DOI: 10.1534/genetics.110.124693

A Hierarchical Bayesian Model for Next-Generation Population Genomics

Zachariah Gompert[1] and C. Alex Buerkle

Department of Botany and Program in Ecology, University of Wyoming, Laramie, Wyoming 82071

Manuscript received October 27, 2010
Accepted for publication December 28, 2010

ABSTRACT

The demography of populations and natural selection shape genetic variation across the genome and understanding the genomic consequences of these evolutionary processes is a fundamental aim of population genetics. We have developed a hierarchical Bayesian model to quantify genome-wide population structure and identify candidate genetic regions affected by selection. This model improves on existing methods by accounting for stochastic sampling of sequences inherent in next-generation sequencing (with pooled or indexed individual samples) and by incorporating genetic distances among haplotypes in measures of genetic differentiation. Using simulations we demonstrate that this model has a low false-positive rate for classifying neutral genetic regions as selected genes (*i.e.*, ϕ_{ST} outliers), but can detect recent selective sweeps, particularly when genetic regions in multiple populations are affected by selection. Nonetheless, selection affecting just a single population was difficult to detect and resulted in a high false-negative rate under certain conditions. We applied the Bayesian model to two large sets of human population genetic data. We found evidence of widespread positive and balancing selection among worldwide human populations, including many genetic regions previously thought to be under selection. Additionally, we identified novel candidate genes for selection, several of which have been linked to human diseases. This model will facilitate the population genetic analysis of a wide range of organisms on the basis of next-generation sequence data.

THE distribution of genetic variants among populations is a fundamental attribute of evolutionary lineages. Population genetic diversity shapes contemporary functional diversity and future evolutionary dynamics and provides a record of past evolutionary and demographic processes. Methods to quantify genetic diversity among populations have a long history (Wright 1951; Holsinger and Weir 2009) and provide a basis to distinguish neutral and adaptive evolutionary histories, to reconstruct migration histories, and to identify genes underlying diseases and other significant traits (Bamshad and Wooding 2003; Tishkoff and Verrelli 2003; Voight *et al.* 2006; Barreiro *et al.* 2008; Lohmueller *et al.* 2008; Novembre *et al.* 2008; Tishkoff *et al.* 2009; Hohenlohe *et al.* 2010). For example, population genetic analyses in humans have resolved a history of natural selection and independent origins of lactase persistence in adults in Europe and East Africa (Tishkoff *et al.* 2007). Similarly, allelic diversity at the Duffy blood group locus is consistent with the action of natural selection, and one allele that has gone to fixation in sub-Saharan populations confers resistance to malaria (Hamblin and Di Rienzo 2000; Hamblin *et al.* 2002). These studies of natural selection, and many others involving a diversity of organisms, utilize contrasts between genetic differentiation at putatively selected loci and the remainder of the genome. Genomic diversity within and among populations is determined primarily by mutation and neutral demographic factors, such as effective population size and rates of migration among populations (Wright 1951; Slatkin 1987). Specifically, these demographic factors determine the rates of genetic drift and population differentiation across the genome. In contrast, selection affects variation in specific regions of the genome, including the direct targets of selection and to a lesser extent genetic regions in linkage disequilibrium with these targets (Maynard-Smith and Haigh 1974; Slatkin and Wiehe 1998; Gillespie 2000; Stajich and Hahn 2005). Thus, the genomic consequences of selection are superimposed on the genomic outcomes of neutral genetic differentiation and the two must be disentangled to identify regions of the genome affected by selection.

A variety of models and methods have been proposed to identify genetic regions that have been affected by selection (Nielsen 2005), and these population genetic methods can be divided into within-population and among-population analyses. The former include widely used neutrality tests based on the site-frequency spectrum for a single locus, such as Tajima's *D* test (Tajima 1989), as well as recently developed tests based on the presence of extended blocks of linkage disequilibrium

Supporting information is available online at http://www.genetics.org/cgi/content/full/genetics.110.124693/DC1.

[1]*Corresponding author:* Department of Botany, 3165, 1000 E. University Ave., University of Wyoming, Laramie, WY 82071.
E-mail: zgompert@uwyo.edu

or reduced haplotype diversity (Andolfatto *et al.* 1999; Sabeti *et al.* 2002; Voight *et al.* 2006). Among-population tests for selection require multilocus data sets and identify nonneutral or outlier loci by contrasting patterns of population divergence among genetic regions (Lewontin and Krakauer 1973; Beaumont and Nichols 1996; Akey *et al.* 2002; Beaumont and Balding 2004; Foll and Gaggiotti 2008; Guo *et al.* 2009; Chen *et al.* 2010). The most commonly employed of these methods is the F_{ST} outlier analysis developed by Beaumont and Nichols (1996). This test contrasts F_{ST} for individual loci with an expected null distribution of F_{ST} on the basis of a neutral, infinite-island, coalescent model (Beaumont and Nichols 1996). Loci with very high levels of among-population differentiation (*i.e.*, high F_{ST}) are considered candidates for positive or divergent selection, whereas loci with exceptionally low F_{ST} are regarded as candidates for balancing selection. However, many F_{ST} outlier analyses can be biased by violations of the assumed demographic history (Flint *et al.* 1999; Excoffier *et al.* 2009). Alternative Bayesian approaches to obtain a null distribution of population genetic differentiation assume that F_{ST}'s for individual loci represent independent draws from a common, underlying distribution that characterizes the genome and that can be estimated directly from multilocus data (Beaumont and Balding 2004; Foll and Gaggiotti 2008; Guo *et al.* 2009). These approaches are more robust to different demographic histories, but might have reduced power to detect selection relative to methods that model the appropriate demographic history when it is known (Guo *et al.* 2009).

Herein we propose a hierarchical Bayesian model for estimating genomic population differentiation and detecting selection. This method improves on current methods for detecting selection in several important ways. First, unlike previous Bayesian outlier detection models (*e.g.*, Beaumont and Balding 2004; Foll and Gaggiotti 2008; Guo *et al.* 2009), the likelihood component of the model captures the stochastic sampling processes inherent in next-generation sequence (NGS) data (Mardis 2008a,b). Specifically, NGS technologies result in uneven coverage among individuals, genetic regions, and homologous gene copies, including missing data for many individuals and loci, and thus, increased uncertainty in the diploid sequence of individuals relative to traditional Sanger sequencing (Lynch 2009; Gompert *et al.* 2010; Hohenlohe *et al.* 2010). This issue is most pronounced when population-level indexing is used (*e.g.*, Gompert *et al.* 2010), but should persist even with individual-level indexing and high coverage data (*i.e.*, even with high mean coverage the sequence coverage for certain combinations of individuals and genetic regions will be low). Moreover, appropriately modeling and accounting for this uncertainty is important and vastly preferable to simply ignoring it or discarding large amounts of sequence data (*e.g.*, Hohenlohe *et al.* 2010). Previous methods to detect selection on the basis of genetic differentiation among populations have focused solely on allele frequency differences, as captured by F_{ST} (Nielsen 2005). Instead the model we propose measures population genetic differentiation using Excoffier *et al.*'s ϕ-statistics, which are DNA sequence-based measures of the proportion of molecular variation partitioned among groups or populations (Excoffier *et al.* 1992). Quantifying genetic differentiation using ϕ-statistics allows us to include the genetic distances among sequences in the measure of differentiation and thus take mutation rate into account, which should increase our ability to accurately identify targets of selection (Kronholm *et al.* 2010). Finally, the model we propose provides a novel criterion for designating outlier loci. Although we do not believe this criterion is inherently superior to alternatives (*e.g.*, Beaumont and Nichols 1996; Foll and Gaggiotti 2008; Guo *et al.* 2009), we believe it accords well with the concept of statistical outliers and is well suited for genome scans for divergent selection. Specifically, the model assumes that the ϕ-statistics for each genetic region are drawn from a common, genome-level distribution and identifies outlier loci, or loci potentially affected by selection, on the basis of the probability of their locus-specific ϕ-statistic given the genome-level probability distribution of ϕ. This genome-level distribution is equivalent to the conditional probability for the locus-specific ϕ-statistics in the model.

We begin this article by fully describing the model, both verbally and mathematically. We then use coalescent simulations to generate data with a known history and investigate the performance of the model for identifying genetic regions affected by selection. To illustrate its capacity to identify exceptional genes that are likely to have experienced selection, we use the proposed model to analyze empirical data from two studies of human population genetic variation. The first of these data sets includes 316 completely sequenced genes from 24 individuals with African ancestry and 23 individuals with European ancestry (SeattleSNPs data set). The second data set was published by Jakobsson *et al.* (2008) and includes genotype data from 525,901 SNPs from 33 human populations distributed worldwide. We analyzed these human data with the *known haplotypes model* (see *Bayesian models for molecular variance*) and found evidence of selection affecting a total of 569 genes including genes previously identified as targets of selection in humans (*e.g.*, CYP3A5, SLC24A5, and PKDREJ) and novel genes not previously thought to be under selection (*e.g.*, FOXA2 and SPATA5L1). Whereas these remarkably large-scale studies do not involve NGS data, the analyses of empirical data identify a large set of genes for additional study and illustrate our analytical approach, which can be applied to a diversity of large-scale genomic data sets, including those that will result

from the anticipated widespread adoption of NGS for population genomics.

METHODS

Bayesian models for molecular variance: Our goal is to use molecular divergence among populations and groups to identify genetic regions that might have been affected by natural selection. Patterns of divergence at individual loci arise from differences in haplotype frequencies and the genetic distance among haplotypes. We assume the distances are fixed and known and we attempt to estimate the haplotype frequencies from sequence data using a hierarchical Bayesian model. The model includes a first-level likelihood for the probability of the observed haplotype counts given population haplotype frequencies, a conditional prior for the haplotype frequencies for each locus in each population given genome-level parameters and genetic distance matrix, and uninformative priors for the genome-level parameters. This conditional prior defines the distribution of ϕ-statistics across the genome, whereas locus-specific ϕ-statistics are derived from our estimate of haplotype frequencies and the genetic distance among haplotypes. From this model we are able to estimate the probability of each locus-specific ϕ-statistic given the estimated genome-level distribution, which serves as a metric for identifying outlier loci that depart from the typical level of differentiation and therefore might have been affected by selection.

First-level likelihood models: We developed three models for the first-level likelihood of the observed haplotype counts ($\mathbf{x}$) given the population haplotype frequencies ($\mathbf{p}$). Each model is applicable to a different category of DNA sequence data that was generated via a different process. These models are presented in order of increasing uncertainty in the true genotype of individuals and thus in the population allele frequencies. The first model (*known haplotype model*) is applicable if the two haplotypes of a diploid individual are known without error, as is generally assumed for phased Sanger sequence data. Under this model the probability of the observed haplotype counts for each locus and population follows a multinomial distribution, such that the complete likelihood is a product of multinomial distributions (for all loci and populations),

$$\mathrm{P}(\mathbf{x}\,|\,\mathbf{p}) = \prod_i \prod_j \frac{n_{ij}!}{x_{ij1}! \cdots x_{ijk}!} p_{ij1}^{x_{ij1}} \cdots p_{ijk}^{x_{ijk}}, \qquad (1)$$

where n_{ij} is the total number of observed sequences at locus i for population j, and x_{ijk} and p_{ijk} are the observed count and population frequency of the kth haplotype in the jth population at the ith locus. The multinomial likelihood model assumes Hardy–Weinberg equilibrium for each locus and linkage equilibrium among loci.

Because additional sampling error occurs when sequence reads are sampled stochastically from DNA templates, NGS sequence data require alternative first-level likelihood models with the specific model depending on the information associated with each sequence. NGS can incorporate individual-level indexing or barcoding (Craig *et al.* 2008; Hohenlohe *et al.* 2010; Meyer and Kircher 2010). With individual-level indexes a sequence can be assigned to a particular individual and locus, but there is uncertainty regarding whether diploid individuals with a single observed haplotype are heterozygous or homozygous. If we assume sequencing errors are negligible or have already been corrected, any individual with two distinct haplotypes sampled is known to be a heterozygote, whereas individuals with one sampled haplotype might be heterozygous for the observed haplotype and an alternative haplotype or homozygous for the observed haplotype. Under these circumstances we propose the following likelihood (*NGS-individual model*) for the observed haplotype counts given the population haplotype frequencies,

$$\mathrm{P}(\mathbf{x}\,|\,\mathbf{p}) = \prod_i \prod_j \prod_l \begin{cases} \left[2p_{ijk_a}p_{ijk_b}\right]\left[\frac{n_{ijl}!}{x_{ijk_al}!x_{ijk_bl}!}0.5^{x_{ijk_al}}0.5^{x_{ijk_bl}}\right] \\ \quad \text{if } h = 2 \\ p_{ijk_a}^2 + \sum_{k\neq k_a}\left[2p_{ijk_a}p_{ijk}\right]\left[0.5^{x_{ijk_al}}\right] \\ \quad \text{if } h = 1, \end{cases} \qquad (2)$$

where the product is across all individuals (l) as well as populations and loci, h is the number of distinct haplotypes observed for an individual at a locus, x_{ijk_al} and x_{ijk_bl} are the observed counts of the one- or two-haplotype sequences for individual l, and p_{ijk_a} and p_{ijk_b} are the frequencies of these haplotypes in population j. The complementary set of haplotypes that exist, but that were not observed for an individual, has counts and frequencies simply denoted by $x_{ijkl} = 0$ and p_{ijk}, respectively.

Alternatively, if NGS sequence data consist of indexed populations of pooled individuals (Gompert *et al.* 2010), rather than indexed individuals, information will be associated with sequences only at the population level. We propose a third likelihood model (*NGS-population model*) for this situation, in which the probability of the observed haplotype frequencies given the population frequencies is described by a multivariate Pólya distribution

$$\mathrm{P}(\mathbf{x}\,|\,\mathbf{p},\nu) = \prod_i \prod_j \frac{n_{ij}!}{\prod_k (x_{ijk}!)} \frac{\Gamma(\sum_k \nu_j p_{ijk} + 1)}{\Gamma(n_{ij} + \sum_k \nu_j p_{ijk} + 1)} \times \prod_k \frac{\Gamma(x_{ijk} + \nu_j p_{ijk} + 1)}{\Gamma(\nu_j p_{ijk} + 1)} \qquad (3)$$

(Gompert *et al.* 2010), where Γ is the gamma function, ν_j is the number of gene copies (*i.e.*, twice the number of

diploid individuals) sampled from population j, and other parameters are as described above. This likelihood model has been used previously by Gompert *et al.* (2010) for population-level NGS data. This likelihood function assumes that the frequency of each haplotype in the sample of individuals from a population can take on any value between zero and one and thus might not be appropriate when very low numbers of individuals are sampled from each population.

Model priors: The Bayesian model includes a conditional prior for the population haplotype frequencies **p** assuming the distance matrix **d** is fixed and known without error. The conditional prior with its estimated parameters provides a null distribution for identifying outliers and inferring selection (see *Designating outlier loci*). The conditional prior we choose depends on whether we are interested in genetic structure among populations or genetic structure among groups of populations (*e.g.*, geographic regions) and among populations within groups. For genetic structure among populations, we assign the following prior to **p**,

$$\mathrm{P}(\mathbf{p} \mid \alpha_{\mathrm{ST}}, \beta_{\mathrm{ST}}, \mathbf{d}) = \prod_i \frac{1}{B(\alpha_{\mathrm{ST}},\ \beta_{\mathrm{ST}})} \frac{(\phi_{\mathrm{ST}_i}+1)^{\alpha_{\mathrm{ST}}-1}(1-\phi_{\mathrm{ST}_i})^{\beta_{\mathrm{ST}}-1}}{2^{\alpha_{\mathrm{ST}}+\beta_{\mathrm{ST}}-1}}, \tag{4}$$

where B is the beta function and ϕ_{ST} denotes ϕ_{ST_i} for locus i calculated on the basis of p_i and d_i following Excoffier *et al.* (1992). This specification of the conditional prior for **p** does not correspond to a standard probability distribution with respect to **p**, but is equivalent to assuming that the locus-specific ϕ_{ST} are distributed Beta($\alpha = \alpha_{\mathrm{ST}}$, $\beta = \beta_{\mathrm{ST}}$, $a = -1$, $b = 1$) with **d** fixed, where α_{ST} and β_{ST} are the shape parameters of a Beta distribution and a and b define the lower and upper bounds of the distribution (*i.e.*, we use a rescaled Beta distribution). Thus, this conditional prior describes the distribution of ϕ_{ST} across the genome, with a mean and standard deviation given by

$$\mu_{\mathrm{ST}} = \frac{2\alpha_{\mathrm{ST}}}{\alpha_{\mathrm{ST}} + \beta_{\mathrm{ST}}} - 1 \tag{5}$$

and

$$\sigma_{\mathrm{ST}} = 2\sqrt{\frac{\alpha_{\mathrm{ST}}\beta_{\mathrm{ST}}}{(\alpha_{\mathrm{ST}} + \beta_{\mathrm{ST}})^2(\alpha_{\mathrm{ST}} + \beta_{\mathrm{ST}} + 1)}}, \tag{6}$$

respectively. We complete this model by assigning uninformative, uniform priors to $\alpha_{\mathrm{ST}} \sim U(0,\ u_\alpha)$ and $\beta_{\mathrm{ST}} \sim U(0,\ u_\beta)$. For all analyses presented in this article we set $u_\alpha = u_\beta = 10^6$, yielding priors with uniform density over all parameter values with nonnegligible support in the posterior; however, alternative values might be necessary for specific data sets. The above specification results in the following hierarchical Bayesian model:

$$\begin{aligned} &P(\mathbf{p}, \alpha_{\mathrm{ST}},\ \beta_{\mathrm{ST}} \mid \mathbf{x}, \mathbf{d}, \nu) \\ &\quad \propto P(\mathbf{x} \mid \mathbf{p}, \nu) P(\mathbf{p} \mid \alpha_{\mathrm{ST}}, \beta_{\mathrm{ST}}, \mathbf{d}) P(\alpha_{\mathrm{ST}}) P(\beta_{\mathrm{ST}}). \end{aligned} \tag{7}$$

We provide an alternative conditional prior for the haplotype frequencies when genetic structure among groups and populations is of interest. Specifically we assume

$$\begin{aligned} P(\mathbf{p} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{d}) = \prod_i &\frac{1}{B(\alpha_{\mathrm{ST}}, \beta_{\mathrm{ST}})} \frac{(\phi_{\mathrm{ST}_i}+1)^{\alpha_{\mathrm{ST}}-1}(1-\phi_{\mathrm{ST}_i})^{\beta_{\mathrm{ST}}-1}}{2^{\alpha_{\mathrm{ST}}+\beta_{\mathrm{ST}}-1}} \\ \times &\frac{1}{B(\alpha_{\mathrm{SC}}, \beta_{\mathrm{SC}})} \frac{(\phi_{\mathrm{SC}_i}+1)^{\alpha_{\mathrm{SC}}-1}(1-\phi_{\mathrm{SC}_i})^{\beta_{\mathrm{SC}}-1}}{2^{\alpha_{\mathrm{SC}}+\beta_{\mathrm{SC}}-1}} \\ \times &\frac{1}{B(\alpha_{\mathrm{CT}}, \beta_{\mathrm{CT}})} \frac{(\phi_{\mathrm{CT}_i}+1)^{\alpha_{\mathrm{CT}}-1}(1-\phi_{\mathrm{CT}_i})^{\beta_{\mathrm{CT}}-1}}{2^{\alpha_{\mathrm{CT}}+\beta_{\mathrm{CT}}-1}} \end{aligned} \tag{8}$$

where ϕ_{SC_i} and ϕ_{CT_i} denote ϕ_{SC} (molecular variation among populations within groups of populations) and ϕ_{CT} (molecular variation among groups of populations relative to total haplotypic diversity) for locus i calculated on the basis of p_i and d_i following Excoffier *et al.* (1992), and other parameters are as described above. This specification estimates genome-level distributions for each ϕ-statistic independently. Because locus-specific ϕ_{ST} are fully specified given locus-specific ϕ_{SC} and ϕ_{CT}, an alternative and equally valid modeling approach would be to treat the mean genome-level ϕ_{ST} as a derived parameter and specify a conditional prior based only on ϕ_{SC} and ϕ_{CT}. This alternative approach would account for the lack of independence among ϕ_{SC}, ϕ_{CT}, and ϕ_{ST} and would be closer to existing decompositions of ϕ-statistics. However, this alternative model prior would not provide posterior estimates of α_{ST} or β_{ST}, which are necessary to specify the posterior distribution for the genome-level distribution of ϕ_{ST}, as opposed to the posterior distribution of the mean genome-level ϕ_{ST}. The former is necessary to identity outlier loci with respect to ϕ_{ST}, which is central to our model. Similar to the model for genetic differentiation among populations only, Equation 8 does not correspond to a standard probability distribution with respect to **p**, but is equivalent to assuming that each of the locus-specific ϕ-statistics follows its own Beta distribution with **d** fixed. As with the model for population structure, it is possible to derive the mean and standard deviation of the genome-level beta distribution using Equations 5 and 6 and substituting the appropriate α- and β-parameters. Finally, as before we assign uninformative, uniform priors to all α- and β-parameters. This specification results in the following Bayesian model:

$$\begin{aligned} &P(\mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\beta} \mid \mathbf{x}, \mathbf{d}, \nu) \\ &\quad \propto P(\mathbf{x} \mid \mathbf{p}, \nu) P(\mathbf{p} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{d}) P(\alpha_{\mathrm{ST}}) P(\beta_{\mathrm{ST}}) P(\alpha_{\mathrm{SC}}) P(\beta_{\mathrm{SC}}) \\ &\qquad \times P(\alpha_{\mathrm{CT}}) P(\beta_{\mathrm{CT}}). \end{aligned} \tag{9}$$

Designating outlier loci: This Bayesian model gives rise to a framework for identifying outlier loci or genetic

regions with an unusual proportion of molecular variation partitioned among populations or groups. Such genetic regions might have experienced selection directly, or indirectly through linkage. Genetic regions with unusual patterns of molecular variation can be identified by contrasting ϕ-statistics for each locus with the genome-level distribution for each ϕ-statistic. Specifically, we identify outlier loci by estimating the posterior probability distribution for the quantile of each locus-specific ϕ-statistic in the genome-level distribution. Formally, we define a_i as the ath quantile of the posterior distribution for the locus-specific ϕ-statistic for locus i and let q_n be the interval with endpoints defined as the nth/2 and $1-$ nth/2 quantiles of the genome-level ϕ-statistic distribution. We then consider locus i an outlier at the ath quantile with respect to a given ϕ-statistic with probability n if the interval q_n does not contain a_i. Values of n and a used for outlier designation will determine the stringency of the analysis, with higher values of a and lower values of n resulting in fewer loci classified as outliers. In this article we set $n = 0.05$ ($q_n =$ [0.025,0.975]) and use two different values of a (0.5 and 0.95 quantiles). These values for a denote the median and 95th quantile of the posterior probability distribution for the quantile of each locus-specific ϕ-statistic. When only differentiation among populations is being considered, a genetic region can be designated an outlier only with respect to ϕ_{ST}, whereas a genetic region could be an outlier with respect to ϕ_{ST}, ϕ_{SC}, or ϕ_{CT} when group and population structure are considered. Outlier loci can be classified as having very low ϕ_{ST}, which could be indicative of balancing or purifying selection, or very high ϕ_{ST}, which could be indicative of positive selection within populations or divergent selection among populations (Beaumont and Balding 2004; Beaumont 2005). Patterns of selection giving rise low or high ϕ_{CT} and ϕ_{SC} might be a bit more complex. For example, high ϕ_{CT} outliers would be expected if divergent selection occurred among groups of populations with the same alleles favored within each group and low ϕ_{SC} outliers might be expected with balancing selection within groups that favored different subsets of alleles among groups.

