A hierarchical Bayesian model for next-generation population genomics

Zachariah Gompert and C. Alex Buerkle

Department of Botany and Program in Ecology, University of Wyoming, Laramie, Wyoming 82071
Running title: Bayesian population genomics

Keywords: Population genomics, selection, outlier analysis, Bayesian modeling, next-generation sequencing

Corresponding author: Zachariah Gompert
Department of Botany, 3165
1000 E. University Ave.
University of Wyoming
Laramie, WY 82071
Phone: (307) 760-3858
Fax: (307) 766-2851
Email: zgompert@uwyo.edu
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., $\phi_{ST}$ 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 based on 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 or reduced haplotype diversity (Andolfatto et al. 1999; Sabeti et al. 2002; Voight et al. 2006). Among-population tests for selection require multi-locus data sets and identify non-neutral 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 $F_{ST}$ based on 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 $F_{ST}$ for individual loci represent independent draws from a common, underlying distribution that characterizes the genome and that can be estimated directly from multi-locus 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 2008). 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 based on 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 \( \phi \) 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 \( \phi \) 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 \( \phi \) statistics for each genetic region are drawn from a common, genome-level distribution, and identifies outlier loci, or loci potentially affected by selection, based on the probability of their locus-specific \( \phi \) statis-
tic given the genome-level probability distribution of $\phi$. This genome-level distribution is equivalent to the conditional probability for the locus-specific $\phi$ statistics in the model.

We begin this paper 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 $\phi$ statistics across the genome, whereas locus-specific $\phi$ 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 $\phi$ 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 ($x$) given the population haplotype frequencies ($p$). 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):

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

where $n_{ij}$ is the total number of observed sequences at locus $i$ for population $j$, $x_{ijk}$ and $p_{ijk}$ are the observed count and population frequency of the $k^{th}$ haplotype in the $j^{th}$ population.
at the $i^{th}$ 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 ($\text{NGS-individual model}$) for the observed haplotype counts given the population haplotype frequencies:

$$P(x|p) = \prod_{i} \prod_{j} \prod_{l} \left\{ \begin{array}{ll} 2p_{ijk_{a}}p_{ijk_{b}} & \left[ \frac{n_{ijkl}}{x_{ijkl}} \right]^{\frac{1}{2}} \left[ 0.5^{x_{ijkl}}(0.5^{x_{ijkl}}) \right] \quad \text{if } h \text{ is two} \\ p_{ijk_{a}}^{2} + \sum_{k \neq k_{a}} \left[ 2p_{ijk_{a}}p_{ijk_{k}} \right] \left[ 0.5^{x_{ijkl}} \right] & \quad \text{if } h \text{ is one} \end{array} \right. \right.$$ (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, $x_{ijkl}$ and $x_{ijkl}$ 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, have 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 only be associated
with sequences at the population level. We propose a third likelihood model \((\text{NGS-population model})\) for this situation, in which the probability of the observed haplotype frequencies given the population frequencies is described by a multivariate Pólya distribution (Gompert et al. 2010):

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

where \(\Gamma\) is the gamma function, \(\nu_j\) 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 \(\mathbf{p}\) assuming the distance matrix \(\mathbf{d}\) 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 \(\mathbf{p}\):

\[
P(\mathbf{p}|\alpha_{ST}, \beta_{ST}, \mathbf{d}) = \prod_i \frac{1}{B(\alpha_{ST}, \beta_{ST})} \frac{(\phi_{STi} + 1)^{\alpha_{ST} - 1}(1 - \phi_{STi})^{\beta_{ST} - 1}}{2^{\alpha_{ST} + \beta_{ST} - 1}},
\]

where \(B\) is the beta function and \(\phi_{STi}\) denotes \(\phi_{ST}\) for locus \(i\) calculated based on \(p_i\) and \(d_i\) following Excoffier et al. (1992). This specification of the conditional prior for \(\mathbf{p}\) does
not correspond to a standard probability distribution with respect to \( p \), but is equivalent to assuming that the locus-specific \( \phi_{ST} \) are distributed Beta\((\alpha = \alpha_{ST}, \beta = \beta_{ST}, a = -1, b = 1)\) with \( d \) fixed, where \( \alpha_{ST} \) and \( \beta_{ST} \) 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 \( \phi_{ST} \) across the genome, with a mean and standard deviation given by:

