Detecting Selection from Linked Sites Using an *F*-Model

Marco Galimberti,*,† Christoph Leuenberger,‡ Beat Wolf,§ Sándor Miklós Szilágyi, Matthieu Foll,†† and Daniel Wegmann*,†,1**

*Department of Biology and ‡Department of Mathematics, University of Fribourg, 1700, Switzerland, †Swiss Institute of Bioinformatics, Fribourg, 1700, Switzerland, §iCoSys, University of Applied Sciences Western Switzerland, Fribourg, 1700 Switzerland, **Department of Informatics, University of Medicine, Pharmacy, Science and Technology of Târgu Mureş, Târgu Mureş, 540139, Romania, and ††International Agency for Research on Cancer (IARC/WHO), Section of Genetics, 69372 Lyon, France

ORCID IDs: 0000-0001-6052-156X (M.G.); 0000-0002-5006-6484 (C.L.); 0000-0002-9307-7212 (B.W.); 0000-0002-5657-7365 (S.M.S.); 0000-0001-9006-8436 (M.F.); 0000-0003-2866-6739 (D.W.)

ABSTRACT Allele frequencies vary across populations and loci, even in the presence of migration. While most differences may be due to genetic drift, divergent selection will further increase differentiation at some loci. Identifying those is key in studying local adaptation, but remains statistically challenging. A particularly elegant way to describe allele frequency differences among populations connected by migration is the *F*-model, which measures differences in allele frequencies by population specific F_{ST} coefficients. This model readily accounts for multiple evolutionary forces by partitioning F_{ST} coefficients into locus- and population-specific components reflecting selection and drift, respectively. Here we present an extension of this model to linked loci by means of a hidden Markov model (HMM), which characterizes the effect of selection on linked markers through correlations in the locus specific component along the genome. Using extensive simulations, we show that the statistical power of our method is up to twofold higher than that of previous implementations that assume sites to be independent. We finally evidence selection in the human genome by applying our method to data from the Human Genome Diversity Project (HGDP).

KEYWORDS Bayesian statistics; F-statistics; hidden Markov model; divergent selection; balancing selection

MIGRATION is a major evolutionary force homogenizing evolutionary trajectories of populations by promoting the exchange of genetic material. At some loci, however, the influx of new genetic material may be modulated by selection. In case of strong local adaptation, for instance, migrants may carry maladapted alleles that are selected against. Identifying loci that contribute to local adaptation is of major interests in evolutionary biology because these loci are thought to constitute the first step toward ecological speciation (*e.g.*, Wu 2001; Feder *et al.* 2012) and allow us to understand the role of selection in shaping phenotypic differences between populations and species (*e.g.*, Bonin *et al.* 2006; Fournier-Level *et al.* 2011).

A simple, yet flexible and useful, approach to identify loci contributing to local adaptation is to scan the genome using statistics that quantify divergence between populations. One frequently used statistic is F_{ST}, which measures population differentiation, and loci with greatly elevated F_{ST} have been reported for many population comparisons (*e.g.*, Jones *et al.* 2012; Andrew and Rieseberg 2013; Stölting *et al.* 2013). While other statistics measuring absolute divergence (Cruickshank and Hahn 2014) or incongruence between a population tree and locus-specific genealogies (Durand *et al.* 2011; Peter 2016) may be more suited in some situations, genome scans suffer from two inherent limitations. First, multiple evolutionary scenarios may explain the deviations in those statistics, making interpretation difficult (*e.g.*, Cruickshank and Hahn 2014; Eriksson and Manica 2012). Second, the definition of outliers is arbitrary, allowing for the detection of candidate loci only. Indeed, loci also vary in

doi: https://doi.org/10.1534/genetics.120.303780
Manuscript received May 17, 2020; accepted for publication October 3, 2020; published Early Online October 16, 2020.
Supplemental material available at figshare: https://doi.org/10.25386/genetics.13077284.

[1]Corresponding author: University of Fribourg, Department of Biology, Chemin du Musée 10, Fribourg, 1700 Switzerland. E-mail: daniel.wegmann@unifr.ch

their divergence between populations that were never subjected to selection, but outlier approaches would still identify outliers.

Multiple methods have thus been developed that explicitly incorporate the stochastic effects of genetic drift. A first important step to improve the reliability of outlier scans was the proposal to compare observed values of such statistics against the distribution expected under a null model. Among the first, Beaumont and Nichols (1996) proposed to obtain the distribution of F_{ST} through simulations performed under an island model. While the idea to evidence selection by comparing F_{ST} to its expectations is far from new (*e.g.*, Lewontin and Krakauer 1973), the difficulty of properly parameterizing the null model was quickly realized (*e.g.*, Nei and Maruyama 1975). The success of the method by Beaumont and Nichols (1996) relies on tailoring the parameters of the underlying island model to match the observed heterozygosity at each locus—an approach that is also easily extended to structured island models (Excoffier *et al.* 2009).

A more formal approach is given by means of the *F*-model (Rannala and Hartigan 1996; Balding 2003; Falush *et al.* 2003; Gaggiotti and Foll 2010), under which allele frequencies are measured by locus and population specific F_{ST}^{lj} coefficients that reflect the amount of drift that occurred in population j at locus l since its divergence from a common ancestral population. In the case of biallelic loci, the current frequencies $\tilde{p}_{jl}$ are then given by a beta distribution (Beaumont and Balding 2004)

$$\tilde{p}_{jl} \sim \mathrm{Beta}(\theta_{lj}p_l, \theta_{lj}(1-p_l)), \qquad (1)$$

where p_l are the frequencies in the ancestral population and θ_{lj} is given by

$$F_{\mathrm{ST}}^{lj} = \frac{1}{1+\theta_{lj}}.$$

It is straightforward to extend this model to account for different evolutionary forces that affect the degree of genetic differentiation. For instance, Beaumont and Balding (2004) proposed to partition the effects of genetic drift and selection into locus-specific and population-specific components α_l and β_j, as well as a locus-by-population specific error term γ_{ij}:

$$\log\left(\frac{1}{\theta_{lj}}\right) = \alpha_l + \beta_j + \gamma_{ij} \qquad (2)$$

Loci with $\alpha_l \neq 0$ are interpreted to be affected by either balancing ($\alpha_l < 0$) or divergent ($\alpha_l > 0$) selection, either because they are targets of selection or through hitch-hiking (Beaumont and Balding 2004). Such loci may be identified by contrasting models with $\alpha_l = 0$ or $\alpha_l \neq 0$ for each locus l, either through Bayesian variable selection (Riebler *et al.* 2008) or via reversible-jump MCMC, as is done in the popular software BayeScan (Foll and Gaggiotti 2008).

A common problem of this, and many other, genome-scan methods is the assumption of independence among loci, which is easily violated when working with genomic data. By evaluating information from multiple linked loci jointly, however, the statistical power to detect outlier regions is likely increased considerably. Indeed, even a weak signal of divergence may become detectable if it is shared among multiple loci. Similarly, false positives may be avoided as their signal is unlikely to be shared with linked loci.

Unfortunately, fully accounting for linkage is often statistically challenging as well as computationally very costly. One solution is to split the problem by first inferring haplotypes for each sample, and then performing selection scans on the haplotype structure. The extended haplotype homozygosity (EHH) and its derived statistics (Sabeti *et al.* 2002; Voight *et al.* 2006; Sabeti *et al.* 2007; Tang *et al.* 2007), for instance, identify shared haplotypes of exceptional length. More recently, Fariello *et al.* (2013) introduced methods that identify haplotype clusters with particularly large frequency differences between populations and showed that using haplotypes rather than single markers increases power substantially.

An alternative solution is to model linkage through the autocorrelation of hierarchical parameters along the genome, which does not require knowledge of the underlying haplotype structure. Boitard *et al.* (2009) and Kern and Haussler (2010), for instance, proposed a genome-scan method in which each locus was classified as selected or neutral, and then used a Hidden Markov Model (HMM) to account for the fact that linked loci likely belonged to the same class, while ignoring autocorrelation in the genetic data itself.