Several approaches have been described for identifying outlier loci using genome scans for differentiation (Beaumont and Balding 2004; Foll and Gaggiotti 2008; Guo *et al.* 2009). Our approach is most similar to that proposed by Guo *et al.* (2009), but also differs from that approach in several important ways. Guo *et al.* (2009) contrast an approximation of the posterior probability distribution for each locus-specific ϕ_{ST} (θ_i's in their model) with the hyperdistribution describing among-locus variation in ϕ_{ST} (equivalent to our genome-level distribution). Approximation of the posterior probability distribution for ϕ_{ST} is achieved by defining a Beta distribution with first and second moments equal to those of the posterior probability distribution (Guo *et al.* 2009). This approximation allows Guo *et al.* (2009) to measure the divergence between the posterior probability distribution for each locus-specific ϕ_{ST} and the genome-level distribution (also a Beta distribution, which is derived from point estimates of genome-level parameters), using Kullback–Leibler divergence (Kullback and Leibler 1951). Following calibration to determine a cutoff value for significance, the Kullback–Leibler divergence measure is used to designate outlier loci. The primary distinction between the outlier detection method proposed by Guo *et al.* (2009) and the outlier detection method we have implemented is that their method tests for a difference between the posterior probability distribution of each locus-specific ϕ_{ST} (this distribution measures uncertainty in the parameter estimate) and a point estimate of the genome-level ϕ_{ST} distribution (this distribution measures expected variation among loci in ϕ_{ST}). In contrast, we test whether locus-specific ϕ-statistics are unlikely given the genome-level distribution (*i.e.*, not that our estimates of these locus-specific ϕ-statistics simply differ significantly from the genome-level distribution). Additionally, our method differs by accounting for uncertainty in the genome-level distribution by taking the marginal distribution of the quantile of each locus-specific ϕ-statistic in the genome-level distribution, rather than using a point estimate of the genome-level distribution, which is necessary for the method of Guo *et al.* (2009). We do not believe that either of these methods is necessarily superior, but rather that they provide alternative criteria and definitions of outlier loci, with our method perhaps more closely reflecting common conceptions of outlier loci among evolutionary biologists and their method utilizing more information from the posterior distribution of locus-specific ϕ_{ST}.

Analysis of simulated data sets: We conducted a series of simulations to determine what proportion of genetic regions would be classified as outliers in the absence of selection. These data sets were simulated for analysis with the population structure model and for each of the three likelihood models (*i.e., known haplotype model, NGS-individual model,* and *NGS-population model*). We simulated sequence data using an infinite-sites coalescent model, using R. Hudson's software *ms* (Hudson 2002). Data sets were simulated with 25 or 500 genetic regions. The simulations assumed five populations split from a common ancestor τ generations in the past, where τ has units of $4N_e$ and was set to 0.25, 0.5, or 1.0. We conducted simulations with migration among the five populations ($N_e m$) set to 0 or 2. 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. 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 resampled the simu-

lated 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 retained only population identification for *NGS-population model* analyses. We generated 10 replicate simulated data sets for each combination of likelihood model, number of genetic regions (25 or 500), τ, and migration rate. Each data set was analyzed using a Markov chain Monte Carlo (MCMC) implementation of the proposed model, using the *bamova* software we have developed (see supporting information, File S1, MCMC algorithm; available from the authors at http://www.uwyo.edu/buerkle/software/ as stand-alone software). We calculated the distance matrix for each locus, using the number of sites by which each pair of sequences differed. Estimation of the posterior probability distribution for all parameters for each data set was based on a single MCMC algorithm run that included a 25,000-iteration burn-in followed by 50,000 samples from the posterior. Sample history plots were monitored to ensure appropriate chain mixing and convergence on the stationary distribution. We classified genetic regions as outliers on the basis of the posterior probability distribution for the quantile of each locus-specific ϕ_{ST} in the genome-level ϕ_{ST} distribution, as previously described (see *Designating outlier loci*). We then calculated the mean proportion of loci classified as outliers across the 10 replicates for each combination of parameters.

We conducted an additional series of simulations to assess the capacity of our analytical model to detect selected loci among a larger set of neutral regions. We began with a set of simulations of population structure, as above. Simulations were conducted under all combinations of conditions described previously for simulations without selection. However, we simulated selective sweeps affecting 2 (8%, for 25-locus simulations) or 25 (5%, for 500-locus simulations) of the simulated loci. We allowed selective sweeps to occur in one, three, or five populations (one set of simulations for each). Selective sweeps were simulated by selecting one haplotype from each affected population at each affected locus and increasing its frequency to 1. This was meant to simulate recent and strong selective sweeps where affected loci were in arbitrary linkage disequilibrium with the gene subject to selection. Thus, it was possible for the same or different haplotypes to be driven to fixation across populations. We calculated the mean proportion of neutral and nonneutral loci classified as outliers across the 10 replicates for each combination of parameters.

Finally, we conducted a series of simulations to determine the extent to which our group structure model could correctly identify genetic regions affected by selection. For these simulations we concentrated on the *NGS-individual model* and a single demographic history. We simulated two groups of five populations, with the divergence time of the populations within each group equal to 0.25 and the divergence of the two groups equal to 0.75. We allowed no migration between groups, but set the within-group migration rate to 4. We simulated data sets of 25 and 500 loci, and as with the population structure simulations, we simulated selective sweeps affecting 2 (25-locus simulations) or 25 (500-locus simulations) genetic regions. Selective sweeps occurred in all five populations within one group ($s_g = 1$) or all five populations in both groups ($s_g = 2$). We simulated 10 replicate data sets for each combination of simulation parameters. MCMC settings were as previously described. We identified outlier loci on the basis of ϕ_{ST} (the partitioning of molecular variation among populations), ϕ_{CT} (the partitioning of molecular variation between the two groups), and ϕ_{SC} (the partitioning of molecular variation among the populations within each group). We determined the mean proportion of neutral and nonneutral loci classified as outliers on the basis of each ϕ-statistic across the 10 replicates for each combination of parameters.

Analysis of human SeattleSNP data: To illustrate the application of the proposed model to real genetic data we analyzed 316 fully sequenced genes (exons and introns) from the SeattleSNPs data set (http://pga.gs.washington.edu; downloaded May 2010). Each gene was sequenced in 24 individuals with African ancestry [either the African-American (AA) panel or the HapMap Yoruba (YRI) population] and 23 individuals with European ancestry [either the Centre d'Étude du Polymorphisme Humain (CEPH) population or the HapMap Uthah residents with European ancestry (CEU) population]. The phase of polymorphisms in these sequences has been estimated statistically using PHASE v2.0 (Stephens *et al.* 2001) to produce haplotypes. Similar to NGS data sets, these data are DNA sequences, but these data lack the sampling uncertainty associated with NGS data and might represent fewer (but much longer) genetic regions than would be typical for current NGS studies. We used these data and the *bamova* software to identify loci with exceptional ϕ_{ST} estimates that were consistent with divergent selection between or balancing selection within these European and African ancestry populations. The mean number of SNPs per gene was 62.98 (SD = 50.73), resulting in a large number of haplotypes per gene. This large number of SNPs per gene, which resulted from these data being complete gene sequences, was considerably greater than typical for the short reads generated by NGS (*e.g.*, Gompert *et al.* 2010; Hohenlohe *et al.* 2010) and resulted in poor MCMC mixing. Therefore we based our analysis on the first five SNPs in each gene (see File S1, Human SeattleSNP data: alternative data subsets, for analyses using other SNPs). All insertion–deletion polymorphisms were ignored. We used the *known haplotypes model* with population structure. We ran a 25,000-iteration burn-in followed by 50,000 iterations to esti-

mate posterior probabilities. We identified outlier loci using the criteria described for the simulated data sets. We then examined variation in genetic differentiation among SNPs within each outlier locus, as well as several loci that had typical, "neutral" ϕ_{ST} estimates, by calculating point estimates of F_{ST} at each SNP.

Analysis of worldwide SNP data: We obtained data for genomic diversity across 33 widely distributed human populations (including Africa, Eurasia, East Asia, Oceania, and America) from Jakobsson *et al.* (2008) (version 1.3; http://neurogenetics.nia.nih.gov/paperdata/public/). These data include statistically phased haplotypes for 597 individuals, based on 525,901 SNPs. These data differ from NGS data as they are not DNA sequences, but the large number of genetic regions in this data set might be similar to what will be acquired in NGS studies. For each SNP in the HumanHap550 genotyping panel we obtained annotations and information for location relative to genes from the manufacturer of the genotyping technology (Illumina, San Diego). To focus our analysis on haplotypic variation within transcribed, genic regions, we retained only those SNPs that were annotated as being within coding, intron, 5′-UTR, 3′-UTR, or UTR regions. If for a particular gene this included >4 SNPs (minimally, for inclusion we required 2 SNPs per locus), we retained SNPs that were annotated as within coding or intron sequence plus any neighboring SNPs to bring the total to 4 SNPs per locus. Finally, if there were >4 SNPs within coding or intron sequence at a locus, we retained the first and last SNP and randomly chose 2 intervening SNPs in coding or intron sequence. We utilized a maximum of 4 SNPs per locus because this was similar to the number of variable sites expected in short NGS data (Gompert *et al.* 2010; Hohenlohe *et al.* 2010) and led to a sufficiently small number of haplotypes across populations, relative to the numbers of individuals per population, to yield informative haplotype frequencies for populations. Some isolated SNPs had annotations to genes elsewhere in the genome and conflicted with neighboring SNPs and were excluded. However, we did allow these SNPs to break genes into separate genic regions for analysis, so that we utilized haplotypic data from 12,649 regions and 11,866 distinct genes. We excluded the limited amount of data from the mitochondrion, the Y chromosome, and the pseudoautosomal region of the X and Y chromosomes and focused on the autosomes and the X chromosome.

We conducted a separate analysis for each chromosome to test for evidence of selection on the basis of levels of molecular differentiation among these human populations, using the *bamova* software. This allowed us to contrast patterns of genetic variation among chromosomes and identify outlier loci relative to levels of genetic differentiation for the chromosome on which they are found. We used the *known haplotypes* likelihood model with population structure and estimated posterior probabilities on the basis of 50,000 MCMC iterations following a 25,000-iteration burn-in. We classified outlier loci for each data set using the criteria described for the simulated data sets.

RESULTS

Results from simulated data sets: When data were simulated in the absence of selection, few genetic regions were classified as outliers and thus as being associated with targets of selection (Table 1). This result suggests that the method has a low false-positive rate (mean proportion = 0.012, SD = 0.030). The proportion of genetic regions classified as outliers was similar across all likelihood models, but tended to be slightly lower when migration among populations was simulated (particularly for high ϕ_{ST} outliers). With $a = 0.5$ the mean proportion of genetic regions classified as outliers across 10 replicate simulations varied from 0.003 (*known haplotype model*, no migration, 500 sequence loci, high ϕ_{ST} outliers) to 0.034 (*NGS-population model*, no migration, 500 sequence loci, low ϕ_{ST} outliers). Using $a = 0.95$, the mean proportion of outliers detected across 10 replicate simulations varied from 0 (many simulation conditions) to 0.004 (*NGS-individual model*, no migration, 500 sequence loci, high ϕ_{ST} outliers).

Simulations that included selection typically resulted in an increased number of genetic regions being classified as outliers. As expected, we found that the ability to correctly classify nonneutral genetic regions as outlier loci was dependent upon the extent of selection. Few outliers were detected when only one population was affected (*e.g.*, high ϕ_{ST}, $a = 0.5$, *NGS-individual model* mean = 0.057, SD = 0.040), but most selected loci were correctly classified as outliers when all five populations were affected (*e.g.*, high ϕ_{ST}, $a = 0.5$, *NGS-individual model* mean = 0.612, SD = 0.310; Table S1, Table S2, and Table S3). Selection was easiest to detect when data were simulated and analyzed in accordance with the *known haplotype model* (in this case, when all five populations were affected by selection, all selected loci were identified as outliers even with $a = 0.95$; Table S1). Selection was generally easier to detect when migration was simulated. For example, under the *known haplotype model* the proportion of selected loci correctly classified as high outliers increased from 0.429 to 0.492 when simulated populations experienced the homogenizing effects of migration (means across all simulation conditions). For the complete set of simulations, the proportion of selected genetic regions that were correctly classified as outlier loci was greater than the proportion of neutral genetic regions incorrectly classified as outlier loci (Table S1, Table S2, and Table S3).

The ability of the method to correctly identify genetic regions affected by selective sweeps in the presence of group structure (*i.e.*, when populations are organized into groups of populations) was highly dependent on the

TABLE 1

Proportion of outlier loci in simulated neutral data

		$m = 0$				$m = 2$			
		$a = 0.5$		$a = 0.95$		$a = 0.5$		$a = 0.95$	
No. loci	τ	Low	High	Low	High	Low	High	Low	High
					Known haplotypes model				
25	0.25	0.012	0.028	0.000	0.000	0.016	0.020	0.000	0.000
	0.50	0.028	0.028	0.000	0.000	0.016	0.012	0.000	0.000
	1.00	0.024	0.004	0.004	0.000	0.012	0.016	0.000	0.000
500	0.25	0.0132	0.0212	0.0004	0.0036	0.0108	0.0170	0.0000	0.0012
	0.50	0.0240	0.0164	0.0028	0.0004	0.0130	0.0158	0.0000	0.0012
	1.00	0.0320	0.0032	0.0080	0.0000	0.0136	0.0188	0.0002	0.0024
					NGS-individual model				
25	0.25	0.012	0.012	0.000	0.000	0.020	0.016	0.000	0.000
	0.50	0.004	0.024	0.000	0.000	0.020	0.012	0.000	0.000
	1.00	0.020	0.024	0.000	0.000	0.040	0.008	0.000	0.000
500	0.25	0.0136	0.0242	0.0005	0.0042	0.0098	0.0156	0.0004	0.0020
	0.50	0.0220	0.0202	0.0022	0.0042	0.0112	0.0182	0.0004	0.0020
	1.00	0.0280	0.0074	0.0068	0.0006	0.0140	0.0176	0.0008	0.0036
					NGS-population model				
25	0.25	0.016	0.020	0.000	0.000	0.008	0.012	0.000	0.000
	0.50	0.024	0.040	0.000	0.000	0.008	0.012	0.000	0.000
	1.00	0.032	0.004	0.000	0.000	0.004	0.008	0.000	0.004
500	0.25	0.0112	0.0202	0.0002	0.0024	0.0054	0.0166	0.0000	0.0018
	0.50	0.0190	0.0164	0.0006	0.0012	0.0100	0.0174	0.0000	0.0012
	1.00	0.0340	0.0064	0.0050	0.0000	0.0088	0.0180	0.0000	0.0022

Mean proportion of loci is shown over 10 replicates identified as outliers in the absence of selection for each of the three different likelihoods (*known haplotypes model, NGS-individual model,* and *NGS-population model*). Three times since divergence (τ) and two levels of migration (*m)* were simulated. Outlier loci were identified as low or high outliers relative to the genome-wide distribution and based on two different quantiles ($a = 0.5$ or 0.95) of their posterior distribution.

extent of the selective sweeps. When selective sweeps were confined to a single group, outliers were most often detected as genetic regions with high differentiation among populations within a group (*i.e.*, high ϕ_{SC}; Table S4). The proportion of swept genetic regions correctly classified as ϕ_{SC} outliers was generally small and did not exceed 0.150, but was much greater than the proportion of neutral genetic regions classified as ϕ_{SC} outliers. In contrast, when both groups of populations were affected by selective sweeps, nearly all swept loci were classified as high ϕ_{ST} and ϕ_{SC} outliers (92.5–100% using $a = 0.5$) and a smaller proportion (0.100–0.225) were classified as low or high ϕ_{CT} outliers. Thus, the selective sweeps were most readily detected on the basis of high differentiation among all populations (ϕ_{ST}) or among populations within each group (ϕ_{SC}), but could also be detected on the basis of low or high differentiation between the two groups (ϕ_{CT}). This variety of outcomes arises from the stochastic nature used to determine the specific haplotypes that were swept to fixation. The false-positive rate for these group analyses was very low (between 0 and 0.032), similar to the population structure analyses (Table S2 and Table S4).

Divergent selection among African and European populations: Mean genome-level ϕ_{ST} between African ancestry and European ancestry populations based on the SeattleSNPs data set was 0.080 [95% equal tail probability interval (ETPI) = 0.065–0.097]. This means that ~8% of DNA sequence variation was partitioned between the African and European ancestry populations.

We classified three genes as high ϕ_{ST} outliers (using $a = 0.5$) on the basis of the first five SNPs data subset (Figure 1 and Figure 2). One of these genes was HSD11B2. Approximately 32% of molecular variation at this gene was partitioned between African and European ancestry populations ($\phi_{ST} = 0.317$, 95% ETPI = 0.159–0.482, Figure 1). Allelic variants of this gene produce an inherited form of hypertension and an end-stage renal disease (Quinkler and Stewart 2003). A weak association has also been detected between an intronic microsatellite in this gene and type 1 diabetes mellitus and diabetic nephropathy (Lavery *et al.* 2002). FOXA2 was also identified as an outlier, with $\phi_{ST} = 0.321$ (95% ETPI= 0.124–0.513, Figure 1). FOXA2 regulates insulin sensitivity and controls hepatic lipid metabolism in fasting and type 2 diabetes mice (Wolfrum *et al.* 2004;

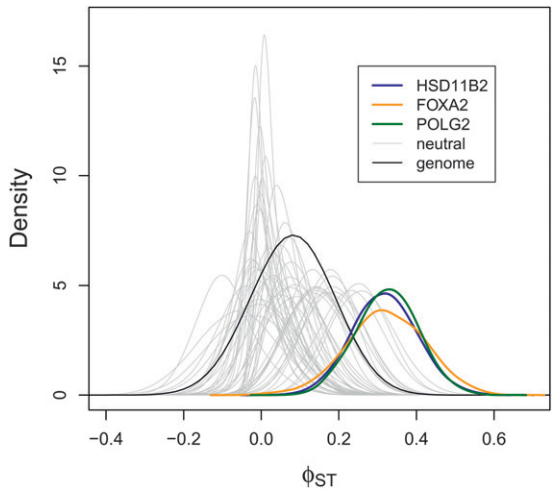

Figure 1.—Locus-specific ϕ_{ST} stimates for Africans and Europeans (SeattleSNPs data set). A point estimate of the genome-level ϕ_{ST} distribution (based on the median from the posterior probability distributions of α_{ST} and β_{ST}) is denoted with a solid black line. The posterior probability distributions for the three outlier loci (colored lines) and 50 additional, randomly chosen genetic regions (gray lines) are also shown. These results are based on the first five SNPs in each gene; additional results are shown in Figure S3.