\[
\mu_{ST} = \frac{2\alpha_{ST}}{\alpha_{ST} + \beta_{ST}} - 1
\]  

and

\[
\sigma_{ST} = 2\sqrt{\frac{\alpha_{ST}\beta_{ST}}{(\alpha_{ST} + \beta_{ST})^2(\alpha_{ST} + \beta_{ST} + 1)}},
\]

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

\[
P(p_{}, \alpha_{ST}, \beta_{ST}|x, d, \nu) \propto P(x|p_{}, \nu)P(p_{}|\alpha_{ST}, \beta_{ST}, d)P(\alpha_{ST})P(\beta_{ST}).
\]  

We provide an alternative conditional prior for the haplotype frequencies when genetic structure among groups and populations is of interest. Specifically we assume:
where $\phi_{SC}$ and $\phi_{CT}$ denote $\phi_{SC}$ (molecular variation among populations within groups of populations) and $\phi_{CT}$ (molecular variation among groups of populations relative to total haplotypic diversity) for locus $i$ calculated based on $p_i$ and $d_i$ following EXCOFFIER et al. (1992), and other parameters are as described above. This specification estimates genome-level distributions for each $\phi$ statistic independently. Because locus-specific $\phi_{ST}$ are fully specified given locus-specific $\phi_{SC}$ and $\phi_{CT}$, an alternative and equally valid modeling approach would be to treat the mean genome-level $\phi_{ST}$ as a derived parameter and specify a conditional prior based only on $\phi_{SC}$ and $\phi_{CT}$. This alternative approach would account for the lack of independence among $\phi_{SC}$, $\phi_{CT}$, and $\phi_{ST}$ and would be closer to existing decompositions of $\phi$ statistics. However, this alternative model prior would not provide posterior estimates of $\alpha_{ST}$ or $\beta_{ST}$, which are necessary to specify the posterior distribution for the genome-level distribution of $\phi_{ST}$, as opposed to the posterior distribution of the mean genome-level $\phi_{ST}$. The former is necessary to identify outlier loci with respect to $\phi_{ST}$, 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 $\phi$ 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 $\alpha$ and $\beta$ parameters. Finally, as before we assign uninformative, uniform priors to all $\alpha$ and $\beta$ parameters. This specification results in the following Bayesian model:
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 \( a^{th} \) 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 \( \frac{n}{2} \) th and \( 1 - \frac{n}{2} \) th quantiles of the genome-level φ statistic distribution. We then consider locus \( i \) an outlier at the \( a^{th} \) 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 paper 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 only be designated an outlier with respect to \( \phi_{ST} \), whereas a genetic region could be an outlier with respect to \( \phi_{ST}, \phi_{SC}, \) or \( \phi_{CT} \) when group and population structure are considered. Outlier loci can be classified as having very low \( \phi_{ST} \), which could be indicative of balancing or purifying selection, or very high \( \phi_{ST} \), 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 $\phi_{CT}$ and $\phi_{SC}$ might be a bit more complex. For example, high $\phi_{CT}$ outliers would be expected if divergent selection occurred among groups of populations with the same alleles favored within each group and low $\phi_{SC}$ 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 this approach in several important ways. Guo et al. (2009) contrast an approximation of the posterior probability distribution for each locus-specific $\phi_{ST}$ ($\theta_i$’s in their model) to the hyperdistribution describing among locus variation in $\phi_{ST}$ (equivalent to our genome-level distribution). Approximation of the posterior probability distribution for $\phi_{ST}$ 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 $\phi_{ST}$ 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 $\phi_{ST}$ (this distribution measures uncertainty in the parameter estimate) and a point estimate of the genome-level $\phi_{ST}$ distribution (this distribution measures expected variation among loci in $\phi_{ST}$). In contrast, we test whether locus-specific $\phi$ statistics are unlikely given the genome-level distribution (i.e., not that our estimates of these locus-specific $\phi$ 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 $\phi$ 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 biologist and their method utilizing more information from the posterior distribution of locus-specific $\phi_{ST}$.