Here, we build on this idea to develop a genome-scan method based on the *F*-model. While an HMM implementation of the *F*-model was previously proposed to deal with linked sites when inferring admixture proportions (Falush *et al.* 2003), we use it here to characterize autocorrelations in the strength of selection α_l among linked markers. As we show using both simulations and an application to human data, aggregating information across loci results in an increase in power of up to twofold at the same false-discovery rate (FDR).

Materials and Methods

A model for genetic differentiation and observations

We assume the classic *F*-model, in which J populations diverged from a common ancestral population. Since divergence, each population experienced genetic drift at a different rate. We quantify this drift of population $j = 1, \ldots, J$ at locus $l = 1, \ldots, L$ by θ_{jl}. We further assume each locus to be biallelic with ancestral frequencies p_l, in which case the current frequencies $\tilde{p}_{jl}$ are given by a beta distribution (Beaumont and Balding 2004), as shown in (2). We thus have

$$\mathbb{P}(\tilde{p}_{jl}|p_l, \theta_{jl}) = \frac{1}{B(\theta_{jl}p_l, \theta_{jl}q_l)}\,(\tilde{p}_{jl})^{\theta_{jl}p_l-1}(\tilde{q}_{jl})^{\theta_{jl}q_l-1}, \qquad (3)$$

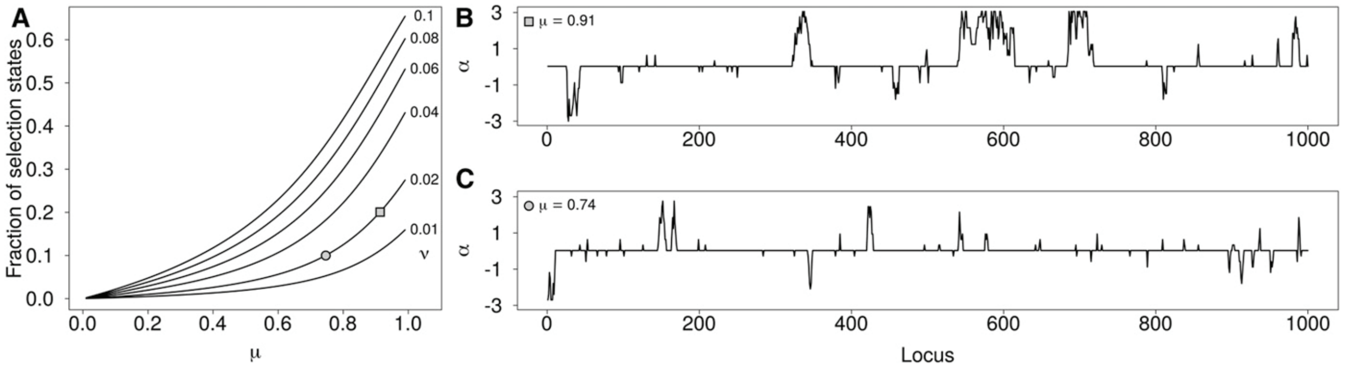

Figure 1 (A) Expected proportion of neutral sites as a function of rates μ and ν. (B, C) Example paths of α_l along 1000 loci simulated at a distance of d_l = 100 with $s_{\max}$ = 10 positive and negative states up to $\alpha_{\max}$ = 3.0. Autocorrelation among loci was simulated with log(κ) = −3.0, ν = 0.02, and μ = 0.91 (B, square) or μ = 0.74 (C, circle), respectively. The two cases correspond to an expected proportion of 20% and 10% of the genome under selection, as marked in (A).

where $q_l = 1 - p_l$, $\tilde{q}_{jl} = 1 - \tilde{p}_{jl}$, $B(x,y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$ and $\Gamma(\cdot)$ is the gamma function.

Let n_{jl} denote the allele counts in a sample of N_{jl} haplotypes from population j at locus l, which is given by a binomial distribution

$$n_{jl} \sim \mathrm{Bin}(\tilde{p}_{jl},\ N_{jl})$$

and hence

$$\mathbb{P}(n_{jl}|\tilde{p}_{jl}) = \binom{N_{jl}}{n_{jl}} (\tilde{p}_{jl})^{n_{jl}} (\tilde{q}_{jl})^{N_{jl}-n_{jl}}. \tag{4}$$

Equations (3) and (4) combine to a beta-binomial distribution

$$\mathbb{P}(n_{jl}|\theta_{jl},\ p_l) = \binom{N_{jl}}{n_{jl}} \frac{B(\theta_{jl}p_l + n_{jl},\ \theta_{jl}q_l + N_{jl} - n_{jl})}{B(\theta_{jl}p_l,\ \theta_{jl}q_l)}. \tag{5}$$

Model of selection

We decompose θ_{jl} into a population-specific component β_j shared by all loci, and a locus-specific component α_l shared by all populations:

$$-\log\theta_{lj} = \alpha_l + \beta_j$$

Here, the locus-specific component α_l quantifies an excess or dearth of differentiation, which is attributed to the effect of either divergent or balancing selection, respectively (Beaumont and Balding 2004). Note that we adopt here the formulation of Foll and Gaggiotti (2008) and omit the error term γ_{ij} of the original model of Beaumont and Balding (2004) shown in (2), as there is generally not enough information to estimate these parameters from the data (Beaumont and Balding 2004).

To account for autocorrelation among the locus-specific component, we propose to discretize $\alpha_l = \alpha\ (s_l)$, where $s_l = -s_{\max}, -s_{\max}+1, \ldots, s_{\max}$ are the states of a ladder-type Markov model with $m = 2s_{\max} + 1$ states such that

$$\alpha(s_l) = \frac{s_l}{s_{\max}}\alpha_{\max} \tag{6}$$

for some positive parameters $\alpha_{\max}$. The transition matrix of this Markov model shall be a finite-state birth-and-death process

$$\mathbf{Q}(d_l) = e^{\kappa d_l \mathbf{L}} \tag{7}$$

with elements $[Q(d_l)]_{ij}$ denoting the probabilities to go from state i at locus $l-1$ to state j at locus l given the strength of autocorrelation measured by the positive scaling parameter κ and the known distance d_l between these loci, either in physical or in recombination space. Here, $\mathbf{L}$ is the $m \times m$ generating matrix

$$\mathbf{L} = \begin{pmatrix} -1 & 1 & 0 & \ldots & 0 & 0 \\ \mu & -1-\mu & 1 & \ldots & 0 & 0 \\ 0 & \mu & -1-\mu & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & -1-\mu & \mu \\ 0 & 0 & 0 & \ldots & 1 & -1 \end{pmatrix}$$

where the middle row at position $s_{\max} + 1$ reflects neutrality, and is given by the element

$$(0 \quad \ldots \quad \nu\mu \quad -2\nu\mu \quad \nu\mu \quad \ldots \quad 0).$$

As exemplified in Figure 1, the two parameters μ and ν control the distribution of sites affected by selection (*i.e.*, having $\alpha_l \neq 0$) in the genome, with ν affecting the number of selected regions and μ their extent and selection strength, with higher values leading to more sites affected selection. It is important to note that we do not assume all sites with $\alpha_l \neq 0$ to be targets of selection. Instead, many will be linked to a target of selection and experience $\alpha_l \neq 0$ due to hitchhiking.