Puigserver and Rodgers 2006). A genome-wide association study detected a SNP near FOXA2 (rs1209523) that was associated with fasting glucose levels in European- and African-Americans (Xing *et al.* 2010). This outlier is particularly interesting as type 2 diabetes is 1.2–2.3 times more common in African-Americans than in European-Americans (Harris 2001). The third outlier locus was POLG2, with an estimated ϕ_{ST} of 0.327 (95% ETPI = 0.175–0.477, Figure 1). This gene was also classified as a target of selection in humans in a recent study (Barreiro *et al.* 2008).

Selection among worldwide human populations: For the worldwide human SNP data, estimates of mean chromosome-level ϕ_{ST} were similar for the autosomal chromosomes and ranged from 0.083 (95% ETPI = 0.0752–0.091; chromosome 22) to 0.113 (95% ETPI = 0.104–0.120; chromosome 16; Table 2 and Figure 3). The estimate of ϕ_{ST} for the X chromosome was considerably higher (0.139, 95% ETPI = 0.128–0.149).

We detected remarkable variation in ϕ_{ST} along each of the chromosomes (Figure 4). We detected 569 unique ϕ_{ST} outlier loci, which we designate as candidate genes for selection among worldwide human populations (Table S5). Of these 569 genes, 518 were high ϕ_{ST} outlier loci (including 222 with $a = 0.95$) and 51 were low ϕ_{ST} outlier loci (including 6 with $a = 0.95$). Outlier loci were detected on all chromosomes, with the greatest number identified on chromosome 1 (67 outlier loci) and the fewest on chromosome 13 (9 outlier loci; Table 2 and Figure 4). Some of the genes that we classified as outliers have previously been implicated as genes experiencing selection in human populations. These include CYP3A4 and CYP3A5, which are cytochrome P450 genes found near one another on chromosome 7 ($a = 0.95$; CYP3A4 $\phi_{ST} = 0.293$, 95% ETPI = 0.245–0.319 and CYP3A5 $\phi_{ST} = 0.359$, 95% ETPI = 0.315–0.401). These genes are important for detoxification of plant secondary compounds and are involved in metabolism of some prescribed drugs. Additional evidence that these genes have experienced positive selection in human populations exists from previous studies based on distortions of the site frequency spectrum in African, European, and Chinese populations; population differentiation between the CEU and YRI populations; and extended haplotype homozygosity in HapMap populations (Carlson *et al.* 2005; Voight *et al.* 2006; Nielsen *et al.* 2007; Chen *et al.* 2010). Another example is the PKDREJ gene on chromosome 22, which is a candidate sperm receptor gene of mammalian egg-coat proteins and had one of the highest estimates of ϕ_{ST} (0.455, 95% ETPI = 0.396–0.504). Variation in this gene is consistent with positive selection among primate lineages, although evidence suggests that balancing selection might act to maintain diversity at this gene in human populations (Hamm *et al.* 2007). We also found evidence of divergent selection acting on SLC24A5, which has been associated with differences in skin pigmentation among humans and was classified as a candidate for positive selection in several studies (Barreiro *et al.* 2008). Finally, three of the high ϕ_{ST} outliers at $a = 0.95$ (RTTN, MSX2, and CDAN1) were among seven human skeletal genes identified by Wu and Zhang as genes with elevated F_{ST} at nonsynonymous SNPs between African and non-African (European and East Asian) populations and candidates for recent positive selection in Europeans and East Asians (Wu and Zhang 2010).

We detected additional genes with very high ϕ_{ST} that have not, to the best of our knowledge, been previously implicated as experiencing selection in human populations but have been found to be associated with disease traits. For example, the estimate of ϕ_{ST} for spermatogenesis-associated 5-like 1 (SPATA5L1) on chromosome 15 was 0.347 (95% ETPI = 0.246–0.430). This gene was classified as a high ϕ_{ST} outlier in the analysis and has been linked to renal function and kidney disease, but has not been identified in previous tests for selection (Köttgen *et al.* 2009).

We also identified novel candidate targets of balancing selection in worldwide human populations. Interestingly, none of the genes that we classified as low ϕ_{ST} outlier loci were implicated as targets of balancing selection in European- and African-American populations in a recent study by Andrés *et al.* (2009). Several of these genes have been linked to diseases. For example, RIF1 was classified as a low ϕ_{ST} outlier at the $a = 0.95$ level ($\phi_{ST} = -0.065$, 95% ETPI = −0.105– −0.023). RIF1 is an anti-apoptotic factor involved in DNA repair that is necessary for S-phase progression and is heavily expressed in breast cancer tumors (Wang *et al.* 2009).

TABLE 2

Summary of worldwide human HapMap data and results

Chromosome	Chromosome ϕ_{ST}		No. loci	High ϕ_{ST}		Low ϕ_{ST}	
				$a = 0.5$	$a = 0.95$	$a = 0.5$	$a = 0.95$
1	0.098	(0.095–0.103)	1402	59	28	8	1
2	0.101	(0.097–0.106)	956	32	11	5	1
3	0.099	(0.095–0.104)	723	29	13	2	0
4	0.100	(0.094–0.106)	563	22	12	1	0
5	0.095	(0.089–0.100)	565	20	10	3	2
6	0.091	(0.087–0.095)	710	23	8	6	0
7	0.098	(0.092–0.103)	679	25	14	2	0
8	0.092	(0.087–0.098)	438	24	6	2	0
9	0.095	(0.089–0.099)	569	22	9	2	0
10	0.092	(0.087–0.098)	526	20	6	3	2
11	0.096	(0.092–0.101)	686	24	7	1	0
12	0.097	(0.092–0.102)	683	36	14	2	0
13	0.090	(0.082–0.099)	243	8	7	1	0
14	0.100	(0.093–0.107)	388	16	8	2	0
15	0.109	(0.101–0.117)	380	20	10	0	0
16	0.113	(0.104–0.120)	426	21	9	1	0
17	0.107	(0.101–0.114)	603	27	12	1	0
18	0.092	(0.084–0.100)	225	8	4	3	0
19	0.099	(0.094–0.104)	729	34	10	4	0
20	0.101	(0.093–0.109)	367	15	8	0	0
21	0.083	(0.075–0.091)	158	6	3	1	0
22	0.102	(0.093–0.111)	283	11	6	1	0
X	0.139	(0.128–0.149)	347	16	7	0	0

Summary of data and results for each chromosome are shown, including chromosome-level ϕ_{ST} (median and 95% ETPI), the number of loci, and the number of loci classified as outliers (data from Jakobsson *et al.* 2008).

Another example is the AKT3 gene found on chromosome 1 ($\phi_{ST} = -0.046$, 95% ETPI $= -0.102$–0.020; outlier at $a = 0.5$). This gene is involved in cell-cycle regulation and is highly expressed in malignant melanoma, but is also important for attainment of normal organ size, including brain size in mice (Stahl *et al.* 2004; Easton *et al.* 2005).

DISCUSSION

We have presented a novel model to quantify genome-wide population genetic structure and identify genetic regions that are likely to have experienced natural selection. Unlike previous methods for quantifying structure and detecting genetic signatures of selection, the proposed method accurately models the stochastic sampling of sequences that is inherent in current NGS instruments and incorporates genetic distances among sequences in estimates of genetic differentiation. Although few population genomics studies based on NGS of individuals have been published to date (hence the analysis of Sanger sequence and SNP human data sets instead of NGS data sets), various large-scale projects are currently underway to obtain these data in large samples of humans (*e.g.*, 1000 genomes project, http://www.1000genomes.org) and a few recent studies suggest that these data will soon be available for many nonhuman (including nonmodel) species (Gompert *et al.* 2010; Hohenlohe *et al.* 2010). However, beyond data acquisition, substantial biological insights will be possible only if accompanied by models and methods designed to take full advantage of these data, while accurately modeling sources of error. Along with a few other recent reports of models for NGS (Guo *et al.* 2009; Lynch 2009; Futschik and Schlötterer 2010; Gompert *et al.* 2010), we believe our analytical methods help to address this need.

Model properties: The analyses of simulated and human genetic data sets suggest that the model provides statistically sound estimates of population differentiation for large sets of loci (see File S1, Figure S1, and Figure S2, Simulations: estimation of ϕ-statistics). For example, 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)]).

The simulation results indicate that the model only rarely identifies neutral loci incorrectly as outliers (*i.e.*, it has a low false-positive rate). Using $a = 0.5$, the false-

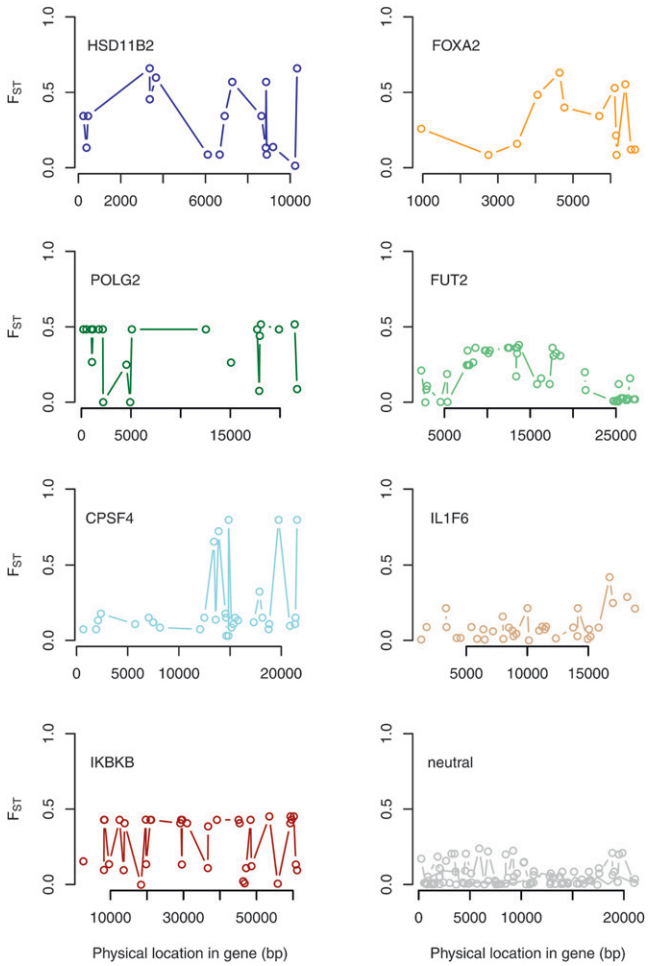

FIGURE 2.—Point estimates of F_{ST} along specific genes (SeattleSNPs data set). Point estimates of F_{ST} are shown at each SNP for seven genes identified as outliers in the comparison of human populations with African and European ancestry as well as three randomly chosen neutral genes.

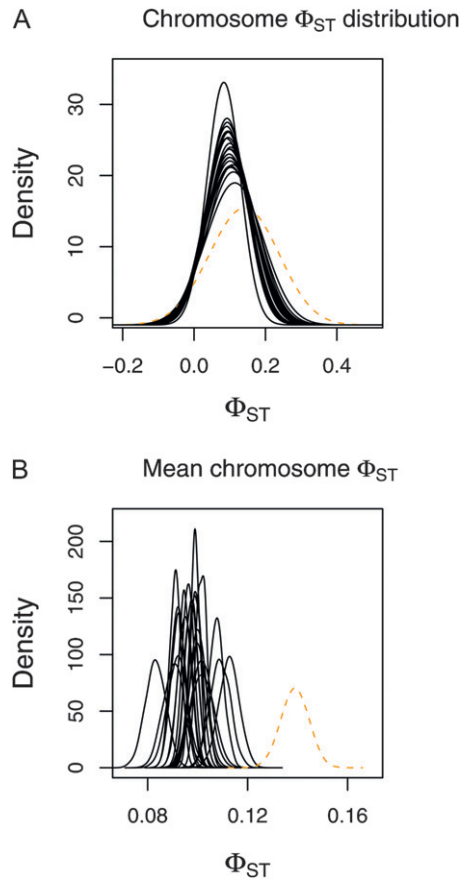

FIGURE 3.—Chromosome-level estimates of ϕ_{ST} for the large sample of human genetic diversity in 33 populations (data from JAKOBSSON *et al.* 2008). (A and B) Point estimates of each chromosome-level ϕ_{ST} distribution (based on the median from the posterior probability distributions of α_{ST} and β_{ST}) are denoted with solid black lines (autosomes) or a dashed orange line (A; X chromosome). (B) Posterior probability distributions for the mean chromosome-level ϕ_{ST} for each chromosome.

positive rate for high or low ϕ_{ST} outliers was never >0.04, and using $a = 0.95$ this rate was never >0.01. Low false-positive rates are particularly important for genome-wide scans in which a large number of genes could otherwise show spurious evidence of selection. In contrast, false-positive rates as high as 0.343 have been reported for F_{ST} outlier analyses of selection that use coalescent simulations under an incorrect demographic history to derive a neutral distribution (EXCOFFIER *et al.* 2009). Moreover, the low false-positive rate for the method held across all simulated demographic scenarios (*i.e.*, different population divergence times, variation in migration rates, and the presence or absence of group structure). This is expected as the null, genome-level distribution we estimate is based on the observed data rather than an assumed demographic history and most demographic histories should be appropriately captured by this genome-level distribution (BEAUMONT and BALDING 2004; GUO *et al.* 2009). Similarly low false-positive rates were reported by GUO *et al.* (2009), using a hierarchical Bayesian model based on F_{ST}. Nonetheless, it is important to note that these hierarchical models generally identify outlier loci as those that are inconsistent with the genome as a whole and thus will not work well if most of the genome has recently experienced selection of a similar magnitude.

Simulation results further indicate that the method has the ability to detect genetic regions under selection and to a greater extent when selective sweeps affect multiple populations or migration occurs. When a simulated genetic region was affected by selection in at least three populations, the selected locus was generally at least 10 times more likely to be classified as an outlier than a randomly chosen neutral locus. This suggests a high true-positive rate (at least under favorable conditions) and that candidate selected genes identified by the method will be enriched substantially for genes actually experiencing selection. This is particularly true when genes are classified as outliers using the more stringent criterion of $a = 0.95$ (although this will also

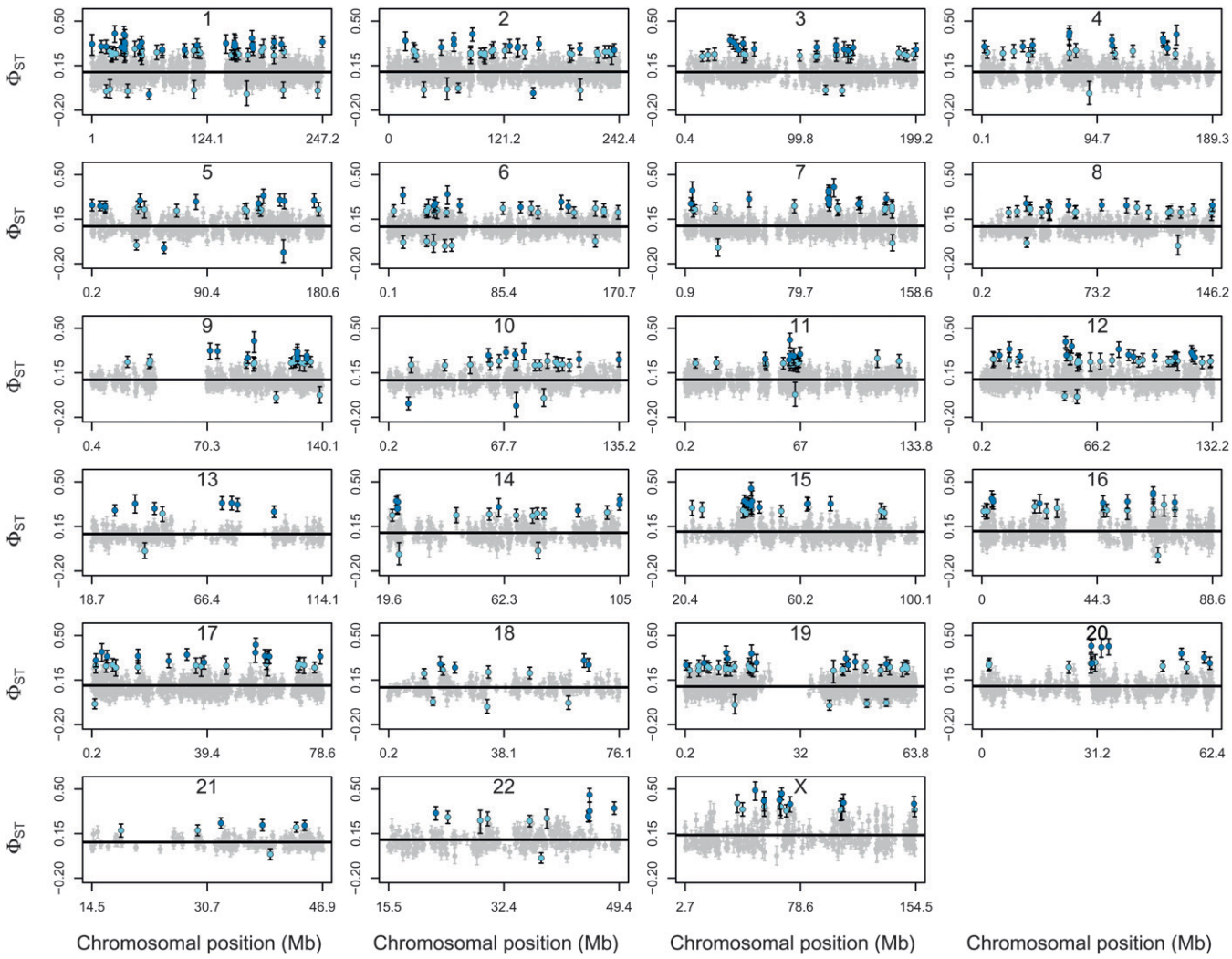

FIGURE 4.—Estimates of ϕ_{ST} across the genome of worldwide human populations (data from JAKOBSSON *et al.* 2008). Each panel depicts a different human chromosome and is labeled accordingly. The solid black line denotes the point estimate (median of the posterior distribution) of the mean chromosome-level ϕ_{ST} for a chromosome. Estimates of ϕ_{ST} for individual genes are shown in gray for nonoutlier genes and light (at $a = 0.5$) or dark blue (at $a = 0.95$) for outlier genes. For each gene the solid circle gives the median from the posterior distribution of ϕ_{ST} for that locus and the bars denote the 95% ETPI.

decrease the total number of selected genes identified relative to $a = 0.5$). One benefit of quantifying genetic divergence on the basis of haplotypes (using ϕ_{ST}) instead of SNP, microsatellite, or AFLP alleles (using F_{ST}) is that multiple selective sweeps should result in increased genetic distances among haplotypes in different populations, making them easier to detect than genes that experienced a single selective sweep, as was used for the simulations.

Despite these generally promising results from the simulations, many selected genes were not classified as outliers using the method and constitute false negatives. There are several synergistic reasons that genes affected by selection might not be classified as outliers, which include inherent limitations in outlier-based tests for selection and difficulties detecting selection (which does not always leave a clear signal), as well as specific details of the simulations (NIELSEN 2005; KELLEY *et al.* 2006). First, as pointed out previously, outlier analyses will detect only genes that stand out from the genome-wide distribution and thus might not detect genes experiencing weak selection or be applicable when much of the genome is under selection (MICHEL *et al.* 2010). However, this particular issue is more likely to affect empirical data than the simulated data sets. The data we simulated experienced selective sweeps with arbitrary linkage between the affected genetic region and the gene under selection. This means that selection could have favored the same or different haplotypes in each population. Thus, a clear molecular signature was not always left by the simulated selective sweeps. We simulated selection in this manner for computational efficiency and because it accurately reflects the effect of a recent and complete selective sweep. Specifically, we believe that this is a realistic model for selection as researchers will often detect selection on the basis

of genetic regions linked to the gene under selection rather than the actual gene under selection and patterns of linkage might vary among populations. Nonetheless, when evidence of selection comes directly from the target of selection or there is tight linkage with consistent patterns of linkage disequilibrium between the selected gene and a sequenced genetic region, a stronger signal of selection should be evident and selection should be easier to detect.

