPs human sequence data and chromosome-level ϕ_{ST} for the worldwide human SNP data (0.080–0.139) were similar to mean levels of genetic differentiation among human populations based on F_{ST} (e.g., F_{ST} = 0.09–0.14 for Yoruba, European, Han Chinese and Japanese populations; Weir *et al.* 2005; Barreiro *et al.* 2008). An important attribute of the model is that it also provides an accurate estimate of the uncertainty in the parameter estimates. This is an attribute not necessarily shared by non-Bayesian methods of parameter estimation, particularly when hierarchical or derived parameters are involved (Link and Baker 2009).

Human SeattleSNP data: alternative data subsets In addition to analysing the SeattleSNPs data set based on the first five SNPs in each gene we analysed four additional subsets of these data: 1) sequences based on the middle five SNPs in each gene, 2) sequences based on the last five SNPs in each gene, 3) sequences based on five SNPs spaced evenly across each gene, 4) and sequences based on every 12^{th} SNP in each gene (mean number of SNPs = 5.24, sd = 0.423). Analyses of these data sets were as described in the main text for the first five SNPs data set.

We classified four genes as high ϕ_{ST} outliers (using $a = 0.5$) in two or more of the data

subsets (Figs. 1, S3). Three of these genes, HSD11B2, FOXA2, and POLG2 were classified as ϕ_{ST} outliers based on the 'first five SNPs' data subset, and are described in the main document. Other outlier gene identified in more than one data subset was CPSF4, which encodes the cleavage polyadenylation specificity factor subunit 4 protein and is an essential component of pre-mRNA 3' processing in mammals (Barabino *et al.* 1997). Estimates of ϕ_{ST} for CPSF4 were as high as 0.382 (95% ETPI 0.262–0.496; 'last five SNPs' data subset, Fig. S3). Four additional genes were identified as high ϕ_{ST} outliers in single subsets of the data: FUT2, IL1F6, EPPB9, and IKBKB. When classified as outliers these genes had ϕ_{ST} estimates similar to the genes detected as outliers more than once (Figs. 1, S3). Interestingly, FUT2 was classified as a candidate gene experiencing balancing selection in European Americans based on levels of polymorphism and intermediate-frequency alleles by Andres *et al.* (Andrés *et al.* 2009) and is generally regarded as a well-established target of balancing selection (contrary to our findings). Variation among data subsets in whether genes were detected as outliers depended both on the distribution of divergent nucleotides along each gene and the extent of divergence at each of these nucleotides (Fig. 2). No genes were identified as low ϕ_{ST} outliers, nor were any genes identified as high ϕ_{ST} outliers using $a = 0.95$.

LITERATURE CITED

Andrés, A. M., M. J. Hubisz, A. Indap, D. G. Torgerson, J. D. Degenhardt, A. R. Boyko, R. N. Gutenkunst, T. J. White, E. D. Green, C. D. Bustamante, A. G. Clark, and R. Nielsen, 2009 Targets of balancing selection in the human genome. Molecular Biology and Evolution **26**: 2755–2764.

Barabino, S. M. L., W. Hubner, A. Jenny, L. MinvielleSebastia, and W. Keller, 1997 The 30-kD subunit of mammalian cleavage and polyadenylation specificity factor and its yeast homolog are RNA-binding zinc finger proteins. Genes and Development **11**: 1703–1716.

Barreiro, L. B., G. Laval, H. Quach, E. Patin, and L. Quintana-Murci, 2008 Natural selection has driven population differentiation in modern humans. Nature Genetics **40**: 340–345.

Galassi, M., J. Davies, J. Theiler, B. Gough, G. Jungman, M. Booth, and F. Rossi, 2009 *GNU Scientific Library: Reference Manual*. Network Theory Ltd.

Gamerman, D. and F. L. Hedibert, 2006 *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman and Hall, New York.

Hudson, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics **18**: 337–338.

Jakobsson, M., S. W. Scholz, P. Scheet, J. R. Gibbs, J. M. VanLiere, H.-C. Fung, Z. A. Szpiech, J. H. Degnan, K. Wang, R. Guerreiro, J. M. Bras, J. C. Schymick, D. G. Hernandez, B. J. Traynor, J. Simon-Sanchez, M. Matarin, A. Britton, J. van de Leemput, I. Rafferty, M. Bucan, H. M. Cann, J. A. Hardy, N. A. Rosenberg, and A. B. Singleton, 2008 Genotype, haplotype and copy-number variation in worldwide human populations. Nature **451**: 998–1003.

Link, W. A. and R. J. Baker, 2009 *Bayesian Inference: with Ecological Applications*. Academic Press, Maryland Heights, MO, USA.

Weir, B. S., L. R. Cardon, A. D. Anderson, D. Nielsen, and W. Hill, 2005 Measures of human population structure show heterogeneity among genomic regions. Genome Research **15**: 1468–1476.