The stationary distribution of this Markov chain is given by

$$\Pi = c \cdot \left(1 \ \frac{1}{\mu} \ \frac{1}{\mu^2} \ \cdots \ \frac{1}{\mu^{s_{\max}-1}} \ \frac{1}{\mu^{s_{\max}}\nu} \ \frac{1}{\mu^{s_{\max}-1}} \ \cdots \ 1 \right),$$

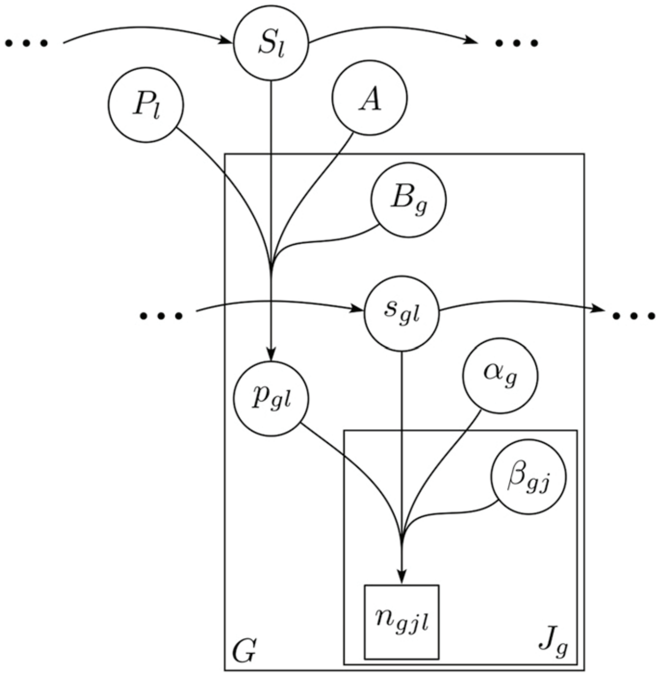

Figure 2 A directed acyclic graph (DAG) of the proposed model with two hierarchical levels.

with

$$c^{-1} = 2\frac{\mu^{s_{\max}} - 1}{\mu^{s_{\max}} - \mu^{s_{\max}-1}} + \frac{1}{\mu^{s_{\max}}\nu}.$$

Note that, as $\kappa \rightarrow \infty$, our model approaches that of Foll and Gaggiotti (2008) implemented in BayeScan, but with discretized α_l.

Hierarchical island models

Hierarchical island models, first introduced by Slatkin and Voelm (1991), address the fact that divergence might vary among groups of populations. They were previously used to infer divergent selection, both using a simulation approach (Excoffier *et al.* 2009) as well as in the case of *F*-models (Foll *et al.* 2014). Here, we describe how our model is readily extended to additional hierarchies.

Consider G groups each subdivided into J_g populations with population-specific allele frequencies $\tilde{p}_{gjl}$ that derive from group-specific frequencies p_{gl} as described above with group-specific parameters μ_g, ν_g, and κ_g. Analogously, we now assume group-specific frequencies to have diverged from a global ancestral frequency P_l according to locus-specific and group-specific parameters Θ_{gl}. Specifically,

$$p_{gl} \sim B(\Theta_{gl}P_l, \Theta_{gl}(1 - P_l))$$

such that

$$\mathbb{P}(p_{gl}|P_l, \Theta_{gl}) = \frac{1}{B(\Theta_{gl}P_l, \Theta_{gl}Q_l)}(p_{gl})^{\Theta_{gl}P_l - 1}(q_{gl})^{\Theta_{gl}Q_l - 1}, \quad (8)$$

Table 1 Parameters used in simulations

Name	J	F_{ST}	N	Log(κ)
Reference	10	0.15	50	−3
Pop-2	2	0.15	50	−3
Pop-5	5	0.15	50	−3
Pop-20	20	0.15	50	−3
Pop-50	50	0.15	50	−3
F_{ST}-0.01	10	0.01	50	−3
F_{ST}-0.05	10	0.05	50	−3
F_{ST}-0.1	10	0.1	50	−3
F_{ST}-0.25	10	0.25	50	−3
Haplo-10	10	0.15	10	−3
Haplo-20	10	0.15	20	−3
Haplo-100	10	0.15	100	−3
Haplo-200	10	0.15	200	−3
log κ-1	10	0.15	50	−1
log κ-5	10	0.15	50	−5
log κ-7	10	0.15	50	−7
log κ-9	10	0.15	50	−9

where $Q_l = 1 - P_l$ and $q_{gl} = 1 - p_{gl}$. The parameter Θ_{gl} is given by

$$-\log\Theta_{gl} = A(S_l) + B_g. \quad (9)$$

As above, B_g quantifies group specific drift, $S_l = -s_{\max}, -s_{\max} + 1, \ldots, s_{\max}$ are the states of a Markov model with m states and transition matrix $\mathbf{Q}_l = e^{\kappa d_l \mathbf{L}}$ with parameters μ and ν, a positive scaling parameter κ, and A (S_l) and $A_{\max}$ defined as in (6). Hence, we assume independent HMM models of the exact same structure at both levels of the hierarchy, as outlined in Figure 2. Additional levels could be added analogously.

Inference

We developed a Bayesian inference scheme for the parameters of the proposed model using a Markov chain Monte Carlo (MCMC) approach with Metropolis–Hastings updates, as detailed in the Supplementary Material. As priors, we used

$$\beta_j, B_g \sim \mathcal{N}(\mu_b, \sigma_b^2)$$

$$p_l \sim \text{Beta}(a_p, b_p)$$

$$\log(a_p), \log(b_p) \sim \mathcal{N}(0, 1)$$

$$\log(\kappa_g), \log(\kappa), \log(\mu), \log(\nu) \sim \mathcal{U}(-\infty, 0).$$

Following Beaumont and Balding (2004), we used $\mu_b = -2$ and $\sigma_b^2 = 1.8$ throughout. We further set $a_p = b_p = 1$.

To identify candidate regions under selection, we used MCMC samples to determine the FDRs

$$q_d(l) = 1 - \mathbb{P}(\alpha_l > 0|\boldsymbol{n}, \boldsymbol{N})$$

$$q_b(l) = 1 - \mathbb{P}(\alpha_l < 0|\boldsymbol{n}, \boldsymbol{N})$$

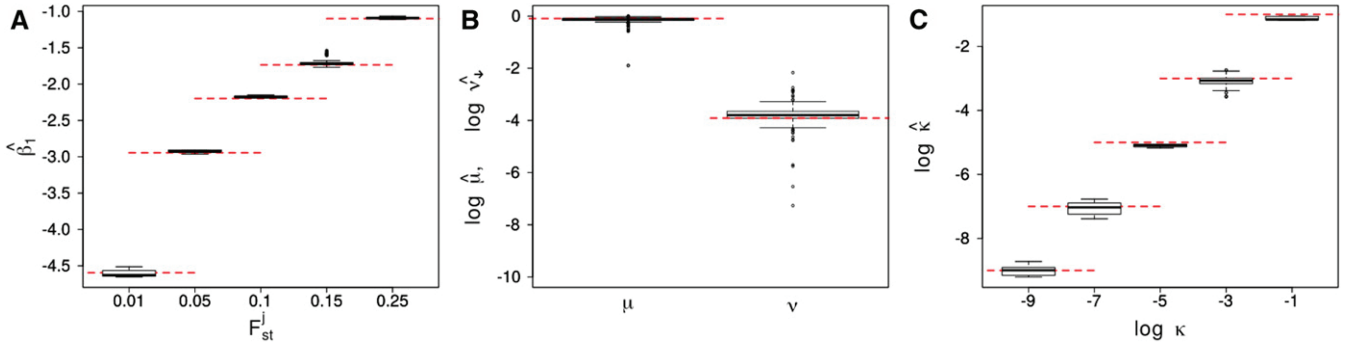

Figure 3 Boxplot of the parameters β_1 (left), ν and μ (center), and $\log(\kappa)$ (right). The values are obtained from the mean of the posterior distributions obtained using Flink on the 10 simulations run for each of the set of parameters reported in Table 1. The red dotted lines show the true values of the respective parameters.

for divergent and balancing selection, respectively, where $\boldsymbol{n} = \{n_{11}, \ldots, n_{JL}\}$ and $\boldsymbol{N} = \{N_{11}, \ldots, N_{JL}\}$ denote the full data.