Evidence of selection in human populations: A diverse array of studies (Akey *et al.* 2002; Nielsen *et al.* 2007; Barreiro *et al.* 2008; Nielsen *et al.* 2009) has found that natural selection has played an important role in shaping functional genetic variation in humans. Analyses based on our model for quantifying the distribution of genetic variation among populations similarly find substantial evidence for the action of natural selection. We classified ~10 times as many genes as candidates for positive or divergent selection (high ϕ_{ST} outlier loci) than as candidates for balancing selection (low ϕ_{ST} outlier loci). However, this does not mean that divergent and positive selection more commonly affects human genetic variation than balancing selection. Evidence from other studies suggests that weak negative selection is prevalent in the human genome and balancing selection is also fairly common (Bustamante *et al.* 2005; Andrés *et al.* 2009). Instead this difference in the prevalence of different forms of selection might indicate that there is more variation in the strength of positive selection, resulting in a greater number of extreme genetic regions, than there is in the strength of negative or balancing selection with many genetic regions weakly affected by these factors. However, this cannot be explicitly addressed by our study. Finally, previous simulation studies suggest that it is difficult to detect balancing selection using outlier analyses (Beaumont and Balding 2004; Guo *et al.* 2009), which might reflect the different molecular signals left by divergent and balancing selection or the bounded nature of F_{ST} and ϕ_{ST}.

Specific genes identified as outliers by our analysis include several genes that have been repeatedly implicated as targets of positive selection in humans, such as POLG2, CYP3A5, and SLC24A5. Nonetheless, we also detected novel candidate genes for selection in humans (*e.g.*, FOXA2) and failed to detect selection on genes expected to have experienced strong selection on the basis of previous studies, such as the lactase (LCT) gene (Bersaglieri *et al.* 2004; Nielsen *et al.* 2007; Tishkoff *et al.* 2007). A lack of complete concordance with earlier studies is not surprising as a general lack of concordance among studies of selection in humans has been noted (Nielsen *et al.* 2007). This lack of concordance likely reflects the sensitivity of different methods to different signatures of selection, which affects whether methods are more likely to detect recent, ongoing, or more ancient selective sweeps, as well as the specific nucleotides and populations analyzed. For example, the lack of evidence from the worldwide human data set for selection on LCT appears to reflect the SNPs included in this study. Previous studies of human genetic variation have detected high values of F_{ST} at specific SNPs in the LCT gene (*e.g.*, 0.53 for SNPs rs4988235 and rs182549), but have also shown that F_{ST} varies across this gene (Bersaglieri *et al.* 2004). Previous estimates of F_{ST} for the two genic SNPs in LCT included in our analysis that were also investigated by Bersaglieri *et al.* (2004) were 0 (rs2874874) and 0.17 (rs2322659), which do not stand out markedly from background levels of differentiation.

Conclusions: Technological advances in DNA sequencing are ushering in a new era of population genomics. Whereas researchers previously were constrained to various, relatively low-throughput molecular markers for genotyping, it is now possible to rapidly generate very large volumes of DNA sequence data for any group of organisms (Mardis 2008a; Gilad *et al.* 2009). This new capability will allow researchers to address long-standing and fundamental questions in evolutionary biology, which were once considered nearly intractable because of limited genetic data (Mardis 2008b). However, most published NGS studies have been primarily descriptive (*e.g.*, transcriptome characterization; Vera *et al.* 2008; Parchman *et al.* 2010) or have been forced to discard valuable data because of analytical limitations (Hohenlohe *et al.* 2010) and have not taken full advantage of the potential of the sequence data. To do so, researchers need robust and accessible methods and models that can be applied to NGS data to test evolutionary hypotheses. The model presented here helps fill this gap, as it allows us to quantify heterogeneous genomic divergence among populations and identify genetic regions affected by selection. The Bayesian analysis of molecular variance illustrates the potential of combining appropriate models and NGS to address important questions in evolutionary biology and genetics and is a critical step toward utilizing the growing abundance of sequence data for population genomics.

We thank J. Fordyce, M. Forister, C. Nice, P. Nosil, T. Parchman, and R. Safran for discussion and comments on previous drafts of this article. Comments from L. Excoffier and two anonymous reviewers helped to improve the article. We are indebted to R. Williamson for his contributions to the development of the bamova software. This work was supported by National Science Foundation Division of Biological Infrastructure (DBI) award 0701757 (to C.A.B.).

LITERATURE CITED

Akey, J. M., G. Zhang, K. Zhang, L. Jin and M. D. Shriver, 2002 Interrogating a high-density SNP map for signatures of natural selection. Genome Res. **12:** 1805–1814.

Andolfatto, P., J. D. Wall and M. Kreitman, 1999 Unusual haplotype structure at the proximal breakpoint of in(2L)t in a natural population of *Drosophila melanogaster*. Genetics **153:** 1297–1311.

Andrés, A. M., M. J. Hubisz, A. Indap, D. G. Torgerson, J. D. Degenhardt *et al.*, 2009 Targets of balancing selection in the human genome. Mol. Biol. Evol. **26:** 2755–2764.

Bamshad, M., and S. P. Wooding, 2003 Signatures of natural selection in the human genome. Nat. Rev. Genet. **4:** 99–111.

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

Beaumont, M. A., 2005 Adaptation and speciation: What can F_{ST} tell us? Trends Ecol. Evol. **20:** 435–440.

Beaumont, M. A., and D. J. Balding, 2004 Identifying adaptive genetic divergence among populations from genome scans. Mol. Ecol. **13:** 969–980.

Beaumont, M. A., and R. A. Nichols, 1996 Evaluating loci for use in the genetic analysis of population structure. Proc. R. Soc. B Biol. Sci. **263:** 1619–1626.

Bersaglieri, T., P. C. Sabeti, N. Patterson, T. Vanderploeg, S. F. Schaffner *et al.*, 2004 Genetic signatures of strong recent positive selection at the lactase gene. Am. J. Hum. Genet. **74:** 1111–1120.

Bustamante, C. D., A. Fledel-Alon, S. Williamson, R. Nielsen, M. T. Hubisz *et al.*, 2005 Natural selection on protein-coding genes in the human genome. Nature **437:** 1153–1157.

Carlson, C. S., D. J. Thomas, M. A. Eberle, J. E. Swanson, R. J. Livingston *et al.*, 2005 Genomic regions exhibiting positive selection identified from dense genotype data. Genome Res. **15:** 1553–1565.

Chen, H., N. Patterson and D. Reich, 2010 Population differentiation as a test for selective sweeps. Genome Res. **20:** 393–402.

Craig, D. W., J. V. Pearson, S. Szelinger, A. Sekar, M. Redman *et al.*, 2008 Identification of genetic variants using bar-coded multiplexed sequencing. Nat. Methods **5:** 887–893.

Easton, R. M., H. Cho, K. Roovers, D. W. Shineman, M. Mizrahi *et al.*, 2005 Role for Akt3/protein kinase B gamma in attainment of normal brain size. Mol. Cell. Biol. **25:** 1869–1878.

Excoffier, L., P. E. Smouse and J. M. Quattro, 1992 Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics **131:** 479–491.

Excoffier, L., T. Hofer and M. Foll, 2009 Detecting loci under selection in a hierarchically structured population. Heredity **103:** 285–298.

Flint, J., J. Bond, D. C. Rees, A. J. Boyce, J. M. Roberts-Thomson *et al.*, 1999 Minisatellite mutational processes reduce F_{ST} estimates. Hum. Genet. **105:** 567–576.

Foll, M., and O. Gaggiotti, 2008 A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics **180:** 977–993.

Futschik, A., and C. Schlötterer, 2010 The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. Genetics **186:** 207–218.

Gilad, Y., J. K. Pritchard and K. Thornton, 2009 Characterizing natural variation using next-generation sequencing technologies. Trends Genet. **25:** 463–471.

Gillespie, J. H., 2000 Genetic drift in an infinite population: the pseudohitchhiking model. Genetics **155:** 909–919.

Gompert, Z., M. L. Forister, J. A. Fordyce, C. C. Nice, R. Williamson *et al.*, 2010 Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of *Lycaeides* butterflies. Mol. Ecol. **19:** 2455–2473.

Guo, F., D. K. Dey and K. E. Holsinger, 2009 A Bayesian hierarchical model for analysis of single-nucleotide polymorphisms diversity in multilocus, multipopulation samples. J. Am. Stat. Assoc. **104:** 142–154.

Hamblin, M. T., and A. Di Rienzo, 2000 Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am. J. Hum. Genet. **66:** 1669–1679.

Hamblin, M. T., E. E. Thompson and A. Di Rienzo, 2002 Complex signatures of natural selection at the Duffy blood group locus. Am. J. Hum. Genet. **70:** 369–383.

Hamm, D., B. S. Mautz, M. F. Wolfner, C. F. Aquadro and W. J. Swanson, 2007 Evidence of amino acid diversity-enhancing selection within humans and among primates at the candidate sperm-receptor gene PKDREJ. Am. J. Hum. Genet. **81:** 44–52.

Harris, M. I., 2001 Racial and ethnic differences in health care access and health outcomes for adults with type 2 diabetes. Diabetes Care **24:** 454–459.

Hohenlohe, P. A., S. Bassham, P. D. Etter, N. Stiffler, E. A. Johnson *et al.*, 2010 Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet. **6:** e1000862.

Holsinger, K. E., and B. S. Weir, 2009 Fundamental concepts in genetics: genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. Nat. Rev. Genet. **10:** 639–650.

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 *et al.*, 2008 Genotype, haplotype and copy-number variation in worldwide human populations. Nature **451:** 998–1003.

Kelley, J. L., J. Madeoy, J. C. Calhoun, W. Swanson and J. M. Akey, 2006 Genomic signatures of positive selection in humans and the limits of outlier approaches. Genome Res. **16:** 980–989.

Köttgen, A., N. L. Glazer, A. Dehghan, S.-J. Hwang, R. Katz *et al.*, 2009 Multiple loci associated with indices of renal function and chronic kidney disease. Nat. Genet. **41:** 712–717.

Kronholm, I., O. Loudet and J. de Meaux, 2010 Influence of mutation rate on estimators of genetic differentiation: lessons from *Arabidopsis thaliana*. BMC Genet. **11:** 33.

Kullback, S., and R. A. Leibler, 1951 On information and sufficiency. Ann. Math. Stat. **22:** 79–86.

Lavery, G. G., C. L. McTernan, S. C. Bain, T. A. Chowdhury, M. Hewison *et al.*, 2002 Association studies between the HSD11B2 gene (encoding human 11 beta-hydroxysteroid dehydrogenase type 2), type 1 diabetes mellitus and diabetic nephropathy. Eur. J. Endocrinol. **146:** 553–558.

Lewontin, R. C., and J. Krakauer, 1973 Distribution of gene frequency as a test of theory of selective neutrality of polymorphisms. Genetics **74:** 175–195.

Lohmueller, K. E., A. R. Indap, S. Schmidt, A. R. Boyko, R. D. Hernandez *et al.*, 2008 Proportionally more deleterious genetic variation in European than in African populations. Nature **451:** 994–997.

Lynch, M., 2009 Estimation of allele frequencies from high-coverage genome-sequencing projects. Genetics **182:** 295–301.

Mardis, E. R., 2008a The impact of next-generation sequencing technology on genetics. Trends Genet. **24:** 133–141.

Mardis, E. R., 2008b Next-generation DNA sequencing methods. Annu. Rev. Genomics Hum. Genet. **9:** 387–402.

Maynard-Smith, J., and J. Haigh, 1974 Hitch-hiking effect of a favorable gene. Genet. Res. **23:** 23–35.

Meyer, M., and M. Kircher, 2010 Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harbor Protoc. **2010:** pdb.prot5448.

Michel, A. P., S. Sim, T. H. Q. Powell, M. S. Taylor, P. Nosil *et al.*, 2010 Widespread genomic divergence during sympatric speciation. Proc. Natl. Acad. Sci. USA **107:** 9724–9729.

Nielsen, R., 2005 Molecular signatures of natural selection. Annu. Rev. Genet. **39:** 197–218.

Nielsen, R., I. Hellmann, M. Hubisz, C. Bustamante and A. G. Clark, 2007 Recent and ongoing selection in the human genome. Nat. Rev. Genet. **8:** 857–868.

Nielsen, R., M. J. Hubisz, I. Hellmann, D. Torgerson, A. M. Andres *et al.*, 2009 Darwinian and demographic forces affecting human protein coding genes. Genome Res. **19:** 838–849.

Novembre, J., T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko *et al.*, 2008 Genes mirror geography within Europe. Nature **456:** 98–101.

Parchman, T. L., K. S. Geist, J. A. Grahnen, C. W. Benkman and C. A. Buerkle, 2010 Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. BMC Genomics **11:** 180.

Puigserver, P., and J. T. Rodgers, 2006 FOXA2, a novel transcriptional regulator of insulin sensitivity. Nat. Med. **12:** 38–39.

Quinkler, M., and P. M. Stewart, 2003 Hypertension and the cortisol–cortisone shuttle. J. Clin. Endocrinol. Metab. **88:** 2384–2392.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Levine, D. J. Richter *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature **419:** 832–837.

Slatkin, M., 1987 Gene flow and the geographic structure of natural-populations. Science **236:** 787–792.

Slatkin, M., and T. Wiehe, 1998 Genetic hitch-hiking in a subdivided population. Genet. Res. **71:** 155–160.

Stahl, J. M., A. Sharma, M. Cheung, M. Zimmerman, J. Q. Cheng *et al.*, 2004 Deregulated Akt3 activity promotes development of malignant melanoma. Cancer Res. **64:** 7002–7010.

Stajich, J. E., and M. H. Hahn, 2005 Disentangling the effects of demography and selection in human history. Mol. Biol. Evol. **22:** 63–73.

Stephens, M., N. J. Smith and P. Donnelly, 2001 A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet. **68:** 978–989.

Tajima, F., 1989 Statistical-method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

Tishkoff, S. A., and B. C. Verrelli, 2003 Patterns of human genetic diversity: implications for human evolutionary history and disease. Annu. Rev. Genomics Hum. Genet. **4:** 293–340.

Tishkoff, S. A., F. A. Reed, A. Ranciaro, B. F. Voight, C. C. Babbitt *et al.*, 2007 Convergent adaptation of human lactase persistence in Africa and Europe. Nat. Genet. **39:** 31–40.

Tishkoff, S. A., F. A. Reed, F. R. Friedlaender, C. Ehret, A. Ranciaro *et al.*, 2009 The genetic structure and history of Africans and African Americans. Science **324:** 1035–1044.

Vera, J. C., C. W. Wheat, H. W. Fescemyer, M. J. Frilander, D. L. Crawford *et al.*, 2008 Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Mol. Ecol. **17:** 1636–1647.

Voight, B. F., S. Kudaravalli, X. Q. Wen and J. K. Pritchard, 2006 A map of recent positive selection in the human genome. PLoS Biol. **4:** 446–458.

Wang, H., A. Zhao, L. Chen, X. Zhong, J. Liao *et al.*, 2009 Human RIF1 encodes an anti-apoptotic factor required for DNA repair. Carcinogenesis **30:** 1314–1319.

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 Res. **15:** 1468–1476.

Wolfrum, C., E. Asilmaz, E. Luca, J. Friedman and M. Stoffel, 2004 FOXA2 regulates lipid metabolism and ketogenesis in the liver during fasting and in diabetes. Nature **432:** 1027–1032.

Wright, S., 1951 The genetical structure of populations. Ann. Eugen. **15:** 323–354.

Wu, D.-D., and Y.-P. Zhang, 2010 Positive selection drives population differentiation in the skeletal genes in modern humans. Hum. Mol. Genet. **19:** 2341–2346.

Xing, C., J. C. Cohen and E. Boerwinkle, 2010 A weighted false discovery rate control procedure reveals alleles at FOXA2 that influence fasting glucose levels. Am. J. Hum. Genet. **86:** 440–446.

Communicating editor: L. Excoffier

GENETICS

Supporting Information
http://www.genetics.org/cgi/content/full/genetics.110.124693/DC1

A Hierarchical Bayesian Model for Next-Generation Population Genomics

Zachariah Gompert and C. Alex Buerkle

DOI: 10.1534/genetics.110.124693

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.

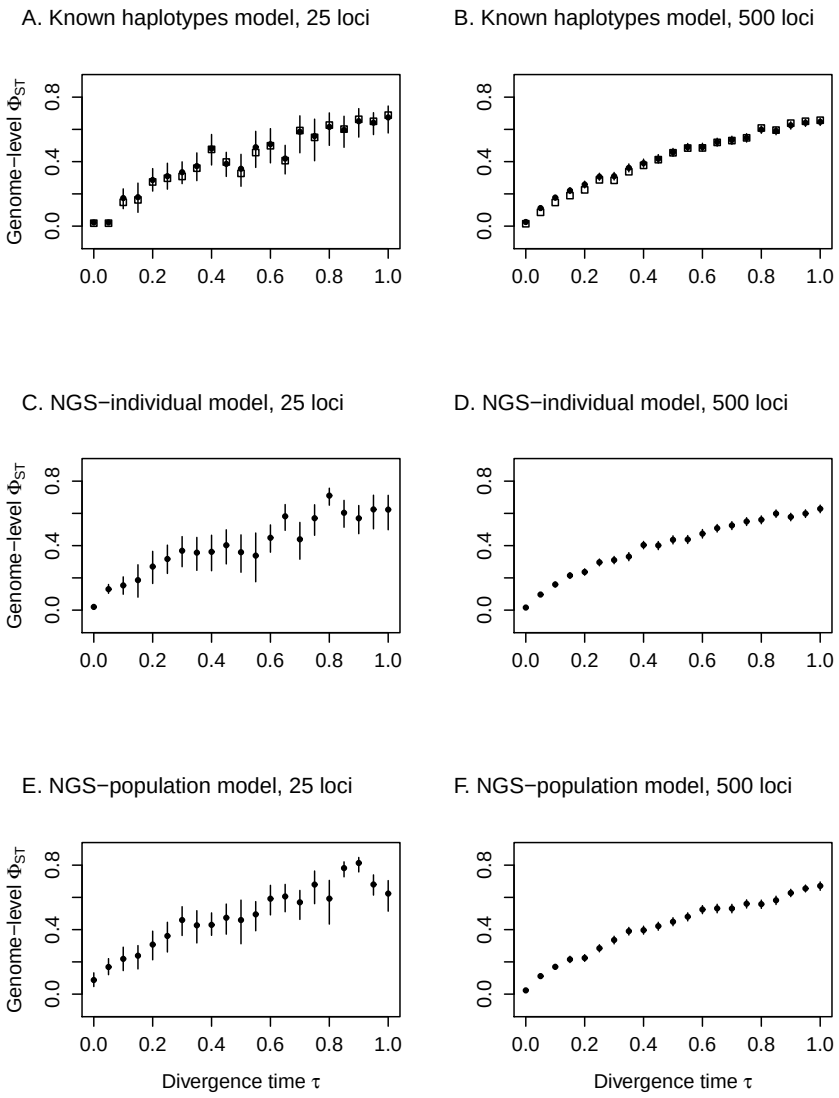

FIGURE S1.—Posterior probability distribution for mean genome-level ϕ_{ST} (μ_{ST}). The median (solid circle), 95% ETPI (vertical lines), and empirical mean ϕ_{ST} (open box, *known haplotypes model*) from a set of simulated data are shown in each plot.

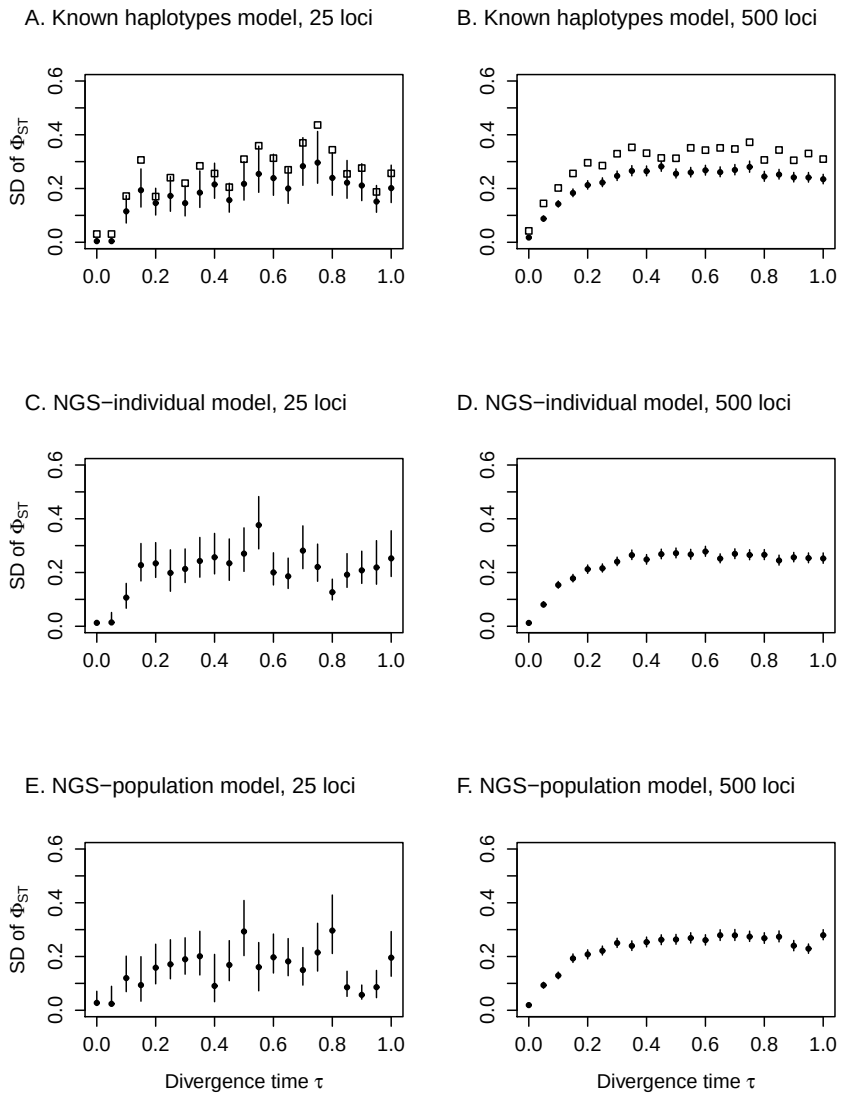

FIGURE S2.—Posterior probability distribution for the standard deviation of the genome-level ϕ_{ST} distribution. The median (solid circle), 95% ETPI (vertical lines), and empirical standard deviation of ϕ_{ST} (open box, known haplotypes model) are shown for each set of simulations.

