

# Detecting selection from linked sites using an $F$ -model

Marco Galimberti<sup>\*,†</sup>, Christoph Leuenberger<sup>‡</sup>, Beat Wolf<sup>§</sup>, Sándor Miklós Szilágyi<sup>\*\*</sup>, Matthieu Foll<sup>††,2</sup> and Daniel Wegmann<sup>\*,†,1</sup>

<sup>\*</sup>Department of Biology and Biochemistry, University of Fribourg, Fribourg, Switzerland, <sup>†</sup>Swiss Institute of Bioinformatics, Fribourg, Switzerland, <sup>‡</sup>Department of Mathematics, University of Fribourg, Fribourg, Switzerland, <sup>§</sup>iCoSys, University of Applied Sciences Western Switzerland, Fribourg, Switzerland,

<sup>\*\*</sup>Department of Informatics, University of Medicine, Pharmacy, Science and Technology of Târgu Mureş, Târgu Mureş, Romania, <sup>††</sup>International Agency for Research on Cancer (IARC/WHO), Section of Genetics, Lyon, France

## 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 two-fold 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 towards 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  $F_{ST}$  that measures population differ-

entiation, and loci with much 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 the 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 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

doi: 10.1534/genetics.XXX.XXXXXX

Manuscript compiled: Friday 16<sup>th</sup> October, 2020

<sup>1</sup>Department of Biology, University of Fribourg, Chemin du Musée 10, 1200 Fribourg, Switzerland, daniel.wegmann@unifr.ch

<sup>2</sup> Where authors are identified as personnel of the International Agency for Research on Cancer / World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer / World Health Organization.

1973), the difficulty to properly parameterize 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 (Balding 2003; Falush et al. 2003; Gaggiotti and Foll 2010; Rannala and Hartigan 1996), 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 bi-allelic loci, the current frequencies  $\tilde{p}_{jl}$  are then given by a beta distribution (Beaumont and Balding 2004)

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

where  $p_l$  are the frequencies in the ancestral population and  $\theta_{lj}$  is given by

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

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

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

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 auto-correlation 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 auto-correlation 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 auto-correlations in the strength of selection  $\alpha_l$  among linked markers. As we show using both simulations and an application to human data, aggregating information across loci results in an up to two-fold increase in power at the same false-discovery rate.

## 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  $j = 1, \dots, J$  at locus  $l = 1, \dots, L$  by  $\theta_{jl}$ . We further assume each locus to be bi-allelic 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

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

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

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

and hence

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

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

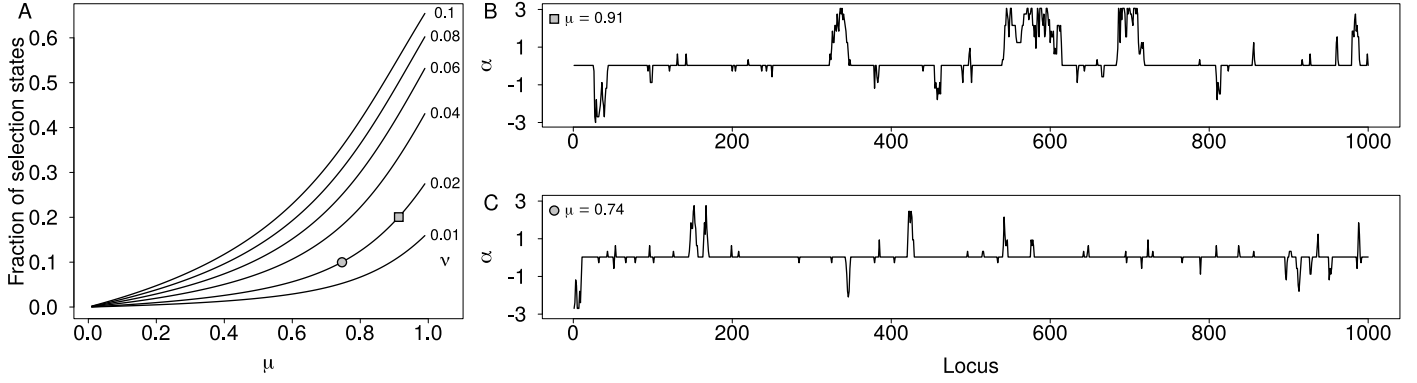
$$\mathbb{P}(n_{jl}|\theta_{jl}, p_l) = \binom{N_{jl}}{n_{jl}} \frac{B(\theta_{jl}p_l + n_{jl}, \theta_{jl}q_l + N_{jl} - n_{jl})}{B(\theta_{jl}p_l, \theta_{jl}q_l)}. \quad (5)$$

### Model of selection

We decompose  $\theta_{jl}$  into a population-specific component  $\beta_j$  shared by all loci, and a locus-specific component  $\alpha_l$  shared by all populations:

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

Here, the locus-specific component  $\alpha_l$  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