Implementation

We implemented the proposed Bayesian inference scheme in the easy-to-use C++ program Flink.

Given the heavy computational burden of the proposed model, we introduce several approximations. Most importantly, we group the distances d_l into $E + 1$ ensembles such that $e_l = \log_2 d_l$, $e_l = 0, \ldots, E$, and use the same transition matrix $\mathbf{Q}(2^e)$ for all loci in ensemble e. We then calculate $\mathbf{Q}(1)$ for the first ensemble using the computationally cheap yet accurate approximation

$$\mathbf{Q}_0 = e^{\kappa d_0 \mathbf{L}} \approx \left(\mathbf{I} + \frac{1}{2^r}\kappa d_0 \mathbf{L}\right)^{2^r}$$

with $r = \log_2(D/3) + 10$, where $D = 2s_{\max} + 1$ is the dimensionality of the transition matrix (Ferrer-Admetlla *et al.* 2016). The transition matrices of all other ensembles is obtained through the recursion $\mathbf{Q}(e) = \mathbf{Q}(e-1)^2$. (See Supplementary Information for other details regarding the implementation).

Data availability

The authors affirm that all data necessary for confirming the conclusions of the article are present within the article or available from repositories as indicated. The source-code of Flink is available through the git repository https://bitbucket.org/wegmannlab/flink, along with detailed information on its usage. Additional scripts used to conduct simulations are found at https://doi.org/10.5281/zenodo.3949763. Supplemental material available at figshare: https://doi.org/10.25386/genetics.13077284.

Results

Comparison with BayeScan

Simulation parameters: To quantify the benefits of accounting for autocorrelation in the locus specific components α_l among linked loci, we first compared our method implemented in Flink against the method implemented in BayeScan (Foll and Gaggiotti 2008) on simulated data. All simulations were conducted under the model laid out above for a single group, using routines available in Flink and with parameter settings similar to those used in (Foll and Gaggiotti 2008). Specifically, we focused on a reference simulation in which we sampled $N = 50$ haplotypes from $J = 10$ populations with β_j chosen such that $F^{lj}_{\mathrm{ST}} = 0.15$ in the neutral case ($\alpha_l = 0$). Following Foll and Gaggiotti (2008), we simulated all $p_l \sim \mathrm{Beta}(0.7, 0.7)$ and about 20% of sites affected by selection (*i.e.*, with $\alpha_l \neq 0$) by setting $\mu = 0.91$ and $\nu = 0.02$. We further set $s_{\max} = 10$ (resulting in $m = 21$ states) and $\alpha_{\max_g} = 3$, and simulated 10^3 loci for each of 10 chromosomes, with a distance of 100 positions between adjacent sites and strength of autocorrelation $\log(\kappa) = -3$. We then varied the number of populations J, the sample size N, F^{lj}_{ST} or the strength of autocorrelation κ individually, while keeping all other parameters constant (Table 1). We further added a case without linkage (*i.e.*, $\kappa \rightarrow \infty$) by simulating each locus on its own chromosome.

To infer parameters with Flink, we set $s_{\max}$ and $\alpha_{\max}$ to the true values and ran the MCMC for $7{\cdot}10^5$ iterations, of which we discarded the first $20{\cdot}10^5$ as burn-in. During the chain, we recorded parameter values every 100 iterations as posterior samples. To infer parameters with Bayescan, we used version 2.1 and set the prior odds for the neutral model to 50, which we found to result in the same power as Flink in the reference simulation (see below) and in the absence of linkage ($\kappa \rightarrow \infty$). We identified loci under selection at an FDR threshold of 5% for both methods.

Power of inference: We first evaluated the power of Flink in inferring the hierarchical parameters β_j, ν, μ, and κ. As shown through the distributions of posterior means across all simulations, these estimates were very accurate and unbiased, regardless of the parameter values used in the simulations (Figure 3). This suggests that the power to identify selected loci is not limited by the number of loci we used to infer hierarchical parameters.

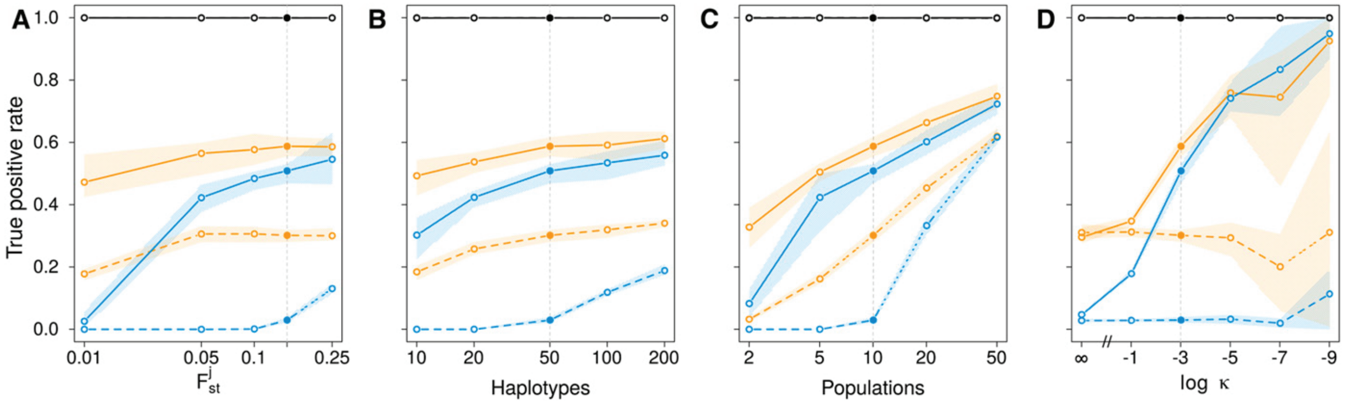

Figure 4 The true positive rate (power) in classifying loci as neutral (black) or under divergent (orange) or balancing selection (blue) as a function of the F_{ST} between populations (A), the number of haplotypes *N* (B), the number of populations *J* (C), and the strength of autocorrelation κ (D). Lines indicate the mean and range of true positive rates obtained with Flink (solid) and BayeScan (dashed) across 10 replicate simulations. Filled dots and the vertical gray line indicate the reference simulation shown in each plot.

We next studied the impact of the sample size and the strength of population differentiation on the power (the true positive rate) to identify loci affected by selection (*i.e.*, loci with $\alpha_l \neq 0$). In line with findings reported by Foll and Gaggiotti (2008), power generally increased with F^j_{ST}, the number of sampled haplotypes, and the number of sampled populations (Figure 4, A–C). Larger sample sizes or stronger differentiation was particularly relevant for detecting loci under balancing selection, for which the power was generally lower and virtually zero at low differentiation ($F^j_{\mathrm{ST}} = 0.01$) or if only few populations were sampled ($J = 2$). Importantly, the FDR was below the chosen 5% threshold in 100% and 98.6% of all simulations conducted for loci identified as affected by divergent and balancing selection, respectively (Supplemental Material, Figure S1). The false positive rates (FPR) for these classes was < 0.1% in 98.6% and 97.1% of all simulations (see Figure 4 for neutral sites).

Compared to BayeScan run on the same set of simulations, Flink had a higher power at the same FDR across all simulations, and often considerably so, unless if very many populations were sampled (Figure 4). If $J = 10$ populations were sampled, for instance, the power of Flink was about 0.2 higher for loci under divergent selection, and even up to 0.4 higher for those under balancing selection (Figure 4, A and B). Importantly, this increase in power described here is fully explained by Flink accounting for autocorrelation among the α_l values as we chose the prior odds in BayeScan to result in the same power if the strength of autocorrelation vanishes (*i.e.*, $\kappa \rightarrow \infty$). Exploiting information from linked sites to identify divergent or balancing selection can thus strongly increase power, certainly if linkage extends to many loci. This is maybe best illustrated by the much higher power of Flink to identify loci under balancing selection at low differentiation ($F^j_{\mathrm{ST}} \leq 0.1$, Figure 4A), in which case even many neutral loci are expected to show virtually no difference in allele frequencies and only an aggregation of loci with a subtle reduction in F^j_{ST} can be interpreted as a reliable signal for a target of selection in the region (Foll and Gaggiotti 2008).