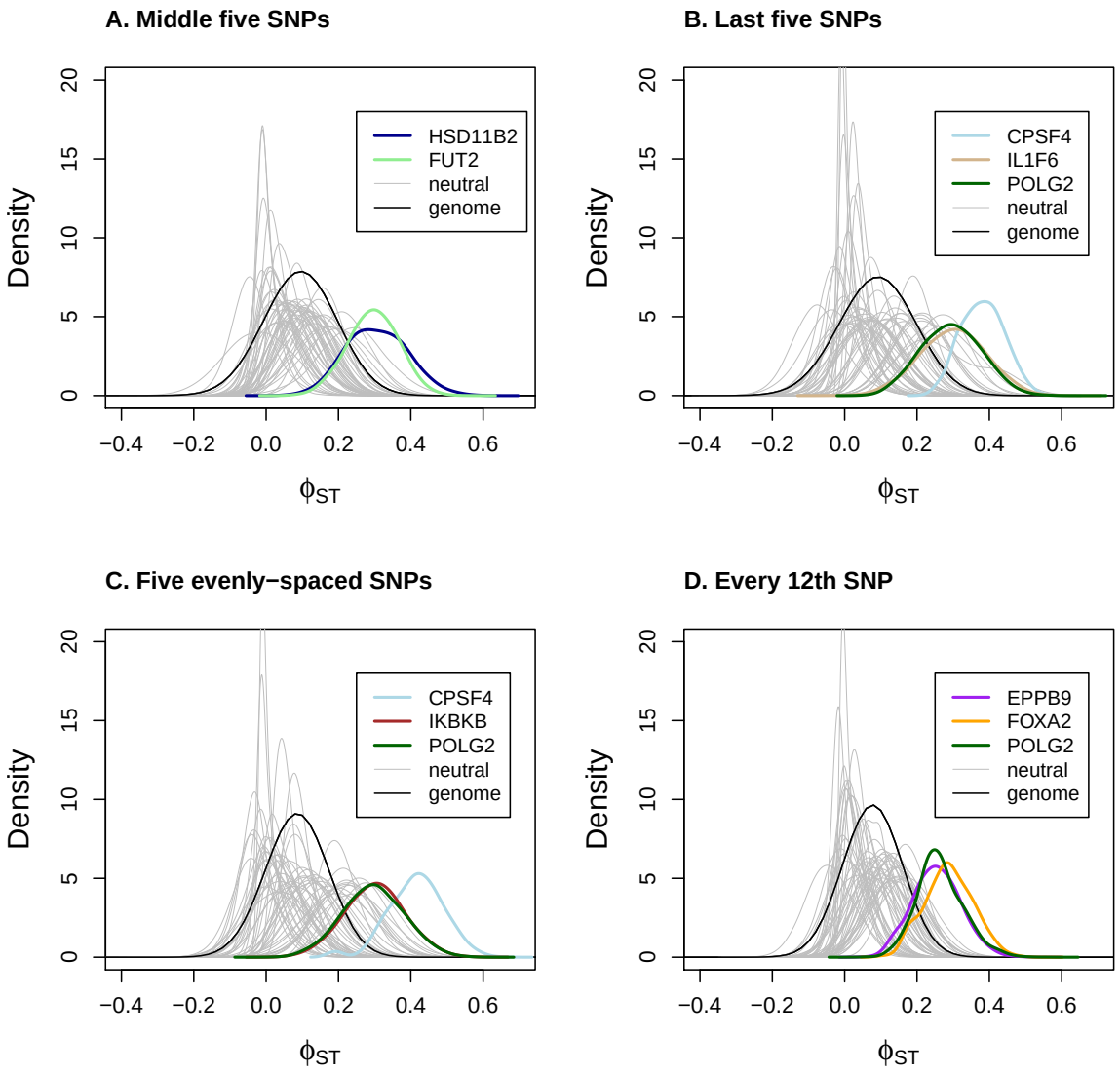

FIGURE S3.—Locus-specific ϕ_{ST} estimates for Africans and Europeans (SeattleSNPs data set). A point estimate of the genome-level ϕ_{ST} distribution (based on the median from the posterior probability distributions of α_{ST} and β_{ST}) is denoted with a solid black line. The posterior probability distributions for the outlier loci (colored lines) and 50 additional randomly chosen genetic regions (gray lines) are shown in each plot. Results from the middle five SNPs in each gene (A), the last five SNPs in each gene (B), five evenly spaced SNPs in each gene (C), and every 12^{th} in each gene (D) are shown.

TABLE S1

Proportion of outlier loci (*known haplotypes model*)

			Neutral								Selective Sweep							
			$m = 0$				$m = 2$				$m = 0$				$m = 2$			
			$a = 0.5$		$a = 0.95$		$a = 0.5$		$a = 0.95$		$a = 0.5$		$a = 0.95$		$a = 0.5$		$a = 0.95$	
S	No. loci	τ	low	high	low	high	low	high	low	high	low	high	low	high	low	high	low	high
1	25	0.25	0.013	0.039	0.000	0.004	0.017	0.004	0.000	0.000	0.000	0.050	0.000	0.000	0.050	0.150	0.000	0.000
		0.50	0.035	0.017	0.000	0.000	0.017	0.017	0.000	0.000	0.150	0.000	0.000	0.000	0.050	0.100	0.000	0.000
		1.00	0.030	0.009	0.000	0.000	0.013	0.026	0.000	0.000	0.050	0.200	0.050	0.000	0.000	0.250	0.000	0.050
	500	0.25	0.017	0.019	0.001	0.001	0.011	0.015	0.000	0.001	0.036	0.072	0.000	0.012	0.020	0.076	0.000	0.040
		0.50	0.022	0.021	0.004	0.002	0.013	0.014	0.001	0.001	0.052	0.084	0.012	0.012	0.028	0.124	0.004	0.044
		1.00	0.029	0.015	0.009	0.000	0.013	0.016	0.000	0.002	0.036	0.052	0.008	0.004	0.072	0.152	0.000	0.076
3	25	0.25	0.013	0.004	0.000	0.000	0.009	0.004	0.000	0.000	0.010	0.300	0.000	0.100	0.000	0.350	0.000	0.250
		0.50	0.030	0.000	0.000	0.000	0.009	0.013	0.000	0.000	0.050	0.300	0.000	0.150	0.000	0.450	0.000	0.100
		1.00	0.026	0.004	0.000	0.000	0.000	0.026	0.000	0.004	0.150	0.100	0.000	0.000	0.100	0.300	0.050	0.000
	500	0.25	0.011	0.008	0.000	0.001	0.009	0.008	0.000	0.001	0.032	0.260	0.000	0.124	0.052	0.324	0.000	0.196
		0.50	0.017	0.010	0.002	0.001	0.009	0.006	0.002	0.001	0.032	0.156	0.000	0.060	0.048	0.376	0.000	0.216
		1.00	0.029	0.012	0.009	0.000	0.009	0.010	0.001	0.001	0.016	0.160	0.004	0.036	0.060	0.340	0.008	0.204
5	25	0.25	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000
		0.50	0.004	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000
		1.00	0.030	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000
	500	0.25	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000
		0.50	0.010	0.003	0.002	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000
		1.00	0.030	0.012	0.010	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000

Mean proportion of neutrally evolving and selected loci over 10 replicates identified as outliers for the *known haplotypes model*.

TABLE S2
Proportion of outlier loci (*NGS–individual model*)

			Neutral								Selective Sweep							
			$m = 0$				$m = 2$				$m = 0$				$m = 2$			
			$a = 0.5$		$a = 0.95$		$a = 0.5$		$a = 0.95$		$a = 0.5$		$a = 0.95$		$a = 0.5$		$a = 0.95$	
S	No. loci	τ	low	high	low	high	low	high	low	high	low	high	low	high	low	high	low	high
1	25	0.25	0.009	0.022	0.000	0.000	0.013	0.009	0.000	0.000	0.000	0.100	0.000	0.000	0.050	0.150	0.000	0.000
		0.50	0.022	0.013	0.000	0.004	0.000	0.013	0.000	0.000	0.000	0.050	0.000	0.000	0.050	0.050	0.000	0.000
		1.00	0.030	0.017	0.009	0.004	0.013	0.017	0.000	0.000	0.100	0.000	0.000	0.000	0.100	0.050	0.000	0.000
	500	0.25	0.012	0.021	0.001	0.003	0.005	0.017	0.000	0.003	0.008	0.044	0.000	0.008	0.012	0.064	0.000	0.012
		0.50	0.023	0.016	0.003	0.002	0.011	0.019	0.000	0.002	0.016	0.036	0.004	0.000	0.008	0.064	0.000	0.000
		1.00	0.031	0.007	0.008	0.001	0.011	0.018	0.000	0.002	0.036	0.004	0.008	0.000	0.028	0.076	0.000	0.008
3	25	0.25	0.004	0.022	0.000	0.000	0.004	0.026	0.000	0.000	0.050	0.250	0.050	0.000	0.000	0.100	0.000	0.050
		0.50	0.018	0.022	0.004	0.004	0.017	0.004	0.000	0.000	0.050	0.000	0.000	0.000	0.000	0.400	0.000	0.100
		1.00	0.043	0.009	0.000	0.000	0.009	0.009	0.000	0.000	0.000	0.000	0.000	0.000	0.100	0.250	0.000	0.0500
	500	0.25	0.009	0.015	0.001	0.002	0.008	0.013	0.000	0.003	0.024	0.160	0.000	0.032	0.012	0.248	0.000	0.072
		0.50	0.018	0.014	0.001	0.002	0.009	0.013	0.000	0.002	0.036	0.076	0.004	0.008	0.044	0.220	0.000	0.056
		1.00	0.032	0.009	0.008	0.001	0.010	0.013	0.001	0.001	0.056	0.032	0.004	0.008	0.036	0.216	0.000	0.044
5	25	0.25	0.000	0.000	0.000	0.000	0.004	0.000	0.000	0.000	0.000	0.650	0.000	0.150	0.000	0.850	0.000	0.350
		0.50	0.013	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.300	0.000	0.100	0.000	0.950	0.000	0.3000
		1.00	0.022	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.200	0.000	0.100	0.000	0.850	0.000	0.400
	500	0.25	0.011	0.006	0.000	0.001	0.006	0.002	0.000	0.000	0.000	0.656	0.000	0.2888	0.000	0.888	0.000	0.440
		0.50	0.018	0.007	0.002	0.001	0.005	0.003	0.000	0.001	0.000	0.308	0.000	0.108	0.000	0.824	0.000	0.416
		1.00	0.031	0.008	0.009	0.001	0.009	0.004	0.000	0.001	0.000	0.060	0.000	0.004	0.000	0.792	0.000	0.324

Mean proportion of neutrally evolving and selected loci over 10 replicates identified as outliers for the *NGS-individual model*.

TABLE S3
Proportion of outlier loci (*NGS-population model*)

			Neutral								Selective Sweep							
			$m = 0$				$m = 2$				$m = 0$				$m = 2$			
			$a = 0.5$		$a = 0.95$		$a = 0.5$		$a = 0.95$		$a = 0.5$		$a = 0.95$		$a = 0.5$		$a = 0.95$	
S	No. loci	τ	low	high	low	high	low	high	low	high	low	high	low	high	low	high	low	high
1	25	0.25	0.030	0.013	0.000	0.000	0.000	0.017	0.000	0.000	0.050	0.100	0.000	0.050	0.000	0.150	0.000	0.000
		0.50	0.013	0.009	0.000	0.000	0.004	0.017	0.000	0.000	0.050	0.050	0.000	0.000	0.000	0.200	0.000	0.000
		1.00	0.052	0.013	0.000	0.000	0.009	0.009	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.050	0.100	0.000
	500	0.25	0.014	0.018	0.000	0.002	0.006	0.012	0.000	0.001	0.024	0.108	0.000	0.016	0.012	0.084	0.000	0.016
		0.50	0.020	0.013	0.001	0.001	0.008	0.015	0.000	0.001	0.036	0.048	0.008	0.004	0.016	0.096	0.000	0.016
		1.00	0.036	0.006	0.004	0.000	0.009	0.017	0.000	0.002	0.044	0.016	0.008	0.000	0.020	0.084	0.000	0.032
3	25	0.25	0.013	0.009	0.000	0.000	0.004	0.000	0.000	0.000	0.000	0.350	0.000	0.050	0.000	0.350	0.000	0.050
		0.50	0.022	0.013	0.000	0.000	0.009	0.004	0.000	0.000	0.000	0.100	0.000	0.000	0.100	0.150	0.000	0.000
		1.00	0.030	0.004	0.009	0.000	0.004	0.004	0.000	0.000	0.000	0.050	0.000	0.000	0.050	0.300	0.000	0.050
	500	0.25	0.019	0.014	0.000	0.002	0.007	0.010	0.000	0.000	0.028	0.248	0.000	0.080	0.044	0.308	0.000	0.132
		0.50	0.016	0.013	0.001	0.000	0.009	0.009	0.000	0.000	0.048	0.140	0.012	0.024	0.048	0.388	0.000	0.160
		1.00	0.033	0.005	0.004	0.000	0.006	0.011	0.000	0.001	0.036	0.040	0.000	0.004	0.028	0.336	0.000	0.136
5	25	0.25	0.004	0.000	0.000	0.000	0.000	0.004	0.000	0.000	0.000	0.800	0.000	0.250	0.000	0.750	0.000	0.350
		0.50	0.022	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.450	0.000	0.050	0.000	0.850	0.000	0.550
		1.00	0.030	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.250	0.000	0.100	0.000	0.900	0.000	0.4000
	500	0.25	0.003	0.001	0.000	0.000	0.001	0.001	0.000	0.000	0.000	0.420	0.000	0.240	0.000	0.724	0.000	0.552
		0.50	0.009	0.000	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.240	0.000	0.160	0.000	0.676	0.000	0.532
		1.00	0.028	0.002	0.005	0.000	0.001	0.000	0.000	0.000	0.000	0.096	0.000	0.052	0.000	0.520	0.000	0.368

Mean proportion of neutrally evolving and selected loci over 10 replicates identified as outliers for the *NGS-population model*.

TABLE S4
Proportion of outlier loci (group structure) with the *NGS–individual model*

			Neutral				Selective Sweep			
			$a = 0.5$	$a = 0.5$	$a = 0.95$	$a = 0.95$	$a = 0.5$	$a = 0.5$	$a = 0.95$	$a = 0.95$
S_G	No. loci	ϕ	low	high	low	high	low	high	low	high
1	25	ϕ_{ST}	0.004	0.030	0.002	0.009	0.000	0.025	0.000	0.000
		ϕ_{CT}	0.009	0.028	0.000	0.004	0.025	0.025	0.000	0.000
		ϕ_{SC}	0.007	0.024	0.000	0.000	0.000	0.150	0.000	0.025
	500	ϕ_{ST}	0.014	0.032	0.003	0.008	0.016	0.012	0.000	0.000
		ϕ_{CT}	0.019	0.032	0.003	0.009	0.048	0.008	0.012	0.000
		ϕ_{SC}	0.009	0.010	0.000	0.000	0.004	0.120	0.000	0.044
2	25	ϕ_{ST}	0.004	0.002	0.000	0.000	0.000	0.925	0.000	0.875
		ϕ_{CT}	0.020	0.017	0.002	0.000	0.225	0.100	0.175	0.025
		ϕ_{SC}	0.000	0.000	0.000	0.000	0.000	0.950	0.000	0.900
	500	ϕ_{ST}	0.006	0.002	0.000	0.000	0.000	1.000	0.000	0.988
		ϕ_{CT}	0.020	0.029	0.004	0.009	0.164	0.108	0.160	0.100
		ϕ_{SC}	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.988

Mean proportion of neutrally evolving and selected loci over 10 replicates identified as outliers with group structure for the *NGS-individual model*.

TABLE S5

Summary of outlier analysis for the worldwide human SNP data (Jakobsson *et al.* 2008). For each genetic region we give the chromosome number (Chrom.), gene ID (Gene), and classification as a low (l) or high (h) ϕ_{ST} outlier as well as whether outlier status was at $a = 0.5$ or $a = 0.95$. Additionally we provide the median and credible intervals for each locus's ϕ_{ST} and quantile in the genome-level distribution.

Chrom.	Gene	Class.	Quant. (lb)	Quant. (median)	Quant. (ub)	ϕ_{ST} (lb)	ϕ_{ST} (median)	ϕ_{ST} (ub)
1	AKT3	l–0.5	0.0015	0.0167	0.1217	-0.1017	-0.0455	0.0201
1	C1orf116	l–0.5	0.0016	0.0190	0.1253	-0.1007	-0.0420	0.0209
1	C1orf117	l–0.5	0.0018	0.0137	0.0711	-0.0981	-0.0508	-0.0009
1	GPR161	l–0.5	0.0001	0.0061	0.1521	-0.1631	-0.0712	0.0297
1	GSTM4	l–0.5	0.0011	0.0205	0.1673	-0.1089	-0.0399	0.0335
1	MYCBP	l–0.5	0.0019	0.0146	0.0828	-0.0969	-0.0493	0.0048
1	PLA2G2D	l–0.5	0.0014	0.0243	0.1920	-0.1036	-0.0351	0.0399
1	ACADM	h–0.5	0.9536	0.9938	0.9995	0.2119	0.2655	0.3168
1	AIM1L	h–0.5	0.9249	0.9851	0.9981	0.1960	0.2441	0.2906
1	AP4B1	h–0.5	0.9489	0.9951	0.9997	0.2087	0.2710	0.3267
1	ASPM	h–0.5	0.8376	0.9752	0.9981	0.1653	0.2304	0.2903
1	ATPAF1	h–0.5	0.9639	0.9971	0.9999	0.2192	0.2824	0.3420
1	C1orf113	h–0.5	0.9074	0.9882	0.9992	0.1883	0.2500	0.3076
1	C1orf114	h–0.5	0.8921	0.9770	0.9973	0.1825	0.2325	0.2834
1	C1orf123	h–0.5	0.9202	0.9903	0.9994	0.1935	0.2550	0.3111
1	C1orf161	h–0.5	0.9433	0.9886	0.9987	0.2055	0.2510	0.2984
1	C1orf83	h–0.5	0.8171	0.9774	0.9987	0.1603	0.2330	0.2983
1	C1orf83	h–0.5	0.9513	0.9960	0.9998	0.2101	0.2755	0.3345
1	CD34	h–0.5	0.9540	0.9915	0.9990	0.2125	0.2581	0.3017
1	ELA3B	h–0.5	0.8788	0.9763	0.9973	0.1782	0.2318	0.2832
1	FCER1A	h–0.5	0.9222	0.9829	0.9975	0.1951	0.2405	0.2846
1	GPR88	h–0.5	0.9547	0.9900	0.9984	0.2127	0.2540	0.2942