**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 $\tau$ generations in the past, where $\tau$ 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 $\mu$ 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 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. We generated 10 replicate simulated data sets for each combination of likelihood model, number of genetic regions...
(25 or 500), \( \tau \), and migration rate. Each data set was analyzed using an Markov chain Monte Carlo (MCMC) implementation of the proposed model using the \textit{bamova} software we have developed (see Supplemental Material \textit{MCMC algorithm}; available from the authors at \url{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 based on the posterior probability distribution for the quantile of each locus-specific \( \phi_{ST} \) in the genome-level \( \phi_{ST} \) distribution, as previously described (see \textit{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 1, 3, or 5 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 non-neutral 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 \textit{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 based on $\phi_{ST}$ (the partitioning of molecular variation among populations), $\phi_{CT}$ (the partitioning of molecular variation between the two groups), $\phi_{SC}$ (the partitioning of molecular variation among the populations within each group). We determined the mean proportion of neutral and non-neutral loci classified as outliers based on each $\phi$ statistic across the 10 replicates for each combination of parameters.

\textbf{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 (\texttt{http://pga.gs.washington.edu}; downloaded May 2010). Each gene was sequenced in 24 individuals with African ancestry (either the AA panel or the HapMap YRI population) and 23 individuals with European Ancestry (either the CEPH population or the HapMap CEU population). The phase of polymorphisms in these sequences has been estimated statistically using PHASE v2.0 (\textit{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 will be typical for current NGS studies. We used these data and the \textit{bamova} software to identify loci with exceptional
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 Supplemental Material 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 burnin followed by 50,000 iterations to estimate 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” \( \phi_{ST} \) 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, CA). 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 more that 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 lo-
cus. Finally, if there were more than 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 12649 regions and 11866 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
based on levels of molecular differentiation among these human populations using the bamova
software. This allowed us to contrast pattern 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 based 50,000 MCMC iterations following a 25,000 iter-
ation burnin. 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
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 $\phi_{ST}$ 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 $\phi_{ST}$ outliers) to 0.034 (NGS–population model, no migration, 500 sequence loci, low $\phi_{ST}$ 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 $\phi_{ST}$ 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 non-neutral 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 $\phi_{ST}$, $a = 0.5$, 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 $\phi_{ST}$, $a = 0.5$, NGS-individual model mean = 0.612, sd = 0.310; Tables S1, S2, 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 (Tables S1, S2, 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 pop-
ulations) was highly dependent on the 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 $\phi_{SC}$; Table S4). The proportion of swept genetic regions correctly classified as $\phi_{SC}$ outliers was generally small and did not exceed 0.150, but was much greater than the proportion of neutral genetic regions classified as $\phi_{SC}$ outliers. In contrast, when both groups of populations were affected by selective sweeps, nearly all swept loci were classified as high $\phi_{ST}$ and $\phi_{SC}$ outliers (92.5–100% using $a = 0.5$) and a smaller proportion (0.100–0.225) were classified as low or high $\phi_{CT}$ outliers. Thus, the selective sweeps were most readily detected based on high differentiation among all populations ($\phi_{ST}$) or among populations within each group ($\phi_{SC}$), but could also be detected based on low or high differentiation between the two groups ($\phi_{CT}$). 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 (Tables S2, S4).

**Divergent selection among African and European populations** Mean genome-level $\phi_{ST}$ 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 approximately 8% of DNA sequence variation was partitioned between the African and European ancestry populations.

We classified three genes as high $\phi_{ST}$ outliers (using $a = 0.5$) based on the first five SNPs data subset (Fig. 1). 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 \( \phi_{ST} \) equal to 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; 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 European Americans (Harris 2001). The third outlier locus was POLG2, with an estimated \( \phi_{ST} \) of 0.327 (95\% ETPI 0.175–0.477, Fig. 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 \( \phi_{ST} \) 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, Fig. 3). The estimate of \( \phi_{ST} \) for the X chromosome was considerably higher (0.139, 95\% ETPI 0.128–0.149).