Runtime: Thanks to careful optimization, there is little to no overhead of our implementation compared to that of BayeScan. On the reference simulation of 10^4 loci from 10 populations, for instance, Flink took on average 130 min on a modern computer if calculations were spread over four CPU cores. On the same data, BayeScan took 361 min. However, we note that comparing the two implementations is difficult due to many settings that strongly impact run times, such as the number of iterations or the use of pilot runs in BayeScan. Without pilot runs, the run time of BayeScan reduced to 182 min on average for the default number of iterations (10^5 including burn-in). In the same time, Flink runs for close to 10^6 iterations, but also requires more to converge.

But since computation times scale linearly with the number of loci, they remain prohibitively slow for whole genome applications in a single run. However, computations are easily spread across many computers by analyzing the genome in independent chunks such as for each chromosome or chromosome arm independently. This is justified because (1) linkage does not persist across chromosome boundaries and is usually weak across the centromere, and (2) because our simulations indicate that 10^4 polymorphic loci were sufficient to estimate the hierarchical parameters accurately.

Effect of model misspecification

The *F*-model makes the explicit assumption that the allele frequencies in a structured population can be characterized by a multinomial Dirichlet distribution. This distribution is appropriate for a wide range of demographic models, but not if some pairs of populations share a more recent ancestry than others (Beaumont and Balding 2004; Excoffier *et al.* 2009). Unsurprisingly, several previous studies found high FPRs when challenging BayeScan with models of isolation-by-distance (IBD), recent range expansions, recent admixture, or asymmetric divergence (*e.g.*, Lotterhos and Whitlock 2014; Luu *et al.* 2017). These high FPRs are partially mitigated by choosing higher prior odds (*e.g.*, 50 as used here, Lotterhos

Table 2 Population groups analyzed

		Divergent			Balancing		
Group	Populations	SNPs (%)	Regions	Length[a]	SNPs (%)	Regions	Length[a]
Africa	Bantu N.E., Biaka Pygmies, Mandenka, Mbuti Pygmies, San, Yoruba	8,020 (1.42)	759	16.8	8,026 (1.42)	433	30.2
Middle East	Mozabite, Palestinian, Druze, Bedouin	14,324 (2.54)	1137	20.6	18,432 (3.27)	848	41.2
Europe	Adygei, French, French Basque, North Italian, Orcadian, Russian, Sardinian, Tuscan	19,128 (3.39)	1466	22.0	37,736 (6.7)	1382	48.3
America	Colombians, Karitiana, Maya, Pima, Surui	33,062 (5.87)	1889	29.8	34,499 (6.12)	1735	39.4
Central Asia	Balochi, Brahui, Burusho, Hazara, Kalash, Makrani, Pathan, Sindhi	16,663 (2.96)	1290	22.6	25,473 (4.52)	1132	44.5
East Asia	Uygur, Dai, Daur, Han, Hezhen, Lahu, Miaozu, Mongola, Naxi, Oroqen, She, Tu, Tujia, Xibo, Yizu	20,528 (3.64)	1832	17.3	33,678 (5.98)	1656	35.2
Higher hierarchy	N/A	24,595 (4.36)	1692	26.8	20,156 (3.58)	1074	31.2

[a] Median length of the regions in kb.

and Whitlock 2014) or when using the hierarchical version of BayeScan (Foll *et al.* 2014), particularly in case of asymmetric divergence. In the case of a recent range expansion or recent admixture, however, the *F*-model is unlikely to be appropriate and other methods have been shown to outperform BayeScan, in particular hapFLK (Fariello *et al.* 2013) and pcadapt (Luu *et al.* 2017).

Here, we investigated how the sensitivity of the linkage-aware implementation of an *F*-model in Flink is affected by such model misspecifications. We focused on the case of a recent range expansion as this model is difficult to accommodate even with a hierarchical *F*-model. Using quantiNemo (Neuenschwander *et al.* 2018), we simulated genomic data from 11 populations with carrying capacity 10^3 each that form a one-dimensional stepping-stone model. Initially, only the left-most population contained individuals that then colonized the remaining populations through symmetric dispersal between neighboring populations at rate 0.1 and with a population growth rate of 0.1. After 10^3 generations, 20 diploid individuals were sampled from each population. We simulated 10 independent chromosomes of 10^4 neutral loci each with initial allele frequencies drawn from a Beta distribution $f_l \sim \text{Beta}(0.7, 0.7)$. We run these simulations for different recombination rates by setting the total length of the genetic map per chromosome to either 1, 10, or 100 cM. We then inferred selection on all loci still polymorphic at the end of the simulations with both BayeScan and Flink for 10 replicates per set.

Across all simulations, BayeScan identified no locus as affected as balancing selection and only 0.16% as affected by divergent selection. This low FPR is consistent with the generally low power of BayeScan to identify loci affected by balancing selection as well as the used prior odds of 50 in favor of the neutral model. Similar results were obtained with Flink on simulations with high recombination (genetic map of 100 cM), in which case no linkage information could be exploited. Across these simulations, Flink identified no locus as affected by balancing selection and only 0.14% as affected by divergent selection. The number of false positives, however, was rising sharply with decreasing recombination rate. At a genetic map of 10 cM, 5.0% and 2.8% of all loci were wrongly inferred as affected by balancing and divergent selection, respectively. At a genetic map of only 1 cM and, hence, tight linkage, the corresponding FPRs were 22.7% and 7.5%, respectively. These results thus highlight that the power gained by Flink in exploiting linkage information also translates into a higher FPR in case the model is misspecified. Under such scenarios, other methods such as hapFLK (Fariello *et al.* 2013) or PCAdapt (Luu *et al.* 2017) are thus more appropriate.

Application to humans

To illustrate the usefulness of Flink, we applied it to SNP data of 46 populations analyzed as part of the HGDP (Rosenberg *et al.* 2002, 2005) and available at https://www.hagsc.org/hgdp/files.html. We then used Plink v1.90 (Chang *et al.* 2015) to transpose the data into vcf files, and used the liftOver tool of the UCSC Genome Browser (Kent *et al.* 2002) to convert the coordinates to the human reference GRCh38.

We divided the 46 populations into six groups (Table 2) of between 4 and 15 populations each according to genetic landscapes proposed by Peter *et al.* (2017). We then inferred divergent and balancing selection using the hierarchical version of Flink on all 22 autosomes, but excluded 5 Mb on each side of the centromere and adjacent to the telomeres. The final data set consisted of 563,589 SNPs. We analyzed each chromosome arm individually with $\alpha_{\max} = 4.0$, $s_{\max} = 10$ and using an MCMC chain with $7{\cdot}10^5$ iterations, of which we discarded the first $2{\cdot}10^5$ as burn-in. Estimates of hierarchical parameters are shown in Figure S2 and the locus-specific FDRs $q_d\,(l)$ and $q_b\,(l)$ are shown for all loci, all groups as well as the higher hierarchy in Figures S4–S42. All regions