Continued on next page

TABLE S5 – continued from previous page

Chrom.	Gene	Class.	Quant. (lb)	Quant. (median)	Quant. (ub)	ϕ_{ST} (lb)	ϕ_{ST} (median)	ϕ_{ST} (ub)
1	HMCN1	h–0.5	0.9608	0.9982	1.0000	0.2174	0.2915	0.3563
1	HNRPR	h–0.5	0.9246	0.9884	0.9989	0.1960	0.2506	0.3016
1	ILF2	h–0.5	0.9176	0.9845	0.9981	0.1926	0.2431	0.2907
1	INSRR	h–0.5	0.8887	0.9753	0.9965	0.1814	0.2304	0.2779
1	KCNT2	h–0.5	0.9559	0.9981	1.0000	0.2131	0.2912	0.3615
1	KIAA0319L	h–0.5	0.8693	0.9802	0.9986	0.1750	0.2366	0.2958
1	LRRC7	h–0.5	0.9286	0.9898	0.9993	0.1978	0.2539	0.3104
1	MEF2D	h–0.5	0.9292	0.9855	0.9982	0.1980	0.2449	0.2914
1	NBL1	h–0.5	0.9640	0.9929	0.9991	0.2204	0.2627	0.3031
1	NBPF3	h–0.5	0.9042	0.9777	0.9967	0.1870	0.2334	0.2789
1	ORC1L	h–0.5	0.9018	0.9832	0.9985	0.1860	0.2410	0.2956
1	PHTF1	h–0.5	0.9478	0.9909	0.9991	0.2078	0.2563	0.3049
1	PSMA5	h–0.5	0.9569	0.9947	0.9996	0.2142	0.2688	0.3209
1	RGL1	h–0.5	0.9513	0.9930	0.9995	0.2103	0.2627	0.3141
1	SNX27	h–0.5	0.9244	0.9779	0.9954	0.1959	0.2337	0.2721
1	TRIM46	h–0.5	0.9376	0.9838	0.9971	0.2023	0.2420	0.2815
1	TM2D1	l–0.95	0.0008	0.0051	0.0273	-0.1154	-0.0755	-0.0320
1	C1orf41	h–0.95	0.9968	0.9998	1.0000	0.2798	0.3322	0.3808
1	CLCNKB	h–0.95	0.9859	0.9988	0.9999	0.2456	0.2997	0.3505
1	CLSPN	h–0.95	0.9975	0.9998	1.0000	0.2855	0.3343	0.3793
1	EIF2C1	h–0.95	0.9999	1.0000	1.0000	0.3387	0.3881	0.4335
1	EIF2C3	h–0.95	0.9961	0.9997	1.0000	0.2760	0.3237	0.3695
1	EPS15	h–0.95	0.9862	0.9989	1.0000	0.2461	0.3010	0.3552
1	FRRS1	h–0.95	0.9710	0.9957	0.9997	0.2256	0.2740	0.3218
1	KHDRBS1	h–0.95	0.9757	0.9984	1.0000	0.2309	0.2947	0.3554
1	KIAA0319L	h–0.95	0.9664	0.9963	0.9998	0.2214	0.2768	0.3306
1	KIAA0907	h–0.95	0.9712	0.9989	1.0000	0.2261	0.3016	0.3673
1	MASP2	h–0.95	0.9899	0.9989	0.9999	0.2540	0.3015	0.3473
1	OSBPL9	h–0.95	0.9796	0.9986	1.0000	0.2354	0.2974	0.3520

Continued on next page

TABLE S5 – continued from previous page

Chrom.	Gene	Class.	Quant. (lb)	Quant. (median)	Quant. (ub)	ϕ_{ST} (lb)	ϕ_{ST} (median)	ϕ_{ST} (ub)
1	PMVK	h–0.95	0.9848	0.9998	1.0000	0.2434	0.3282	0.4002
1	POLR3GL	h–0.95	0.9871	0.9997	1.0000	0.2475	0.3256	0.3922
1	RABGAP1L	h–0.95	0.9755	0.9974	0.9999	0.2311	0.2844	0.3359
1	RABGGTB	h–0.95	0.9757	0.9961	0.9996	0.2311	0.2757	0.3174
1	RBBP5	h–0.95	0.9943	0.9996	1.0000	0.2673	0.3194	0.3679
1	RC3H1	h–0.95	0.9912	0.9994	1.0000	0.2571	0.3129	0.3632
1	SH3BP5L	h–0.95	0.9982	0.9999	1.0000	0.2922	0.3371	0.3799
1	SHE	h–0.95	0.9889	0.9990	0.9999	0.2514	0.3028	0.3492
1	SLC16A1	h–0.95	0.9896	0.9993	1.0000	0.2533	0.3095	0.3599
1	SLC9A11	h–0.95	0.9987	1.0000	1.0000	0.2977	0.3642	0.4257
1	SYF2	h–0.95	0.9998	1.0000	1.0000	0.3325	0.4019	0.4630
1	UBE2J2	h–0.95	0.9767	0.9996	1.0000	0.2321	0.3199	0.3941
1	YY1AP1	h–0.95	0.9811	0.9996	1.0000	0.2375	0.3189	0.3886
1	ZBTB41	h–0.95	0.9975	0.9998	1.0000	0.2856	0.3355	0.3850
1	ZMYM4	h–0.95	0.9860	0.9982	0.9999	0.2458	0.2920	0.3372
1	ZMYM6	h–0.95	0.9997	1.0000	1.0000	0.3267	0.3962	0.4585
2	EIF2AK2	l–0.5	0.0010	0.0145	0.1055	-0.0956	-0.0384	0.0216
2	MGC39518	l–0.5	0.0003	0.0125	0.1879	-0.1204	-0.0421	0.0451
2	SMYD5	l–0.5	0.0049	0.0219	0.0819	-0.0632	-0.0278	0.0123
2	XPO1	l–0.5	0.0003	0.0164	0.2286	-0.1175	-0.0352	0.0541
2	ALK	h–0.5	0.8903	0.9772	0.9971	0.1800	0.2279	0.2737
2	CAB39	h–0.5	0.9596	0.9938	0.9996	0.2123	0.2590	0.3075
2	CHST10	h–0.5	0.8857	0.9832	0.9988	0.1785	0.2357	0.2910
2	CREG2	h–0.5	0.8386	0.9791	0.9986	0.1645	0.2304	0.2885
2	CTDSP1	h–0.5	0.9642	0.9928	0.9990	0.2160	0.2556	0.2937
2	FLJ20758	h–0.5	0.9261	0.9873	0.9987	0.1937	0.2427	0.2892
2	KYNU	h–0.5	0.9414	0.9933	0.9997	0.2005	0.2571	0.3138
2	LOC51252	h–0.5	0.9550	0.9897	0.9984	0.2096	0.2476	0.2849
2	MALL	h–0.5	0.9563	0.9971	0.9999	0.2097	0.2747	0.3367

Continued on next page

TABLE S5 – continued from previous page

Chrom.	Gene	Class.	Quant. (lb)	Quant. (median)	Quant. (ub)	ϕ_{ST} (lb)	ϕ_{ST} (median)	ϕ_{ST} (ub)
2	MKI67IP	h–0.5	0.9526	0.9959	0.9998	0.2076	0.2674	0.3228
2	ORMDL1	h–0.5	0.8793	0.9895	0.9997	0.1765	0.2470	0.3128
2	REEP1	h–0.5	0.8450	0.9847	0.9994	0.1661	0.2382	0.3043
2	REEP1	h–0.5	0.9332	0.9911	0.9994	0.1968	0.2507	0.3046
2	RHBDD1	h–0.5	0.9498	0.9937	0.9995	0.2065	0.2583	0.3075
2	RNF149	h–0.5	0.9336	0.9899	0.9989	0.1970	0.2478	0.2930
2	SELI	h–0.5	0.9476	0.9964	0.9999	0.2044	0.2702	0.3299
2	STK11IP	h–0.5	0.8709	0.9848	0.9991	0.1736	0.2381	0.2956
2	SULT1C2	h–0.5	0.9477	0.9948	0.9997	0.2048	0.2625	0.3164
2	TMEFF2	h–0.5	0.9186	0.9854	0.9981	0.1904	0.2391	0.2841
2	TRPM8	h–0.5	0.8220	0.9842	0.9995	0.1605	0.2373	0.3055
2	ZNF512	h–0.5	0.9559	0.9948	0.9997	0.2100	0.2627	0.3143
2	RIF1	l–0.95	0.0006	0.0048	0.0267	-0.1051	-0.0646	-0.0226
2	AOX2	h–0.95	0.9840	0.9980	0.9999	0.2371	0.2823	0.3275
2	BIN1	h–0.95	0.9919	0.9994	1.0000	0.2525	0.3044	0.3556
2	C2orf13	h–0.95	0.9994	1.0000	1.0000	0.3041	0.3569	0.4085
2	CMKOR1	h–0.95	0.9707	0.9967	0.9998	0.2216	0.2720	0.3213
2	KIAA1212	h–0.95	0.9818	0.9990	1.0000	0.2333	0.2946	0.3490
2	PROKR1	h–0.95	0.9948	0.9997	1.0000	0.2625	0.3165	0.3658
2	PSCDBP	h–0.95	0.9968	0.9998	1.0000	0.2725	0.3220	0.3735
2	RAB3GAP1	h–0.95	0.9625	0.9986	1.0000	0.2137	0.2881	0.3572
2	SMC6L1	h–0.95	0.9960	1.0000	1.0000	0.2677	0.3464	0.4152
2	SMYD1	h–0.95	1.0000	1.0000	1.0000	0.3423	0.3974	0.4457
2	ZRANB3	h–0.95	0.9765	0.9992	1.0000	0.2268	0.2993	0.3597
3	CEP63	l–0.5	0.0015	0.0108	0.0543	-0.0848	-0.0445	-0.0017
3	GPR156	l–0.5	0.0023	0.0122	0.0503	-0.0767	-0.0415	-0.0044
3	ARL6	h–0.5	0.9189	0.9826	0.9977	0.1864	0.2296	0.2720
3	B3GNT5	h–0.5	0.9435	0.9935	0.9996	0.1977	0.2522	0.3032
3	BCL6	h–0.5	0.9323	0.9882	0.9987	0.1927	0.2388	0.2832

Continued on next page

TABLE S5 – continued from previous page

Chrom.	Gene	Class.	Quant. (lb)	Quant. (median)	Quant. (ub)	ϕ_{ST} (lb)	ϕ_{ST} (median)	ϕ_{ST} (ub)
3	C3orf17	h–0.5	0.9370	0.9900	0.9991	0.1943	0.2426	0.2906
3	C3orf25	h–0.5	0.9158	0.9838	0.9983	0.1847	0.2315	0.2776
3	DNAH1	h–0.5	0.9268	0.9860	0.9985	0.1898	0.2347	0.2800
3	FLJ25996	h–0.5	0.9350	0.9882	0.9986	0.1935	0.2389	0.2821
3	MAPKAPK3	h–0.5	0.8826	0.9899	0.9997	0.1736	0.2424	0.3076
3	NGLY1	h–0.5	0.9304	0.9885	0.9990	0.1910	0.2392	0.2876
3	NR2C2	h–0.5	0.9094	0.9770	0.9961	0.1831	0.2227	0.2617
3	PCAF	h–0.5	0.9187	0.9856	0.9985	0.1866	0.2340	0.2793
3	PLCXD2	h–0.5	0.9041	0.9751	0.9963	0.1803	0.2207	0.2623
3	PLS1	h–0.5	0.9424	0.9894	0.9989	0.1976	0.2415	0.2861
3	RASA2	h–0.5	0.8803	0.9806	0.9986	0.1725	0.2271	0.2824
3	RBP2	h–0.5	0.8786	0.9843	0.9991	0.1718	0.2322	0.2907
3	UNQ846	h–0.5	0.9097	0.9806	0.9975	0.1829	0.2271	0.2704
3	BTLA	h–0.95	0.9910	0.9993	1.0000	0.2450	0.2953	0.3438
3	C3orf23	h–0.95	0.9959	0.9997	1.0000	0.2624	0.3097	0.3539
3	C3orf37	h–0.95	0.9697	0.9977	0.9999	0.2157	0.2733	0.3248
3	C3orf58	h–0.95	0.9842	0.9992	1.0000	0.2326	0.2921	0.3468
3	COPG	h–0.95	0.9941	0.9997	1.0000	0.2541	0.3098	0.3629
3	GNAI2	h–0.95	0.9980	0.9999	1.0000	0.2761	0.3260	0.3715
3	GORASP1	h–0.95	0.9995	1.0000	1.0000	0.2997	0.3492	0.3956
3	LMLN	h–0.95	0.9778	0.9981	0.9999	0.2234	0.2769	0.3268
3	LTF	h–0.95	0.9766	0.9978	0.9999	0.2225	0.2743	0.3243
3	MSL2L1	h–0.95	0.9707	0.9983	1.0000	0.2165	0.2792	0.3377
3	NPCDR1	h–0.95	0.9794	0.9985	0.9999	0.2254	0.2813	0.3323
3	RBP1	h–0.95	0.9772	0.9971	0.9998	0.2232	0.2689	0.3134
3	TRAK1	h–0.95	0.9983	1.0000	1.0000	0.2783	0.3344	0.3882
4	KLHL8	l–0.5	0.0001	0.0080	0.1424	-0.1536	-0.0674	0.0264
4	C4orf18	h–0.5	0.9279	0.9814	0.9966	0.2011	0.2425	0.2824
4	CCDC4	h–0.5	0.9555	0.9926	0.9993	0.2178	0.2660	0.3140

Continued on next page

TABLE S5 – continued from previous page

Chrom.	Gene	Class.	Quant. (lb)	Quant. (median)	Quant. (ub)	ϕ_{ST} (lb)	ϕ_{ST} (median)	ϕ_{ST} (ub)
4	HCAP-G	h–0.5	0.9144	0.9843	0.9984	0.1941	0.2475	0.2979
4	LEF1	h–0.5	0.9273	0.9826	0.9974	0.2010	0.2444	0.2878
4	MOBKL1A	h–0.5	0.9038	0.9853	0.9990	0.1899	0.2490	0.3071
4	NUDT6	h–0.5	0.9562	0.9924	0.9992	0.2182	0.2652	0.3109
4	NUP54	h–0.5	0.9501	0.9935	0.9996	0.2135	0.2693	0.3228
4	STX18	h–0.5	0.9399	0.9877	0.9986	0.2078	0.2536	0.3022
4	TBC1D19	h–0.5	0.9401	0.9911	0.9993	0.2076	0.2618	0.3156
4	TLR10	h–0.5	0.8653	0.9763	0.9978	0.1768	0.2357	0.2920
4	ARHGAP10	h–0.95	0.9977	0.9998	1.0000	0.2930	0.3374	0.3778
4	CENTD1	h–0.95	0.9902	0.9984	0.9999	0.2597	0.3007	0.3447
4	DCK	h–0.95	0.9999	1.0000	1.0000	0.3526	0.4099	0.4624
4	ENAM	h–0.95	0.9993	1.0000	1.0000	0.3171	0.3827	0.4396
4	ESSPL	h–0.95	0.9755	0.9970	0.9998	0.2361	0.2863	0.3362
4	ETFDH	h–0.95	0.9992	1.0000	1.0000	0.3127	0.3954	0.4679
4	LRBA	h–0.95	0.9765	0.9975	0.9999	0.2361	0.2906	0.3455
4	NPNT	h–0.95	0.9859	0.9989	1.0000	0.2508	0.3070	0.3613
4	PHF22	h–0.95	0.9985	0.9999	1.0000	0.3024	0.3501	0.3943
4	POLN	h–0.95	0.9869	0.9984	0.9999	0.2516	0.3001	0.3454
4	SLC30A9	h–0.95	0.9916	0.9991	1.0000	0.2641	0.3107	0.3573
4	TMEM34	h–0.95	0.9989	1.0000	1.0000	0.3075	0.3612	0.4108
5	DNAJA5	l–0.5	0.0010	0.0071	0.0386	-0.0917	-0.0542	-0.0132
5	C6	h–0.5	0.8581	0.9869	0.9996	0.1593	0.2280	0.2928
5	CD180	h–0.5	0.8830	0.9792	0.9981	0.1674	0.2171	0.2665
5	GMCL1L	h–0.5	0.9080	0.9871	0.9991	0.1753	0.2284	0.2809
5	HSPA4	h–0.5	0.9065	0.9840	0.9984	0.1748	0.2236	0.2694
5	IL4	h–0.5	0.9003	0.9792	0.9975	0.1729	0.2173	0.2615
5	LOC51334	h–0.5	0.9439	0.9887	0.9986	0.1908	0.2315	0.2708
5	LOC90624	h–0.5	0.9509	0.9917	0.9993	0.1951	0.2384	0.2826
5	MARCH6	h–0.5	0.9528	0.9897	0.9985	0.1963	0.2336	0.2703

Continued on next page

TABLE S5 – continued from previous page

Chrom.	Gene	Class.	Quant. (lb)	Quant. (median)	Quant. (ub)	ϕ_{ST} (lb)	ϕ_{ST} (median)	ϕ_{ST} (ub)
5	SKP2	h–0.5	0.9493	0.9936	0.9996	0.1934	0.2439	0.2916
5	SRFBP1	h–0.5	0.8352	0.9788	0.9988	0.1541	0.2166	0.2760
5	GPBP1	l–0.95	0.0002	0.0023	0.0206	-0.1175	-0.0774	-0.0293
5	RBM22	l–0.95	0.0000	0.0004	0.0386	-0.1900	-0.1092	-0.0118
5	HINT1	h–0.95	0.9818	0.9987	1.0000	0.2207	0.2745	0.3265
5	LOC92270	h–0.95	0.9862	0.9994	1.0000	0.2273	0.2886	0.3431
5	MGC23985	h–0.95	0.9958	0.9998	1.0000	0.2533	0.3026	0.3497
5	MSX2	h–0.95	0.9929	0.9997	1.0000	0.2417	0.2991	0.3511
5	NSUN2	h–0.95	0.9796	0.9961	0.9996	0.2176	0.2540	0.2929
5	PITX1	h–0.95	0.9988	1.0000	1.0000	0.2753	0.3344	0.3864
5	ROPN1L	h–0.95	0.9688	0.9952	0.9996	0.2080	0.2498	0.2925
5	SDHA	h–0.95	0.9773	0.9974	0.9998	0.2159	0.2619	0.3059
5	SLC36A2	h–0.95	0.9893	0.9996	1.0000	0.2325	0.2929	0.3505
5	WDR70	h–0.95	0.9943	0.9997	1.0000	0.2455	0.2979	0.3491
6	FRS3	l–0.5	0.0003	0.0041	0.0388	-0.1045	-0.0605	-0.0098
6	GCM2	l–0.5	0.0014	0.0173	0.1098	-0.0787	-0.0300	0.0207
6	KIFC1	l–0.5	0.0003	0.0099	0.1468	-0.1083	-0.0425	0.0312
6	MYCT1	l–0.5	0.0026	0.0239	0.1380	-0.0682	-0.0222	0.0287
6	TDRD6	l–0.5	0.0004	0.0053	0.0448	-0.1009	-0.0555	-0.0063
6	ZNF187	l–0.5	0.0030	0.0223	0.1170	-0.0654	-0.0239	0.0232
6	C6orf146	h–0.5	0.9155	0.9867	0.9989	0.1696	0.2166	0.2631
6	C6orf70	h–0.5	0.8524	0.9768	0.9985	0.1509	0.2040	0.2573
6	CDC40	h–0.5	0.8677	0.9762	0.9979	0.1548	0.2034	0.2521
6	CYB5R4	h–0.5	0.9572	0.9951	0.9997	0.1886	0.2369	0.2827
6	FNDC1	h–0.5	0.9051	0.9779	0.9969	0.1658	0.2051	0.2448
6	IHPK3	h–0.5	0.8933	0.9791	0.9975	0.1624	0.2062	0.2490
6	IL20RA	h–0.5	0.8823	0.9753	0.9972	0.1591	0.2024	0.2471
6	PEX6	h–0.5	0.9103	0.9771	0.9958	0.1682	0.2044	0.2388
6	PGBD1	h–0.5	0.8928	0.9843	0.9990	0.1622	0.2132	0.2650