We detected remarkable variation in \( \phi_{ST} \) along each of the chromosomes (Figure 4). We detected 569 unique \( \phi_{ST} \) outlier loci, which we designate as candidate genes for selection among worldwide human populations. Of these 569 genes, 518 were high \( \phi_{ST} \) outlier loci (including 222 with \( a = 0.95 \)) and 51 were low \( \phi_{ST} \) outlier loci (including six 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, Fig. 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 $\phi_{ST}$ (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 $\phi_{ST}$ 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 non-synonymous 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 $\phi_{ST}$ 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 $\phi_{ST}$ for SPATA5L1 (spermatogenesis associated 5-like 1) on chromosome 15 was 0.347 (95% ETPI 0.246–0.430). This gene was classified as a high $\phi_{ST}$ 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 $\phi_{ST}$ 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 low $\phi_{ST}$ 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). 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 non-human (including non-model) species (Gompert et al. 2010; Hohenlohe et al. 2010). However, beyond data acquisition, substantial biological insights will only be possible 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; Gompert et al. 2010; Futschik and Schlötterer 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 Supplemental Material *Simulations: estimation of φ statistics*). For example, the estimates of genome-level $\phi_{ST}$ for the SeattleSNPs human sequence data and chromosome-level $\phi_{ST}$ 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 positive rate for high or low $\phi_{ST}$ outliers was never greater than 0.04, and using $a = 0.95$ this rate was never greater than 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 decrease the total number of selected genes identified relative to $a = 0.5$). One benefit of quantifying genetic divergence based on haplotypes (using $\phi_{ST}$) 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 only detect 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 based on 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 approximately 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 affect 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 based on previous studies, such as the LCT (lactase) 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 2008; 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 2008). 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.

ACKNOWLEDGEMENTS

We would like to thank J. Fordyce, M. Forister, C. Nice, P. Nosil, T. Parchman, and R. Safran for discussion and comments on previous drafts of this manuscript. Comments from L. Excoffier and two anonymous reviewers helped to improve the manuscript. We are indebted to R. Williamson for his contributions to the development of the bamova software. This work was supported by NSF DBI award 0701757 to CAB.

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,


Tishkoff, S. A., F. A. Reed, F. R. Friedlaender, C. Ehret, A. Ran- ciaro, A. Froment, J. B. Hirbo, A. A. Awomoyi, J.-M. Bodo, O. Doumbo,


Figure 1.—Locus-specific $\phi_{ST}$ estimates for Africans and Europeans (SeattleSNPs data set). A point estimate of the genome-level $\phi_{ST}$ distribution (based on the median from the posterior probability distributions of $\alpha_{ST}$ and $\beta_{ST}$) 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.
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 $\phi_{ST}$ for the large sample of human genetic diversity in 33 populations (data from Jakobsson et al. 2008). Point estimates of each chromosome-level $\phi_{ST}$ distribution (based on the median from the posterior probability distributions of $\alpha_{ST}$ and $\beta_{ST}$) are denoted with solid black lines (autosomes) or a dashed orange line (X-chromosome; A). Posterior probability distributions for the mean chromosome-level $\phi_{ST}$ for each chromosome (B).
<table>
<thead>
<tr>
<th>Chromosomal position (Mb)</th>
<th>Chromosomal position (Mb)</th>
<th>Chromosomal position (Mb)</th>
<th>Chromosomal position (Mb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>−0.20 0.15 0.50</td>
<td>−0.20 0.15 0.50</td>
<td>−0.20 0.15 0.50</td>
<td>−0.20 0.15 0.50</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0.4</td>
<td>0.1</td>
</tr>
<tr>
<td>124.1</td>
<td>121.2</td>
<td>99.8</td>
<td>94.7</td>
</tr>
<tr>
<td>247.2</td>
<td>242.4</td>
<td>199.2</td>
<td>189.3</td>
</tr>
</tbody>
</table>