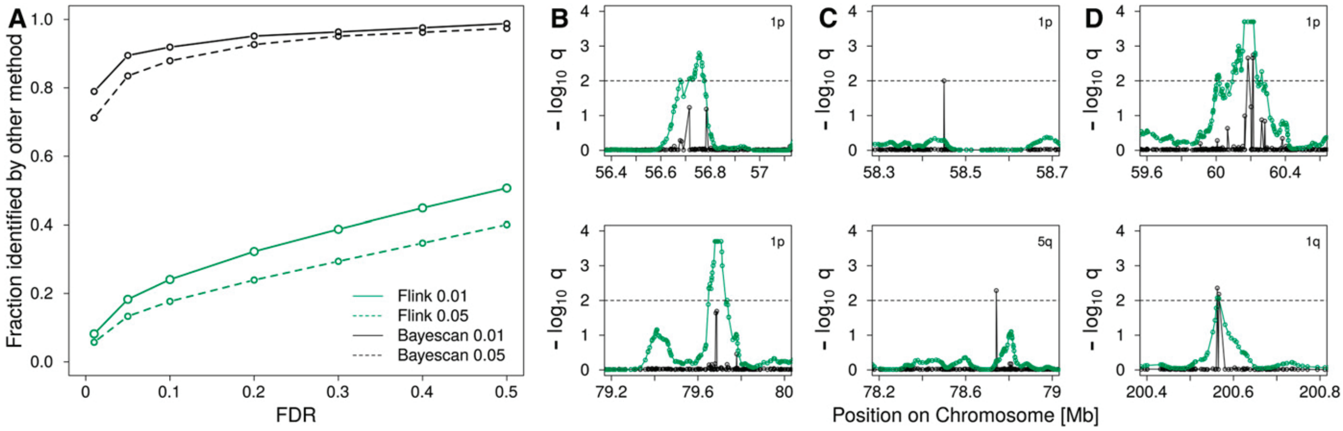

Figure 5 (A) The fraction of regions identified as divergent among Europeans by Flink (green) and BayescanH (black) at a false discovery rate (FDR) of 0.01 (solid) and 0.05 (dashed) also identified by the other method at different FDR. (B–D) Examples of regions found under divergent selection by Flink (B), BayeScanH (C), or both (D) among Europeans. Dashed lines indicate the 0.01 FDR threshold.

identified as potential targets for selection are further detailed in the Supplementary Files. As summarized in Table 2, we discovered between 759 and 1889 and between 433 and 1735 candidate regions for divergent and balancing selection, respectively, spanning together about 10% of the genome.

Comparison with BayeScan

We first validated our results by running BayeScan on the same data. We then identified divergent regions as continuous sets of SNP markers that passed an FDR threshold of 0.01 or 0.05 for each method, and determined the FDR threshold necessary to identify at least one locus within these regions by the other method. To ensure the observed differences between methods is due to accounting for linkage only, we used the hierarchical version BayeScanH (Foll *et al.* 2014) that also implements the same hierarchical island model as Flink.

As shown in Figure 5A for selected regions among Europeans, the majority of regions identified by BayeScanH were replicated by Flink at small FDR thresholds. In contrast, most of the regions identified by Flink were not replicated by BayeScanH, in line with a higher statistical power for the former. Visual inspection indeed revealed that for most regions identified by Flink but not BayeScanH, the latter also showed a signal of selection at multiple markers, each of which not passing the FDR threshold individually (see Figure 5B for examples). In contrast, sites identified by BayeScanH but not Flink usually consisted of a signal at a single site, suggesting many of those are likely false positives (Figure 5C).

Results were similar for the other groups (Figure S3), but the correspondence between the methods was higher for the African group and considerably lower for the American group, likely due to the different patterns of divergence among populations (Figure S2).

Comparison with a recent scan for selective sweeps

Positive selection acting in a subset of populations may also lead to an increase in population differentiation (Nielsen 2005). We therefore compared our outlier regions also to those of a recent scan for positive selection that combined multiple test for selection using a machine learning approach (Sugden *et al.* 2018). Among the 593 candidate loci reported for the CEU population of the 1000 Genomes Project (1000 Genomes Project Consortium *et al.* 2015) and overlapping the chromosomal segments studied here, 293 loci (49.4%) fall within a region we identified as under divergent selection either among European populations (154 loci), at the higher hierarchy (132 loci), or both (7 loci).

To test if this overlap exceeds random expectations, we generated 10,000 bootstrapped data sets by randomly sampling the same amount of loci among all those found polymorphic in the 1000 Genome Project CEU samples and within the chromosomal segments studied here. We then determined the overlap with our outlier regions for each data set. On average, 46.6 loci overlapped with our regions identified among European populations or at the higher hierarchy. Importantly, the largest overlap observed among the bootstrapped data set (72 loci) was much smaller than that observed (293 loci, $P < 10^{-4}$).

Example: the LCT region

As an illustration, we show the FDRs q_d (l) and q_b (l) for 30 Mb around the *LCT* gene in Figure 6 for the higher hierarchy as well as the European, Middle Eastern, and East Asian group. The *LCT* gene is a well studied target of positive selection that has acted to increase lactase persistence in several human populations, including Europeans (Bersaglieri *et al.* 2004; Burger *et al.* 2020). Lactase persistence varies among Europeans and decreases on a roughly north–south cline (Bersaglieri *et al.* 2004; Burger *et al.* 2007; Leonardi *et al.* 2012), consistent with the signal of divergent selection we detected among European populations (Figure 6). In line with previous findings (*e.g.*, Grossman *et al.* 2013), we detected a signal of divergent selection among Europeans also in various genes around *LCT*, most notably in *R3HDM1* but also *MIR128-1*, *UBXN4* and *DARS*. In contrast, we detected no such signal for the other groups.

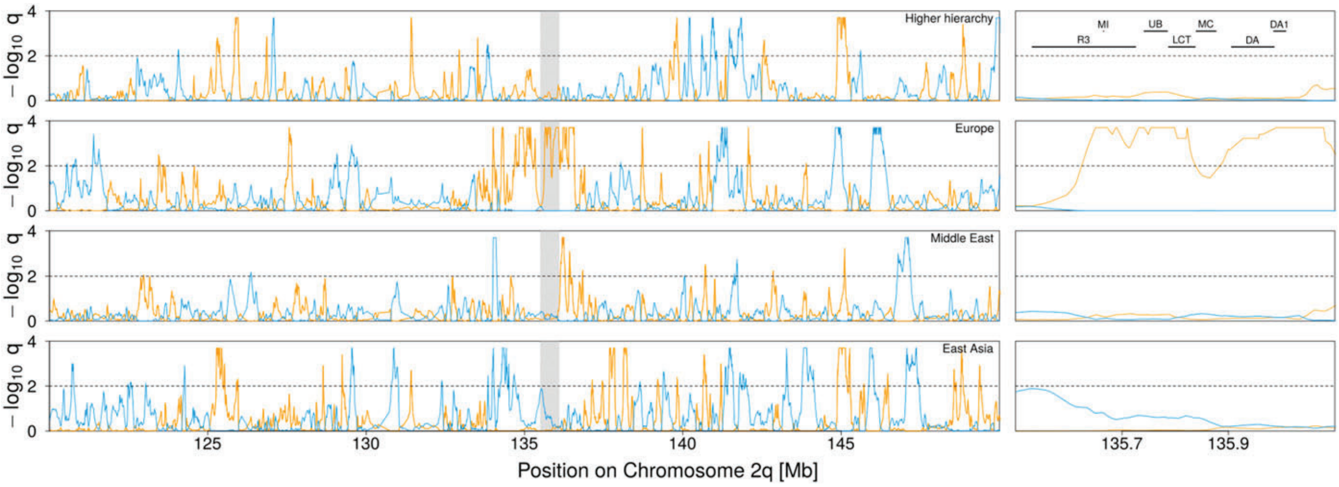

Figure 6 Signal of selection around the *LCT* gene on Chromosome 2q. The orange and blue lines indicate the locus-specific FDR for divergent (orange) and balancing (blue) selection, respectively. The black dashed line shows the 1% FDR threshold. A zoom of the highlighted region is shown on the right indicating the position of several genes: *R3HDM1* (R3), *MIR128-1* (MI), *UBXN4* (UB), *MCM6* (MC), *DARS* (DA), and *DARS-AS1* (DA1). The entire Chromosome 2q is shown in Figure S7.

Discussion

Genome scans are common methods to identify loci that contribute to local adaptation among populations. Here, we extend the particularly powerful method implemented in BayeScan (Foll and Gaggiotti 2008) to linked sites.