Continued on next page

TABLE S5 – continued from previous page

Chrom.	Gene	Class.	Quant. (lb)	Quant. (median)	Quant. (ub)	ϕ_{ST} (lb)	ϕ_{ST} (median)	ϕ_{ST} (ub)
6	POPDC3	h–0.5	0.9417	0.9948	0.9998	0.1799	0.2359	0.2923
6	SRPK1	h–0.5	0.8672	0.9891	0.9997	0.1547	0.2210	0.2831
6	TAGAP	h–0.5	0.9185	0.9893	0.9993	0.1706	0.2214	0.2709
6	UBD	h–0.5	0.9104	0.9962	0.9999	0.1672	0.2418	0.3070
6	VIP	h–0.5	0.9282	0.9951	0.9999	0.1743	0.2367	0.2935
6	ZNF435	h–0.5	0.9373	0.9855	0.9978	0.1788	0.2147	0.2508
6	C6orf125	h–0.95	0.9783	0.9984	0.9999	0.2060	0.2567	0.3038
6	C6orf167	h–0.95	0.9725	0.9968	0.9998	0.2003	0.2447	0.2880
6	C6orf206	h–0.95	0.9992	1.0000	1.0000	0.2669	0.3475	0.4119
6	GSTA5	h–0.95	0.9781	0.9986	1.0000	0.2057	0.2587	0.3096
6	KIAA0408	h–0.95	0.9932	0.9998	1.0000	0.2305	0.2864	0.3374
6	PAK1IP1	h–0.95	0.9994	1.0000	1.0000	0.2722	0.3395	0.3991
6	SNRPC	h–0.95	0.9881	0.9994	1.0000	0.2194	0.2732	0.3227
6	VNN3	h–0.95	0.9711	0.9976	0.9999	0.1988	0.2501	0.2996
7	IGF2BP3	l–0.5	0.0001	0.0043	0.0569	-0.1411	-0.0728	-0.0046
7	PIP	l–0.5	0.0012	0.0186	0.1446	-0.0982	-0.0374	0.0290
7	C7orf34	h–0.5	0.8282	0.9763	0.9989	0.1589	0.2243	0.2905
7	CREB3L2	h–0.5	0.9083	0.9757	0.9961	0.1840	0.2240	0.2661
7	EPHB6	h–0.5	0.9386	0.9912	0.9994	0.1967	0.2488	0.3030
7	FAM3C	h–0.5	0.9403	0.9912	0.9993	0.1978	0.2488	0.2993
7	MDH2	h–0.5	0.9449	0.9929	0.9995	0.2003	0.2537	0.3039
7	PBEF1	h–0.5	0.9085	0.9791	0.9969	0.1837	0.2276	0.2701
7	PSCD3	h–0.5	0.8635	0.9754	0.9976	0.1680	0.2233	0.2764
7	NYD-SP18	h–0.5	0.8941	0.9879	0.9994	0.1778	0.2413	0.3028
7	RPA3	h–0.5	0.9227	0.9784	0.9956	0.1896	0.2269	0.2628
7	SP4	h–0.5	0.9258	0.9846	0.9982	0.1911	0.2354	0.2805
7	TRIM24	h–0.5	0.9181	0.9852	0.9985	0.1871	0.2364	0.2843
7	CYP3A43	h–0.95	0.9678	0.9963	0.9998	0.2160	0.2674	0.3187
7	CYP3A4	h–0.95	0.9896	0.9990	0.9999	0.2453	0.2928	0.3384

Continued on next page

TABLE S5 – continued from previous page

Chrom.	Gene	Class.	Quant. (lb)	Quant. (median)	Quant. (ub)	ϕ_{ST} (lb)	ϕ_{ST} (median)	ϕ_{ST} (ub)
7	CYP3A5	h–0.95	0.9997	1.0000	1.0000	0.3147	0.3589	0.4011
7	DDX56	h–0.95	0.9896	0.9996	1.0000	0.2449	0.3094	0.3660
7	FLJ10324	h–0.95	0.9736	0.9969	0.9998	0.2214	0.2714	0.3188
7	FLJ12571	h–0.95	0.9949	0.9997	1.0000	0.2615	0.3126	0.3597
7	ING3	h–0.95	0.9839	0.9980	0.9999	0.2349	0.2804	0.3245
7	KCND2	h–0.95	0.9762	0.9966	0.9998	0.2249	0.2692	0.3172
7	MYH16	h–0.95	0.9936	0.9997	1.0000	0.2565	0.3131	0.3663
7	SMURF1	h–0.95	0.9961	0.9997	1.0000	0.2674	0.3123	0.3575
7	SVH	h–0.95	0.9999	1.0000	1.0000	0.3252	0.4036	0.4681
7	TRIAD3	h–0.95	0.9990	1.0000	1.0000	0.2925	0.3765	0.4434
7	TRIM4	h–0.95	0.9893	0.9995	1.0000	0.2441	0.3060	0.3639
7	ZFP95	h–0.95	0.9999	1.0000	1.0000	0.3291	0.3757	0.4177
8	EXTL3	l–0.5	0.0017	0.0111	0.0564	-0.0708	-0.0363	0.0030
8	ZHX1	l–0.5	0.0000	0.0037	0.1125	-0.1293	-0.0582	0.0243
8	ADCK5	h–0.5	0.8071	0.9781	0.9991	0.1410	0.2044	0.2643
8	ASH2L	h–0.5	0.8827	0.9764	0.9972	0.1596	0.2027	0.2447
8	C8orf72	h–0.5	0.8684	0.9757	0.9982	0.1553	0.2021	0.2531
8	CHRNA6	h–0.5	0.8063	0.9830	0.9995	0.1409	0.2101	0.2734
8	CNOT7	h–0.5	0.8986	0.9773	0.9968	0.1640	0.2036	0.2418
8	FLJ23356	h–0.5	0.8522	0.9829	0.9991	0.1508	0.2100	0.2644
8	KIAA0196	h–0.5	0.8861	0.9781	0.9975	0.1599	0.2044	0.2474
8	KIAA1967	h–0.5	0.9152	0.9847	0.9986	0.1693	0.2124	0.2558
8	LRP12	h–0.5	0.8505	0.9780	0.9986	0.1508	0.2044	0.2569
8	POTE8	h–0.5	0.8039	0.9887	0.9998	0.1404	0.2188	0.2901
8	SAMD12	h–0.5	0.9250	0.9859	0.9984	0.1738	0.2142	0.2533
8	SDCBP	h–0.5	0.8948	0.9832	0.9986	0.1628	0.2105	0.2566
8	SLA	h–0.5	0.9488	0.9922	0.9994	0.1847	0.2264	0.2685
8	SLC30A8	h–0.5	0.8587	0.9754	0.9979	0.1530	0.2019	0.2495
8	TEX15	h–0.5	0.9448	0.9947	0.9998	0.1815	0.2340	0.2851

Continued on next page

TABLE S5 – continued from previous page

Chrom.	Gene	Class.	Quant. (lb)	Quant. (median)	Quant. (ub)	ϕ_{ST} (lb)	ϕ_{ST} (median)	ϕ_{ST} (ub)
8	TSPYL5	h–0.5	0.9590	0.9968	0.9999	0.1892	0.2428	0.2921
8	UBXD6	h–0.5	0.9613	0.9926	0.9991	0.1916	0.2274	0.2617
8	ZNF34	h–0.5	0.9154	0.9893	0.9994	0.1692	0.2201	0.2701
8	C8orf77	h–0.95	0.9849	0.9988	1.0000	0.2131	0.2595	0.3041
8	CRISPLD1	h–0.95	0.9872	0.9989	1.0000	0.2174	0.2610	0.3049
8	IKBKB	h–0.95	0.9750	0.9982	1.0000	0.2016	0.2532	0.3034
8	TCEA1	h–0.95	0.9806	0.9987	1.0000	0.2077	0.2587	0.3090
8	TMEM64	h–0.95	0.9744	0.9986	1.0000	0.2008	0.2575	0.3133
8	ZNF395	h–0.95	0.9868	0.9995	1.0000	0.2152	0.2740	0.3292
9	C9orf152	l–0.5	0.0009	0.0089	0.0591	-0.0869	-0.0446	0.0029
9	PMPCA	l–0.5	0.0009	0.0204	0.1845	-0.0880	-0.0255	0.0422
9	C5	h–0.5	0.9506	0.9921	0.9993	0.1909	0.2349	0.2771
9	CDC14B	h–0.5	0.8811	0.9847	0.9992	0.1639	0.2203	0.2762
9	CDKN2A	h–0.5	0.9539	0.9925	0.9994	0.1933	0.2355	0.2798
9	CTSL2	h–0.5	0.9390	0.9887	0.9987	0.1855	0.2271	0.2674
9	DAB2IP	h–0.5	0.9607	0.9987	1.0000	0.1964	0.2680	0.3335
9	DBC1	h–0.5	0.9401	0.9892	0.9989	0.1860	0.2279	0.2691
9	GARNL3	h–0.5	0.8711	0.9861	0.9995	0.1609	0.2225	0.2838
9	IARS	h–0.5	0.9418	0.9869	0.9980	0.1871	0.2237	0.2596
9	KIAA1539	h–0.5	0.9371	0.9869	0.9985	0.1844	0.2235	0.2645
9	LAMC3	h–0.5	0.9604	0.9935	0.9993	0.1979	0.2381	0.2778
9	NDUFA8	h–0.5	0.9580	0.9927	0.9992	0.1959	0.2358	0.2757
9	NPR2	h–0.5	0.9598	0.9960	0.9998	0.1968	0.2479	0.2943
9	NR6A1	h–0.5	0.8937	0.9890	0.9995	0.1670	0.2275	0.2846
9	C9orf114	h–0.95	0.9843	0.9979	0.9998	0.2203	0.2594	0.2969
9	C9orf89	h–0.95	0.9873	0.9987	0.9999	0.2247	0.2679	0.3117
9	FLJ16636	h–0.95	0.9999	1.0000	1.0000	0.3092	0.4013	0.4717
9	PRG-3	h–0.95	0.9976	1.0000	1.0000	0.2567	0.3238	0.3828
9	RABGAP1	h–0.95	0.9970	1.0000	1.0000	0.2525	0.3159	0.3742

Continued on next page

TABLE S5 – continued from previous page

Chrom.	Gene	Class.	Quant. (lb)	Quant. (median)	Quant. (ub)	ϕ_{ST} (lb)	ϕ_{ST} (median)	ϕ_{ST} (ub)
9	STRBP	h–0.95	0.9809	0.9984	0.9999	0.2151	0.2649	0.3138
9	TBC1D13	h–0.95	0.9953	0.9995	1.0000	0.2463	0.2850	0.3239
9	TRPM6	h–0.95	0.9984	1.0000	1.0000	0.2647	0.3209	0.3727
9	ZBTB26	h–0.95	0.9937	0.9998	1.0000	0.2394	0.2991	0.3526
10	IFIT2	l–0.5	0.0002	0.0085	0.1270	-0.1137	-0.0483	0.0254
10	ANKRD2	h–0.5	0.8747	0.9838	0.9991	0.1592	0.2166	0.2709
10	CDNF	h–0.5	0.9584	0.9935	0.9994	0.1939	0.2360	0.2780
10	BMPR1A	h–0.5	0.8985	0.9796	0.9975	0.1673	0.2110	0.2530
10	C10orf70	h–0.5	0.9145	0.9863	0.9989	0.1722	0.2202	0.2691
10	C10orf99	h–0.5	0.8952	0.9784	0.9974	0.1658	0.2098	0.2530
10	CHUK	h–0.5	0.8947	0.9844	0.9990	0.1653	0.2173	0.2697
10	GDF2	h–0.5	0.8063	0.9813	0.9993	0.1428	0.2133	0.2757
10	HECTD2	h–0.5	0.9484	0.9957	0.9999	0.1872	0.2441	0.2992
10	ITGB1	h–0.5	0.8937	0.9778	0.9976	0.1650	0.2092	0.2540
10	KIAA1754	h–0.5	0.8782	0.9786	0.9982	0.1605	0.2098	0.2602
10	PPP3CB	h–0.5	0.9238	0.9923	0.9997	0.1754	0.2327	0.2890
10	REEP3	h–0.5	0.9600	0.9956	0.9998	0.1944	0.2437	0.2926
10	SEPHS1	h–0.5	0.8384	0.9810	0.9992	0.1499	0.2126	0.2729
10	TTC18	h–0.5	0.8960	0.9774	0.9972	0.1660	0.2085	0.2511
10	ECHDC3	l–0.95	0.0000	0.0009	0.0123	-0.1395	-0.0916	-0.0406
10	USP54	l–0.95	0.0000	0.0003	0.0486	-0.1953	-0.1090	-0.0052
10	ADD3	h–0.95	0.9648	0.9979	1.0000	0.1972	0.2573	0.3121
10	CTNNA3	h–0.95	0.9988	0.9999	1.0000	0.2673	0.3114	0.3530
10	ECHS1	h–0.95	0.9660	0.9975	0.9999	0.1983	0.2546	0.3083
10	IPMK	h–0.95	0.9927	0.9997	1.0000	0.2339	0.2876	0.3400
10	OIT3	h–0.95	0.9966	0.9998	1.0000	0.2490	0.2946	0.3391
10	POLR3A	h–0.95	0.9976	1.0000	1.0000	0.2552	0.3211	0.3815
11	PRDX5	l–0.5	0.0002	0.0230	0.3628	-0.1124	-0.0215	0.0759
11	BAD	h–0.5	0.8636	0.9865	0.9994	0.1606	0.2250	0.2827

Continued on next page

TABLE S5 – continued from previous page

Chrom.	Gene	Class.	Quant. (lb)	Quant. (median)	Quant. (ub)	ϕ_{ST} (lb)	ϕ_{ST} (median)	ϕ_{ST} (ub)
11	C11orf49	h–0.5	0.8937	0.9779	0.9977	0.1696	0.2135	0.2592
11	C11orf56	h–0.5	0.9421	0.9864	0.9983	0.1886	0.2250	0.2633
11	CCS	h–0.5	0.9536	0.9896	0.9983	0.1950	0.2306	0.2647
11	CKAP5	h–0.5	0.9338	0.9890	0.9990	0.1845	0.2295	0.2747
11	FADS3	h–0.5	0.9348	0.9859	0.9980	0.1854	0.2241	0.2624
11	FLJ12529	h–0.5	0.8833	0.9794	0.9981	0.1662	0.2152	0.2637
11	GTF2H1	h–0.5	0.9245	0.9887	0.9991	0.1803	0.2288	0.2756
11	MGC34821	h–0.5	0.8774	0.9786	0.9981	0.1645	0.2144	0.2633
11	NRXN2	h–0.5	0.9228	0.9851	0.9983	0.1798	0.2228	0.2661
11	SART1	h–0.5	0.8403	0.9882	0.9996	0.1548	0.2278	0.2900
11	SDHD	h–0.5	0.9500	0.9981	1.0000	0.1924	0.2637	0.3260
11	SF1	h–0.5	0.9482	0.9906	0.9990	0.1919	0.2328	0.2729
11	SF3B2	h–0.5	0.9418	0.9911	0.9993	0.1883	0.2342	0.2795
11	SIAE	h–0.5	0.9499	0.9943	0.9997	0.1929	0.2433	0.2934
11	SLC43A1	h–0.5	0.9208	0.9863	0.9988	0.1792	0.2247	0.2705
11	SYT7	h–0.5	0.9650	0.9946	0.9996	0.2019	0.2443	0.2873
11	DDB1	h–0.95	1.0000	1.0000	1.0000	0.3472	0.4089	0.4610
11	DKFZP564J0863	h–0.95	0.9754	0.9993	1.0000	0.2105	0.2809	0.3416
11	FADS1	h–0.95	0.9958	0.9995	1.0000	0.2500	0.2868	0.3201
11	FBXL11	h–0.95	0.9935	0.9998	1.0000	0.2406	0.2968	0.3522
11	FLJ20294	h–0.95	0.9689	0.9976	0.9999	0.2052	0.2592	0.3103
11	HRASLS5	h–0.95	0.9878	0.9995	1.0000	0.2271	0.2846	0.3406
11	MGC2574	h–0.95	0.9694	0.9965	0.9998	0.2061	0.2527	0.2988
12	CCDC65	l–0.5	0.0062	0.0243	0.0830	-0.0681	-0.0341	0.0044
12	RAB5B	l–0.5	0.0022	0.0206	0.1205	-0.0915	-0.0387	0.0195
12	BAZ2A	h–0.5	0.9248	0.9887	0.9992	0.1925	0.2472	0.3020
12	C12orf44	h–0.5	0.9647	0.9947	0.9995	0.2167	0.2644	0.3092
12	C12orf51	h–0.5	0.9439	0.9917	0.9993	0.2025	0.2545	0.3048
12	CUTL2	h–0.5	0.9607	0.9944	0.9996	0.2139	0.2633	0.3122

Continued on next page

TABLE S5 – continued from previous page

Chrom.	Gene	Class.	Quant. (lb)	Quant. (median)	Quant. (ub)	ϕ_{ST} (lb)	ϕ_{ST} (median)	ϕ_{ST} (ub)
12	DPY19L2	h–0.5	0.8293	0.9834	0.9994	0.1603	0.2373	0.3063
12	EP400NL	h–0.5	0.9365	0.9867	0.9984	0.1984	0.2430	0.2880
12	EPS8	h–0.5	0.9156	0.9898	0.9993	0.1883	0.2495	0.3068
12	IFT81	h–0.5	0.9400	0.9850	0.9975	0.2006	0.2399	0.2790
12	KERA	h–0.5	0.9284	0.9831	0.9973	0.1948	0.2368	0.2782
12	KIAA0789	h–0.5	0.9388	0.9844	0.9974	0.1999	0.2389	0.2787
12	LOC196463	h–0.5	0.8520	0.9869	0.9995	0.1665	0.2433	0.3109
12	LYZ	h–0.5	0.8838	0.9856	0.9993	0.1761	0.2410	0.3033
12	MBD6	h–0.5	0.8855	0.9814	0.9986	0.1773	0.2343	0.2910
12	NAP1L1	h–0.5	0.9429	0.9892	0.9989	0.2018	0.2483	0.2944
12	NOL1	h–0.5	0.9249	0.9929	0.9997	0.1923	0.2578	0.3162
12	NR4A1	h–0.5	0.9583	0.9956	0.9997	0.2116	0.2686	0.3207
12	OACT5	h–0.5	0.9479	0.9928	0.9994	0.2048	0.2578	0.3082
12	SLC15A4	h–0.5	0.9033	0.9789	0.9973	0.1834	0.2310	0.2790
12	SLCO1C1	h–0.5	0.9353	0.9838	0.9972	0.1984	0.2378	0.2772
12	SMARCC2	h–0.5	0.8367	0.9758	0.9985	0.1627	0.2273	0.2900
12	UBE2N	h–0.5	0.9420	0.9906	0.9992	0.2014	0.2513	0.3010
12	ULK1	h–0.5	0.9188	0.9853	0.9984	0.1901	0.2405	0.2883
12	ACTR6	h–0.95	0.9832	0.9981	0.9999	0.2378	0.2858	0.3319
12	ANAPC5	h–0.95	0.9906	0.9994	1.0000	0.2514	0.3080	0.3631
12	AQP6	h–0.95	0.9872	0.9983	0.9999	0.2447	0.2882	0.3300
12	C12orf30	h–0.95	0.9764	0.9966	0.9997	0.2285	0.2743	0.3197
12	CART1	h–0.95	0.9881	0.9986	0.9999	0.2464	0.2916	0.3334
12	CEP290	h–0.95	0.9750	0.9969	0.9998	0.2269	0.2763	0.3260
12	CMAS	h–0.95	0.9786	0.9976	0.9999	0.2304	0.2813	0.3307
12	EIF4B	h–0.95	0.9993	1.0000	1.0000	0.3036	0.3599	0.4132
12	GABARAPL1	h–0.95	0.9846	0.9983	0.9999	0.2399	0.2879	0.3336
12	MPHOSPH9	h–0.95	0.9687	0.9962	0.9997	0.2200	0.2720	0.3211
12	PAWR	h–0.95	0.9962	0.9999	1.0000	0.2718	0.3351	0.3930

Continued on next page

TABLE S5 – continued from previous page

Chrom.	Gene	Class.	Quant. (lb)	Quant. (median)	Quant. (ub)	ϕ_{ST} (lb)	ϕ_{ST} (median)	ϕ_{ST} (ub)
12	SPATS2	h–0.95	0.9999	1.0000	1.0000	0.3465	0.3944	0.4380
12	STRAP	h–0.95	0.9966	0.9999	1.0000	0.2746	0.3368	0.3933
12	WDR66	h–0.95	0.9826	0.9985	0.9999	0.2362	0.2901	0.3448
13	WBP4	l–0.5	0.0008	0.0147	0.1256	-0.1006	-0.0426	0.0212
13	RB1	h–0.5	0.9494	0.9963	0.9999	0.1894	0.2510	0.3053
13	C13orf12	h–0.95	0.9902	0.9991	1.0000	0.2324	0.2771	0.3200
13	FAM48A	h–0.95	0.9972	1.0000	1.0000	0.2557	0.3291	0.3980
13	FBXL3	h–0.95	0.9992	1.0000	1.0000	0.2792	0.3340	0.3856
13	GTF2F2	h–0.95	0.9942	0.9996	1.0000	0.2444	0.2911	0.3377
13	KLF5	h–0.95	0.9993	1.0000	1.0000	0.2817	0.3346	0.3838
13	NDFIP2	h–0.95	0.9984	1.0000	1.0000	0.2685	0.3216	0.3695
13	TGDS	h–0.95	0.9835	0.9983	0.9999	0.2202	0.2661	0.3100
14	C14orf165	l–0.5	0.0001	0.0066	0.1247	-0.1503	-0.0683	0.0226
14	FOS	l–0.5	0.0012	0.0187	0.1340	-0.1041	-0.0414	0.0246
14	C14orf174	h–0.5	0.9462	0.9882	0.9984	0.2094	0.2514	0.2926
14	DLK1	h–0.5	0.9429	0.9921	0.9994	0.2068	0.2611	0.3118
14	DLST	h–0.5	0.9431	0.9890	0.9986	0.2075	0.2531	0.2954
14	GPHN	h–0.5	0.9125	0.9801	0.9974	0.1924	0.2376	0.2848
14	KIAA0423	h–0.5	0.8789	0.9807	0.9985	0.1792	0.2385	0.2965
14	PSEN1	h–0.5	0.8692	0.9765	0.9980	0.1762	0.2332	0.2901
14	RAB2B	h–0.5	0.9103	0.9801	0.9974	0.1917	0.2376	0.2840
14	SEC10L1	h–0.5	0.9293	0.9856	0.9983	0.1999	0.2461	0.2923
14	BRF1	h–0.95	0.9960	0.9996	1.0000	0.2777	0.3227	0.3675
14	EFS	h–0.95	0.9841	0.9975	0.9997	0.2458	0.2863	0.3260
14	IL17E	h–0.95	0.9890	0.9985	0.9999	0.2549	0.2970	0.3367
14	JPH4	h–0.95	0.9972	0.9999	1.0000	0.2839	0.3437	0.3976
14	MTA1	h–0.95	0.9992	1.0000	1.0000	0.3105	0.3607	0.4061
14	RBM23	h–0.95	0.9993	0.9999	1.0000	0.3142	0.3511	0.3843
14	SIX4	h–0.95	0.9705	0.9989	1.0000	0.2264	0.3029	0.3691