| 5                        | 6                        | 7                        | 8                        |
| −0.20 0.15 0.50          | −0.20 0.15 0.50          | −0.20 0.15 0.50          | −0.20 0.15 0.50          |
| 0                        | 121.2                    | 99.8                     | 94.7                     |
| 0.4                      | 99.8                     | 199.2                    | 189.3                    |
| 124.1                    | 242.4                    | 199.2                    | 189.3                    |

| 9                        | 10                       | 11                       | 12                       |
| −0.20 0.15 0.50          | −0.20 0.15 0.50          | −0.20 0.15 0.50          | −0.20 0.15 0.50          |
| 0.4                      | 70.3                     | 140.1                    | 250.1                    |
| 70.3                     | 140.1                    | 250.1                    | 360.1                    |
| 13                       | 14                       | 15                       | 16                       |
| −0.20 0.15 0.50          | −0.20 0.15 0.50          | −0.20 0.15 0.50          | −0.20 0.15 0.50          |
| 0                        | 18.7                     | 30.7                     | 42.8                     |
| 18.7                     | 30.7                     | 42.8                     | 55.1                     |

**Figure 4.**—Estimates of $\phi_{ST}$ 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 $\phi_{ST}$ for a chromosome. Estimates of $\phi_{ST}$ for individual genes are shown in grey for non-outlier genes and light (at $a = 0.5$) or dark blue (at $a = 0.95$) for outlier genes. For each gene the closed circle gives the median from the posterior distribution of $\phi_{ST}$ for that locus and the bars denote the 95% ETPI.
**TABLE 1**

Proportion of outlier loci in simulated neutral data

<table>
<thead>
<tr>
<th>Model</th>
<th>Model No. loci</th>
<th>( \tau )</th>
<th>( a = 0.5 )</th>
<th>( a = 0.95 )</th>
<th>( m = 0 )</th>
<th>( m = 2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>low</td>
<td>high</td>
<td>low</td>
<td>high</td>
</tr>
<tr>
<td>Known haplotypes model</td>
<td>25</td>
<td>0.25</td>
<td>0.012</td>
<td>0.028</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.50</td>
<td>0.028</td>
<td>0.028</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.00</td>
<td>0.024</td>
<td>0.004</td>
<td>0.004</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.25</td>
<td>0.0132</td>
<td>0.0212</td>
<td>0.0004</td>
<td>0.0036</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.50</td>
<td>0.0240</td>
<td>0.0164</td>
<td>0.0028</td>
<td>0.0004</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.00</td>
<td>0.0320</td>
<td>0.0032</td>
<td>0.0006</td>
<td>0.0000</td>
</tr>
<tr>
<td>NGS–individual model</td>
<td>25</td>
<td>0.25</td>
<td>0.012</td>
<td>0.012</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.50</td>
<td>0.004</td>
<td>0.024</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.00</td>
<td>0.020</td>
<td>0.024</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.25</td>
<td>0.0136</td>
<td>0.0242</td>
<td>0.0005</td>
<td>0.0042</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.50</td>
<td>0.0220</td>
<td>0.0202</td>
<td>0.0022</td>
<td>0.0042</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.00</td>
<td>0.0280</td>
<td>0.0074</td>
<td>0.0068</td>
<td>0.0006</td>
</tr>
<tr>
<td>NGS–population model</td>
<td>25</td>
<td>0.25</td>
<td>0.016</td>
<td>0.020</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.50</td>
<td>0.024</td>
<td>0.040</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.00</td>
<td>0.032</td>
<td>0.004</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.25</td>
<td>0.0112</td>
<td>0.0202</td>
<td>0.0002</td>
<td>0.0024</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.50</td>
<td>0.0190</td>
<td>0.0164</td>
<td>0.0006</td>
<td>0.0012</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.00</td>
<td>0.0340</td>
<td>0.0064</td>
<td>0.0050</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

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

Summary of data and results for each chromosome, including chromosome–level $\phi_{ST}$ (median and 95% ETPI), the number of loci, and the number of loci classified as outliers (data from Jakobsson et al. 2008).