Accounting for linkage in population genetic methods, while desirable, is often computationally hard. We propose to alleviate this problem by modeling the dependence among linked sites through autocorrelation among hierarchical parameters, rather than the population allele frequencies or haplotypes themselves. In the context of genome scans, this has been previously used successfully to classify each locus as selected or neutral using HMMs (Boitard *et al.* 2009; Kern and Haussler 2010). Here, we extend this idea by modeling autocorrelation among the degree of spatial differentiation acting at individual loci. While ignoring autocorrelation at the genetic level certainly leads to a loss of information, the resulting method remains computationally tractable. And as we showed here with simulations and an application to human data, the resulting method features much improved statistical power compared to BayeScan—a similar method that ignores linkage completely.

This is particularly evident for loci with more similar allele frequencies among populations than expected by the genome-wide divergence. These loci are generally interpreted as being under balancing selection (Beaumont and Balding 2004; Foll and Gaggiotti 2008), but may also be the result of purifying selection restricting alleles from reaching high allele frequencies. Given the large number of loci we inferred in this class from the HGDP data (about 5% of the genome), we speculate that balancing selection is unlikely the main driver, and caution against overinterpreting these results. But we note that the empirical FDR for loci under balancing selection was extremely low in our simulations, except if the assumptions underlying the *F*-model was violated.

A benefit of accounting for autocorrelation among locus-specific effects was previously postulated by Guo *et al.* (2009), who proposed a conditional autoregressive (CAR) prior on α_l such that

$$\alpha_l|\boldsymbol{\alpha}_{-l} \sim \mathcal{N}\left(\frac{1}{\overline{w_i}}\sum_{m\neq l} w_{lm}\alpha_m,\ \frac{1}{\tau\overline{w_i}}\right),$$

where $\boldsymbol{\alpha}_{-l}$ denotes the collection of all other α_m, $m \neq l$, w_{lm} indicates the covariance between loci l and m, which is assumed to decrease exponentially with distance, and $\overline{w_i} = \sum_{m\neq l} w_{ml}$. While Guo *et al.* (2009) did not evaluate the benefit of their CAR implementation on the power of selection inference, they found that it was a better fit to high resolution data. Here, we show that the power increase by exploiting autocorrelation among loci is substantial: of all regions identified as under divergent selection by Flink, less than half were also identified by BayeScan, despite evidence that these consist mostly of true outliers.

In this context, it is important to note that due to computational challenges, Guo *et al.* (2009) suggested to run their method on low-resolution data with few markers first, and then to apply the CAR version on inferred candidate regions only. As our analysis suggests, such an approach would likely fail to harvest the full benefit of accounting for autocorrelation among locus-specific parameters. Running Flink on high-resolution data are possible because the first-order Markov assumption on locus-specific effects α_l allows for cheap MCMC updates at a single locus that does not require a recalculation of the prior on the full vector $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_L\}$. Unfortunately, however, no implementation of the method by Guo *et al.* (2009) is available for a direct comparison.

Our proposed model has yet another computational advantage: while the hierarchical parameters of the exponential decay in the model by Guo *et al.* (2009) need to be fixed upfront due to numerical instabilities, the hierarchical parameters of the discrete Markov model proposed here are all estimated well if sufficient sites are provided. Our simulations indicated that 10^4 polymorphic loci were sufficient, based on which we decided to parallelize the analysis of the human data by chromosome arm. Smaller windows may be considered, but the model may struggle to differentiate between population-specific and locus-specific components if too few consecutive loci are used. The window analyzed should therefore span significantly more loci than are expected to be affected by selection within an outlier region. But note that the model does not make any assumption regarding the spacing of loci within the analyzed window, nor does it assume that all individuals have data: it accounts for both the distances between loci as well as the locus-specific sample size explicitly. Hence, Flink may well be used on data obtained with reduced representation techniques such as RAD-seq, albeit with little benefit over BayeScan if loci are in weak linkage only.

Another major difference between Flink and the CAR method of Guo *et al.* (2009) is that the former discretizes the locus-specific effects α_l. While such a discretization leads to a loss of precision in estimating locus specific effects, it allows to directly calculate a FDR to identify outlier loci at any desired level of confidence, similar to BayeScan or the method of Riebler *et al.* (2008). In contrast, the method by Guo *et al.* (2009) identifies outliers indirectly as those for which the posterior distributions on θ_l are significantly different from the distribution of θ_l values under the inferred hyper-parameters. Importantly, the discretization seems to come at no cost on power: in our simulations, Flink and BayeScan had virtually identical power if we simulated unlinked data.

An obvious drawback of modeling the locus-specific selection coefficients as a discrete Markov Chain is that, for most candidate regions we detected, multiple loci showed a strong signal of selection, making it difficult to identify the causal variant. However, once a region is identified, estimates of F_{ST} can be obtained for each locus individually to identify the locus with the strongest signal. Complementary methods such as SWIF(r) (Sugden *et al.* 2018) may further be used on the identified regions to infer locus-specific selection coefficients or other statistics informative about the targets of selection.

Finally, we note that the implementation provided through Flink allows to group populations hierarchically. Accounting for multiple hierarchies was previously shown to reduce the number of false positives in F_{ST} based genome scans (Excoffier *et al.* 2009) and also applied in an *F*-model setting (Foll *et al.* 2014). Aside from accounting for structure more accurately, a hierarchical implementation also allows for genome-wide association studies (GWAS) with population samples. In such a setting, each sampling location would constitute a "group" of, say, two "populations", one for each phenotype (*e.g.*, cases and controls). The parameters at the higher hierarchy will then accurately describe population structure and loci associated with the phenotype will be identified as those highly divergent between the two "populations". A natural assumption would then be that the locus-specific coefficients α_l are shared among all groups, *i.e.*, that they are governed by a single HMM. While we have not made use of such a setting here, we note that it is readily available as an option in Flink.

Acknowledgments

We are grateful for the very constructive feedback of two anonymous reviewers and the editor Nick Barton on an earlier version of this manuscript. This study was supported by two Swiss National Foundation grants to DW with numbers 31003A_149920 and 31003A_173062.

Literature Cited

1000 Genomes Project Consortium; A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison *et al.*, 2015 A global reference for human genetic variation. Nature 526: 68–74. https://doi.org/10.1038/nature15393

Andrew, R. L., and L. H. Rieseberg, 2013 Divergence is focused on few genomic regions early in speciation: incipient speciation of sunflower ecotypes. Evolution 67: 2468–2482. https://doi.org/10.1111/evo.12106

Balding, D. J., 2003 Likelihood-based inference for genetic correlation coefficients. Theor. Popul. Biol. 63: 221–230. https://doi.org/10.1016/S0040-5809(03)00007-8

Beaumont, M., and R. A. Nichols, 1996 Evaluating loci for use in the genetic analysis of population structure. Proc. Biol. Sci. 263: 1619–1626. https://doi.org/10.1098/rspb.1996.0237

Beaumont, M. A., and D. J. Balding, 2004 Identifying adaptive genetic divergence among populations from genome scans. Mol. Ecol. 13: 969–980. https://doi.org/10.1111/j.1365-294X.2004.02125.x

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. https://doi.org/10.1086/421051

Boitard, S., C. Schlötterer, and A. Futschik, 2009 Detecting selective sweeps: a new approach based on hidden Markov models. Genetics 181: 1567–1578. https://doi.org/10.1534/genetics.108.100032

Bonin, A., P. Taberlet, C. Miaud, and F. Pompanon, 2006 Explorative genome scan to detect candidate loci for adaptation along a gradient of altitude in the common frog (Rana temporaria). Mol. Biol. Evol. 23: 773–783. https://doi.org/10.1093/molbev/msj087

Burger, J., M. Kirchner, B. Bramanti, W. Haak, and M. G. Thomas, 2007 Absence of the lactase-persistence-associated allele in early Neolithic Europeans. Proc. Natl. Acad. Sci. USA 104: 3736–3741. https://doi.org/10.1073/pnas.0607187104