Continued on next page

TABLE S5 – continued from previous page

Chrom.	Gene	Class.	Quant. (lb)	Quant. (median)	Quant. (ub)	ϕ_{ST} (lb)	ϕ_{ST} (median)	ϕ_{ST} (ub)
14	TDP1	h–0.95	0.9675	0.9959	0.9998	0.2245	0.2759	0.3272
15	DYX1C1	h–0.5	0.9427	0.9866	0.9978	0.2258	0.2705	0.3146
15	FURIN	h–0.5	0.9098	0.9795	0.9971	0.2090	0.2584	0.3074
15	HERC2	h–0.5	0.9329	0.9915	0.9994	0.2197	0.2828	0.3389
15	ARPP-19	h–0.5	0.9645	0.9928	0.9990	0.2425	0.2869	0.3301
15	PDIA3	h–0.5	0.9641	0.9923	0.9989	0.2422	0.2854	0.3267
15	PEX11A	h–0.5	0.9113	0.9877	0.9990	0.2085	0.2730	0.3310
15	SLC28A2	h–0.5	0.9498	0.9913	0.9991	0.2300	0.2826	0.3326
15	SNRPN	h–0.5	0.9585	0.9948	0.9997	0.2368	0.2954	0.3536
15	TMEM87A	h–0.5	0.9278	0.9893	0.9991	0.2169	0.2767	0.3334
15	ZNF690	h–0.5	0.9388	0.9757	0.9923	0.2251	0.2533	0.2824
15	ARIH1	h–0.95	0.9877	0.9989	0.9999	0.2728	0.3290	0.3828
15	CDAN1	h–0.95	0.9959	0.9996	1.0000	0.3016	0.3500	0.3959
15	DUOX2	h–0.95	1.0000	1.0000	1.0000	0.3899	0.4480	0.4977
15	FRMD5	h–0.95	0.9892	0.9988	0.9999	0.2784	0.3275	0.3731
15	NIP	h–0.95	0.9729	0.9973	0.9999	0.2507	0.3107	0.3657
15	RBPMS2	h–0.95	0.9911	0.9989	0.9999	0.2832	0.3304	0.3761
15	SLC24A5	h–0.95	0.9757	0.9959	0.9996	0.2545	0.3008	0.3466
15	SPATA5L1	h–0.95	0.9700	0.9995	1.0000	0.2462	0.3471	0.4295
15	TP53BP1	h–0.95	0.9939	0.9992	0.9999	0.2933	0.3369	0.3800
15	TRIP4	h–0.95	0.9866	0.9987	0.9999	0.2710	0.3257	0.3788
16	HAS3	l–0.5	0.0011	0.0091	0.0560	-0.1321	-0.0778	-0.0154
16	ADAT1	h–0.5	0.9233	0.9847	0.9983	0.2271	0.2832	0.3398
16	ARHGAP17	h–0.5	0.8884	0.9775	0.9973	0.2099	0.2710	0.3289
16	C16orf34	h–0.5	0.8907	0.9787	0.9975	0.2115	0.2726	0.3301
16	CCL22	h–0.5	0.8749	0.9819	0.9985	0.2047	0.2782	0.3424
16	FLJ20581	h–0.5	0.9630	0.9931	0.9992	0.2543	0.3059	0.3564
16	LONPL	h–0.5	0.9167	0.9857	0.9984	0.2237	0.2851	0.3405
16	LRRC36	h–0.5	0.9125	0.9859	0.9987	0.2214	0.2854	0.3454

Continued on next page

TABLE S5 – continued from previous page

Chrom.	Gene	Class.	Quant. (lb)	Quant. (median)	Quant. (ub)	ϕ_{ST} (lb)	ϕ_{ST} (median)	ϕ_{ST} (ub)
16	MGC33367	h–0.5	0.9190	0.9803	0.9968	0.2247	0.2753	0.3238
16	MPFL	h–0.5	0.9211	0.9761	0.9947	0.2263	0.2692	0.3106
16	SPIN1	h–0.5	0.9085	0.9892	0.9994	0.2188	0.2934	0.3623
16	TRAF7	h–0.5	0.8984	0.9771	0.9970	0.2147	0.2706	0.3267
16	ZNF23	h–0.5	0.9333	0.9962	0.9999	0.2322	0.3217	0.3969
16	ABCC12	h–0.95	0.9821	0.9981	0.9999	0.2796	0.3377	0.3919
16	CCL17	h–0.95	0.9877	0.9987	0.9999	0.2904	0.3469	0.3991
16	CORO7	h–0.95	0.9854	0.9986	0.9999	0.2849	0.3454	0.4005
16	CREBBP	h–0.95	0.9904	0.9995	1.0000	0.2973	0.3667	0.4290
16	GABARAPL2	h–0.95	0.9817	0.9983	0.9999	0.2788	0.3404	0.3962
16	GLIS2	h–0.95	0.9931	0.9992	0.9999	0.3074	0.3586	0.4062
16	PLEKHG4	h–0.95	0.9993	1.0000	1.0000	0.3621	0.4125	0.4598
16	POLR3E	h–0.95	0.9665	0.9977	0.9999	0.2579	0.3339	0.4015
16	SLC9A5	h–0.95	0.9974	0.9999	1.0000	0.3318	0.3985	0.4603
17	TUSC5	l–0.5	0.0047	0.0207	0.0685	-0.0761	-0.0376	0.0015
17	ADORA2B	h–0.5	0.9079	0.9800	0.9978	0.2013	0.2512	0.3036
17	CASC3	h–0.5	0.9093	0.9876	0.9991	0.2019	0.2643	0.3230
17	CD300LF	h–0.5	0.9182	0.9817	0.9975	0.2062	0.2535	0.3021
17	CDR2L	h–0.5	0.9264	0.9908	0.9994	0.2094	0.2717	0.3298
17	DDX5	h–0.5	0.8575	0.9949	1.0000	0.1829	0.2859	0.3739
17	ERN1	h–0.5	0.9634	0.9917	0.9988	0.2343	0.2743	0.3163
17	ET	h–0.5	0.9080	0.9885	0.9993	0.2016	0.2662	0.3271
17	GRB2	h–0.5	0.9599	0.9926	0.9992	0.2309	0.2771	0.3224
17	KIAA1618	h–0.5	0.9183	0.9776	0.9960	0.2066	0.2480	0.2901
17	NDEL1	h–0.5	0.8524	0.9774	0.9984	0.1820	0.2477	0.3106
17	NLGN2	h–0.5	0.9421	0.9899	0.9989	0.2189	0.2693	0.3183
17	RSAD1	h–0.5	0.8919	0.9862	0.9990	0.1947	0.2614	0.3193
17	SCARF1	h–0.5	0.9176	0.9853	0.9986	0.2056	0.2597	0.3129
17	TTC25	h–0.5	0.8478	0.9839	0.9992	0.1797	0.2571	0.3250

Continued on next page

TABLE S5 – continued from previous page

Chrom.	Gene	Class.	Quant. (lb)	Quant. (median)	Quant. (ub)	ϕ_{ST} (lb)	ϕ_{ST} (median)	ϕ_{ST} (ub)
17	ZNF232	h–0.5	0.9295	0.9856	0.9983	0.2115	0.2604	0.3080
17	AATF	h–0.95	0.9977	0.9998	1.0000	0.3051	0.3484	0.3962
17	C17orf64	h–0.95	0.9999	1.0000	1.0000	0.3710	0.4266	0.4791
17	CENTA2	h–0.95	0.9763	0.9973	0.9999	0.2473	0.3002	0.3529
17	CSNK1D	h–0.95	0.9919	0.9995	1.0000	0.2754	0.3355	0.3891
17	DERL2	h–0.95	0.9941	0.9995	1.0000	0.2832	0.3350	0.3831
17	G6PC	h–0.95	0.9675	0.9953	0.9996	0.2367	0.2881	0.3382
17	GNA13	h–0.95	0.9927	0.9995	1.0000	0.2782	0.3348	0.3885
17	ITGAE	h–0.95	0.9968	0.9999	1.0000	0.2965	0.3706	0.4349
17	PRPF8	h–0.95	0.9816	0.9978	0.9999	0.2539	0.3047	0.3557
17	TTC19	h–0.95	0.9934	0.9996	1.0000	0.2813	0.3379	0.3916
17	USP32	h–0.95	0.9955	0.9999	1.0000	0.2885	0.3654	0.4355
17	WDR68	h–0.95	0.9936	0.9996	1.0000	0.2810	0.3392	0.3897
18	ANKRD30B	l–0.5	0.0043	0.0219	0.0782	-0.0497	-0.0194	0.0130
18	C18orf10	l–0.5	0.0001	0.0031	0.0419	-0.1105	-0.0594	-0.0030
18	SERPINB12	l–0.5	0.0008	0.0138	0.1152	-0.0809	-0.0295	0.0253
18	GATA6	h–0.5	0.9591	0.9941	0.9996	0.1882	0.2291	0.2702
18	GNAL	h–0.5	0.8948	0.9777	0.9970	0.1617	0.2019	0.2391
18	KIAA1328	h–0.5	0.9065	0.9869	0.9992	0.1653	0.2128	0.2610
18	ME2	h–0.5	0.8918	0.9802	0.9983	0.1593	0.2045	0.2502
18	KIAA1772	h–0.95	0.9898	0.9997	1.0000	0.2174	0.2765	0.3330
18	PSMA8	h–0.95	0.9796	0.9980	0.9999	0.2043	0.2484	0.2909
18	RTTN	h–0.95	0.9898	0.9995	1.0000	0.2183	0.2688	0.3188
18	TXNDC10	h–0.95	0.9982	1.0000	1.0000	0.2497	0.3032	0.3559
19	FLJ23447	l–0.5	0.0005	0.0130	0.1581	-0.1151	-0.0449	0.0347
19	MYBPC2	l–0.5	0.0073	0.0239	0.0674	-0.0575	-0.0288	0.0017
19	SFRS16	l–0.5	0.0050	0.0195	0.0612	-0.0651	-0.0343	-0.0013
19	ZNF599	l–0.5	0.0018	0.0110	0.0486	-0.0873	-0.0490	-0.0082
19	BPY2IP1	h–0.5	0.9530	0.9955	0.9998	0.2065	0.2650	0.3216

Continued on next page

TABLE S5 – continued from previous page

Chrom.	Gene	Class.	Quant. (lb)	Quant. (median)	Quant. (ub)	ϕ_{ST} (lb)	ϕ_{ST} (median)	ϕ_{ST} (ub)
19	C19orf25	h–0.5	0.8690	0.9769	0.9980	0.1713	0.2264	0.2813
19	CBLC	h–0.5	0.9506	0.9935	0.9995	0.2049	0.2568	0.3073
19	DDX49	h–0.5	0.8839	0.9773	0.9973	0.1760	0.2269	0.2747
19	EEF2	h–0.5	0.8921	0.9756	0.9969	0.1795	0.2249	0.2717
19	EMR1	h–0.5	0.9506	0.9897	0.9986	0.2059	0.2463	0.2866
19	EPN1	h–0.5	0.9660	0.9949	0.9996	0.2165	0.2622	0.3069
19	FCAR	h–0.5	0.8873	0.9835	0.9987	0.1772	0.2351	0.2895
19	GTF2F1	h–0.5	0.9585	0.9937	0.9994	0.2102	0.2578	0.3025
19	LIM2	h–0.5	0.9221	0.9916	0.9996	0.1906	0.2511	0.3098
19	LOC115098	h–0.5	0.9420	0.9908	0.9992	0.2004	0.2493	0.2969
19	LOC126208	h–0.5	0.9280	0.9912	0.9994	0.1930	0.2501	0.3027
19	MAP4K1	h–0.5	0.9053	0.9799	0.9973	0.1840	0.2300	0.2741
19	MGC11271	h–0.5	0.9626	0.9926	0.9992	0.2137	0.2541	0.2957
19	PGPEP1	h–0.5	0.9147	0.9859	0.9988	0.1873	0.2390	0.2894
19	PIAS4	h–0.5	0.9618	0.9946	0.9996	0.2131	0.2611	0.3074
19	PPP1R15A	h–0.5	0.9009	0.9758	0.9963	0.1824	0.2253	0.2681
19	PRKCSH	h–0.5	0.8978	0.9827	0.9984	0.1808	0.2339	0.2852
19	RASGRP4	h–0.5	0.9242	0.9823	0.9972	0.1918	0.2332	0.2741
19	TNPO2	h–0.5	0.9010	0.9833	0.9986	0.1820	0.2346	0.2871
19	TYROBP	h–0.5	0.6729	0.9752	0.9994	0.1284	0.2245	0.3048
19	ZNF444	h–0.5	0.9335	0.9857	0.9980	0.1960	0.2388	0.2799
19	ZNF560	h–0.5	0.9269	0.9905	0.9994	0.1925	0.2485	0.3039
19	ZNF653	h–0.5	0.9179	0.9932	0.9998	0.1884	0.2560	0.3192
19	ARHGEF1	h–0.95	0.9832	0.9990	1.0000	0.2348	0.2938	0.3495
19	CDC34	h–0.95	0.9784	0.9961	0.9995	0.2290	0.2677	0.3053
19	CNN1	h–0.95	0.9996	1.0000	1.0000	0.3092	0.3651	0.4170
19	FKBP8	h–0.95	0.9974	1.0000	1.0000	0.2755	0.3567	0.4256
19	LOC112703	h–0.95	0.9694	0.9975	0.9999	0.2187	0.2771	0.3310
19	PSMC4	h–0.95	0.9700	0.9961	0.9997	0.2196	0.2678	0.3142

Continued on next page

TABLE S5 – continued from previous page

Chrom.	Gene	Class.	Quant. (lb)	Quant. (median)	Quant. (ub)	ϕ_{ST} (lb)	ϕ_{ST} (median)	ϕ_{ST} (ub)
19	SAMD4B	h–0.95	0.9939	0.9998	1.0000	0.2580	0.3197	0.3779
19	ZNF440L	h–0.95	0.9952	0.9998	1.0000	0.2639	0.3204	0.3741
19	ZNF93	h–0.95	0.9660	0.9985	1.0000	0.2155	0.2869	0.3506
19	ZNRF4	h–0.95	0.9748	0.9983	0.9999	0.2243	0.2847	0.3394
20	BCAS4	h–0.5	0.9511	0.9871	0.9982	0.2194	0.2588	0.3027
20	CST9	h–0.5	0.9204	0.9826	0.9976	0.2023	0.2507	0.2971
20	DEFB123	h–0.5	0.9249	0.9947	0.9999	0.2038	0.2810	0.3526
20	FLJ33706	h–0.5	0.9489	0.9961	0.9999	0.2175	0.2880	0.3500
20	PDYN	h–0.5	0.9562	0.9901	0.9985	0.2238	0.2653	0.3060
20	RNPC1	h–0.5	0.9103	0.9812	0.9974	0.1979	0.2481	0.2950
20	STK35	h–0.5	0.9651	0.9938	0.9993	0.2313	0.2767	0.3207
20	AHCY	h–0.95	0.9995	1.0000	1.0000	0.3291	0.4079	0.4767
20	BCL2L1	h–0.95	0.9709	0.9941	0.9992	0.2373	0.2781	0.3197
20	C20orf4	h–0.95	0.9998	1.0000	1.0000	0.3469	0.4151	0.4748
20	CBLN4	h–0.95	0.9986	0.9999	1.0000	0.3127	0.3579	0.3995
20	HM13	h–0.95	0.9999	1.0000	1.0000	0.3653	0.4175	0.4648
20	PSMA7	h–0.95	0.9946	0.9994	1.0000	0.2818	0.3277	0.3707
20	SRMS	h–0.95	0.9688	0.9952	0.9996	0.2337	0.2834	0.3308
20	TPX2	h–0.95	0.9697	0.9956	0.9997	0.2356	0.2848	0.3349
21	BRWD1	l–0.5	0.0016	0.0211	0.1346	-0.0542	-0.0123	0.0316
21	CHODL	h–0.5	0.7998	0.9759	0.9990	0.1226	0.1753	0.2261
21	NDUFV3	h–0.5	0.9511	0.9941	0.9997	0.1615	0.2003	0.2384
21	USP16	h–0.5	0.8572	0.9769	0.9984	0.1334	0.1761	0.2187
21	C21orf33	h–0.95	0.9733	0.9976	0.9999	0.1743	0.2142	0.2534
21	DSCR8	h–0.95	0.9715	0.9981	1.0000	0.1715	0.2174	0.2606
21	MRAP	h–0.95	0.9902	0.9994	1.0000	0.1914	0.2332	0.2736
22	PDGFB	l–0.5	0.0049	0.0232	0.0820	-0.0823	-0.0425	-0.0000
22	ADRBK2	h–0.5	0.9624	0.9939	0.9994	0.2303	0.2791	0.3269
22	CARD10	h–0.5	0.9206	0.9815	0.9969	0.2042	0.2501	0.2928

Continued on next page

TABLE S5 – continued from previous page

Chrom.	Gene	Class.	Quant. (lb)	Quant. (median)	Quant. (ub)	ϕ_{ST} (lb)	ϕ_{ST} (median)	ϕ_{ST} (ub)
22	FAM83F	h–0.5	0.8972	0.9915	0.9997	0.1924	0.2711	0.3406
22	LIF	h–0.5	0.7590	0.9832	0.9995	0.1523	0.2531	0.3337
22	ZNF278	h–0.5	0.9363	0.9899	0.9992	0.2118	0.2670	0.3202
22	FLJ10945	h–0.95	0.9759	0.9956	0.9995	0.2442	0.2871	0.3283
22	FLJ20699	h–0.95	0.9802	0.9994	1.0000	0.2476	0.3275	0.3993
22	FLJ27365	h–0.95	0.9739	0.9937	0.9990	0.2419	0.2787	0.3157
22	MMP11	h–0.95	0.9848	0.9986	0.9999	0.2565	0.3119	0.3625
22	PKDREJ	h–0.95	1.0000	1.0000	1.0000	0.3964	0.4547	0.5044
22	ZBED4	h–0.95	0.9974	0.9998	1.0000	0.3005	0.3498	0.3996
X	CRSP2	h–0.5	0.9387	0.9837	0.9971	0.2888	0.3411	0.3935
X	EDA2R	h–0.5	0.9130	0.9907	0.9996	0.2697	0.3607	0.4439
X	FLJ20298	h–0.5	0.8873	0.9848	0.9990	0.2563	0.3435	0.4233
X	HEPH	h–0.5	0.9445	0.9910	0.9991	0.2930	0.3618	0.4247
X	IGBP1	h–0.5	0.9262	0.9779	0.9951	0.2806	0.3297	0.3778
X	ITIH5L	h–0.5	0.9379	0.9912	0.9993	0.2863	0.3625	0.4313
X	LOC158957	h–0.5	0.9322	0.9817	0.9967	0.2833	0.3369	0.3886
X	NRK	h–0.5	0.8801	0.9806	0.9981	0.2524	0.3348	0.4065
X	PRRG1	h–0.5	0.9661	0.9961	0.9998	0.3144	0.3877	0.4564
X	AR	h–0.95	0.9983	0.9998	1.0000	0.4152	0.4645	0.5098
X	GNL3L	h–0.95	0.9762	0.9980	0.9999	0.3276	0.4068	0.4786
X	GPKOW	h–0.95	0.9983	1.0000	1.0000	0.4117	0.4910	0.5563
X	MSN	h–0.95	0.9834	0.9985	0.9999	0.3410	0.4138	0.4804
X	PHKA1	h–0.95	0.9713	0.9954	0.9995	0.3207	0.3828	0.4396
X	RPL10	h–0.95	0.9726	0.9957	0.9996	0.3227	0.3850	0.4414
X	TSC22D3	h–0.95	0.9724	0.9968	0.9998	0.3217	0.3930	0.4590