Burger, J., V. Link, J. Blöcher, A. Schulz, C. Sell *et al.*, 2020 Low prevalence of lactase persistence in bronze age Europe indicates ongoing strong selection over the last 3,000 years. Curr. Biol. https://doi.org/10.1016/j.cub.2020.08.033

Chang, C. C., C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell *et al.*, 2015 Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4: 7. https://doi.org/10.1186/s13742-015-0047-8

Cruickshank, T. E., and M. W. Hahn, 2014 Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Mol. Ecol. 23: 3133–3157. https://doi.org/10.1111/mec.12796

Durand, E. Y., N. Patterson, D. Reich, and M. Slatkin, 2011 Testing for ancient admixture between closely related populations. Mol. Biol. Evol. 28: 2239–2252. https://doi.org/10.1093/molbev/msr048

Eriksson, A., and A. Manica, 2012 Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. Proc. Natl. Acad. Sci. USA 109: 13956–13960. https://doi.org/10.1073/pnas.1200567109

Excoffier, L., T. Hofer, and M. Foll, 2009 Detecting loci under selection in a hierarchically structured population. Heredity 103: 285–298. https://doi.org/10.1038/hdy.2009.74

Falush, D., M. Stephens, and J. K. Pritchard, 2003 Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164: 1567–1587.

Fariello, M. I., S. Boitard, H. Naya, M. SanCristobal, and B. Servin, 2013 Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. Genetics 193: 929–941. https://doi.org/10.1534/genetics.112.147231

Feder, J. L., S. P. Egan, and P. Nosil, 2012 The genomics of speciation-with-gene-flow. Trends Genet. 28: 342–350. https://doi.org/10.1016/j.tig.2012.03.009

Ferrer-Admetlla, A., C. Leuenberger, J. D. Jensen, and D. Wegmann, 2016 An approximate Markov model for the Wright-Fisher diffusion and its application to time series data. Genetics 203: 831–846. https://doi.org/10.1534/genetics.115.184598

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. https://doi.org/10.1534/genetics.108.092221

Foll, M., O. E. Gaggiotti, J. T. Daub, A. Vatsiou, and L. Excoffier, 2014 Widespread signals of convergent adaptation to high altitude in Asia and America. Am. J. Hum. Genet. 95: 394–407. https://doi.org/10.1016/j.ajhg.2014.09.002

Fournier-Level, A., A. Korte, M. D. Cooper, M. Nordborg, J. Schmitt *et al.*, 2011 A map of local adaptation in Arabidopsis thaliana. Science 334: 86–89. https://doi.org/10.1126/science.1209271

Gaggiotti, O. E., and M. Foll, 2010 Quantifying population structure using the F-model. Mol. Ecol. Resour. 10: 821–830. https://doi.org/10.1111/j.1755-0998.2010.02873.x

Grossman, S. R., K. G. Andersen, I. Shlyakhter, S. Tabrizi, S. Winnicki *et al.*, 2013 Identifying recent adaptations in large-scale genomic data. Cell 152: 703–713. https://doi.org/10.1016/j.cell.2013.01.035

Guo, F., D. K. Dey, and K. E. Holsinger, 2009 A Bayesian hierarchical model for analysis of SNP diversity in multilocus, multipopulation samples. J. Am. Stat. Assoc. 104: 142–154. https://doi.org/10.1198/jasa.2009.0010

Jones, F. C., M. G. Grabherr, Y. F. Chan, P. Russell, E. Mauceli *et al.*, 2012 The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484: 55–61. https://doi.org/10.1038/nature10944

Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle *et al.*, 2002 The human genome browser at UCSC. Genome Res. 12: 996–1006. https://doi.org/10.1101/gr.229102

Kern, A. D., and D. Haussler, 2010 A population genetic hidden Markov model for detecting genomic regions under selection. Mol. Biol. Evol. 27: 1673–1685. https://doi.org/10.1093/molbev/msq053

Leonardi, M., P. Gerbault, M. G. Thomas, and J. Burger, 2012 The evolution of lactase persistence in Europe. A synthesis of archaeological and genetic evidence. Int. Dairy J. 22: 88–97. https://doi.org/10.1016/j.idairyj.2011.10.010

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

Lotterhos, K. E., and M. C. Whitlock, 2014 Evaluation of demographic history and neutral parameterization on the performance of fst outlier tests. Mol. Ecol. 23: 2178–2192. https://doi.org/10.1111/mec.12725

Luu, K., E. Bazin, and M. G. B. Blum, 2017 pcadapt: an r package to perform genome scans for selection based on principal component analysis. Mol. Ecol. Resour. 17: 67–77. https://doi.org/10.1111/1755-0998.12592

Nei, M., and T. Maruyama, 1975 Lewontin-Krakauer test for neutral genes. Genetics 80: 395.

Neuenschwander, S., F. Michaud, and J. Goudet, 2018 QuantiNemo 2: a Swiss knife to simulate complex demographic and genetic scenarios, forward and backward in time. Bioinformatics 35: 886–888. https://doi.org/10.1093/bioinformatics/bty737

Nielsen, R., 2005 Molecular signatures of natural selection. Annu. Rev. Genet. 39: 197–218. https://doi.org/10.1146/annurev.genet.39.073003.112420

Peter, B. M., 2016 Admixture, population structure, and f-statistics. Genetics 202: 1485–1501. https://doi.org/10.1534/genetics.115.183913

Peter, B. M., D. Petkova, and J. Novembre, 2017 Genetic landscapes reveal how human genetic diversity aligns with geography. *Mol. Biol. Evol.* 37: 943–951.

Rannala, B. H., and J. A. Hartigan, 1996 Estimating gene flow in island populations. Genet. Res. 67: 147–158. https://doi.org/10.1017/S0016672300033607

Riebler, A., L. Held, and W. Stephan, 2008 Bayesian variable selection for detecting adaptive genomic differences among populations. Genetics 178: 1817–1829. https://doi.org/10.1534/genetics.107.081281

Rosenberg, N. A., J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd *et al.*, 2002 Genetic structure of human populations. Science 298: 2381–2385. https://doi.org/10.1126/science.1078311

Rosenberg, N. A., S. Mahajan, S. Ramachandran, C. Zhao, J. K. Pritchard, *et al.*, 2005 Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genetics 1: e70. https://doi.org/10.1371/journal.pgen.0010070

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832–837. https://doi.org/10.1038/nature01140

Sabeti, P. C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter *et al.*, 2007 Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913–918. https://doi.org/10.1038/nature06250

Slatkin, M., and L. Voelm, 1991 FST in a hierarchical island model. Genetics 127: 627–629.

Stölting, K. N., R. Nipper, D. Lindtke, C. Caseys, S. Waeber *et al.*, 2013 Genomic scan for single nucleotide polymorphisms reveals patterns of divergence and gene flow between ecologically divergent species. Mol. Ecol. 22: 842–855. https://doi.org/10.1111/mec.12011

Sugden, L. A., E. G. Atkinson, A. P. Fischer, S. Rong, B. M. Henn *et al.*, 2018 Localization of adaptive variants in human genomes using averaged one-dependence estimation. Nat. Commun. 9: 703. https://doi.org/10.1038/s41467-018-03100-7

Tang, K., K. R. Thornton, and M. Stoneking, 2007 A new approach for using genome scans to detect recent positive selection in the human genome. PLoS Biol. 5: e171. https://doi.org/10.1371/journal.pbio.0050171

Voight, B. F., S. Kudaravalli, X. Wen, and J. K. Pritchard, 2006 A map of recent positive selection in the human genome. PLoS Biol. 4: e72 Corrigenda: PLoS Biol. 5 e147 (2007). .https://doi.org/10.1371/journal.pbio.0040072

Wu, C. I., 2001 The genic view of the process of speciation. J. Evol. Biol. 14: 851–865. https://doi.org/10.1046/j.1420-9101.2001.00335.x