**Figure 1** (A) expected proportion of neutral sites as a function of rates  $\mu$  and  $\nu$ . (B and C) Example paths of  $\alpha_l$  along 1,000 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  $\log(\kappa) = -3.0$ ,  $\nu = 0.02$  and  $\mu = 0.91$  (B, square) or  $\mu = 0.74$  (C, circle), respectively. The two cases correspond to an expected proportion of 20% and 10% of the genome under selection, as marked in A.

of Foll and Gaggiotti (2008) and omit the error term  $\gamma_{ij}$  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 auto-correlation among the locus-specific component, we propose to discretize  $\alpha_l = \alpha(s_l)$ , where  $s_l = -s_{\max}, -s_{\max} + 1, \dots, s_{\max}$  are the states of a ladder-type Markov model with  $m = 2s_{\max} + 1$  states such that

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

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

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

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 auto-correlation measured by the positive scaling parameter  $\kappa$  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

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

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

$$\left( 0 \quad \dots \quad \nu\mu \quad -2\nu\mu \quad \nu\mu \quad \dots \quad 0 \right).$$

As exemplified in Figure 1, the two parameters  $\mu$  and  $\nu$  control the distribution of sites affected by selection (i.e. having  $\alpha_l \neq 0$ ) in the genome, with  $\nu$  affecting the number of selected regions and  $\mu$  their extent and selection strength, with higher values leading to more sites affected by selection. It is important to note that we do not assume all sites with  $\alpha_l \neq 0$  to be the targets

of selection. Instead, many will be linked to a target of selection and experience  $\alpha_l \neq 0$  due to hitch-hiking.

The stationary distribution of this Markov chain is given by

$$\Pi = c \cdot \left( 1 \quad \frac{1}{\mu} \quad \frac{1}{\mu^2} \quad \dots \quad \frac{1}{\mu^{s_l-1}} \quad \frac{1}{\mu^{s_l\nu}} \quad \frac{1}{\mu^{s_l-1}} \quad \dots \quad 1 \right),$$

with

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

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

### 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  $\hat{p}_{gjl}$  that derive from group-specific frequencies  $p_{gl}$  as described above with group-specific parameters  $\mu_g, \nu_g$  and  $\kappa_g$ . 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  $\Theta_{gl}$ . Specifically,

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

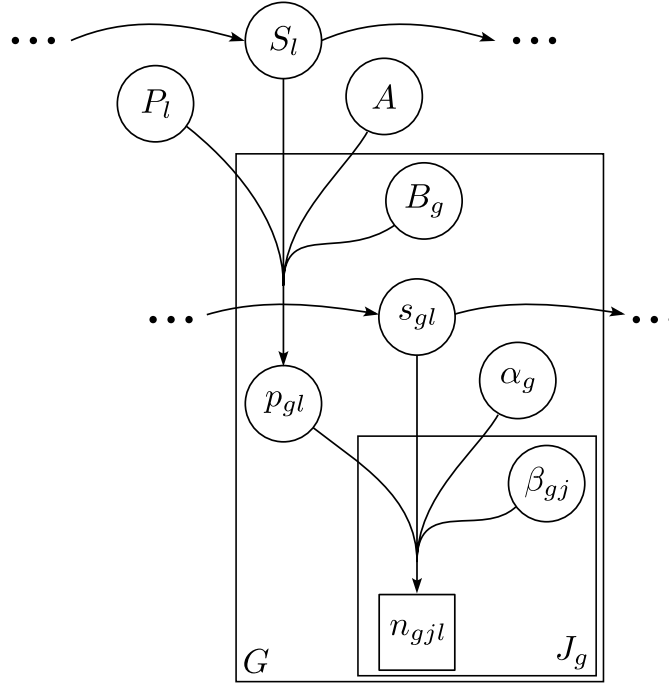
such that

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

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

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

As above,  $B_g$  quantifies group specific drift,  $S_l = -s_{\max}, -s_{\max} + 1, \dots, s_{\max}$  are the states of a Markov model



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

1 with  $m$  states and transition matrix  $\mathbf{Q}_l = e^{\kappa d_l \mathbf{L}}$  with parameters  
 2  $\mu$  and  $\nu$ , a positive scaling parameter  $\kappa$  and  $A(S_l)$  and  $A_{\max}$   
 3 defined as in (6). Hence, we assume independent HMM models  
 4 of the exact same structure at both levels of the hierarchy, as out-  
 5 lined in Figure 2. Additional levels could be added analogously.

### 6 Inference

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

$$\begin{aligned} \beta_j, B_g &\sim \mathcal{N}(\mu_b, \sigma_b^2) \\ p_l &\sim \text{Beta}(a_p, b_p) \\ \log(a_p), \log(b_p) &\sim \mathcal{N}(0, 1) \\ \log(\kappa_g), \log(\kappa), \log(\mu), \log(\nu) &\sim \mathcal{U}(-\infty, 0). \end{aligned}$$

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

13 To identify candidate regions under selection, we used  
 14 MCMC samples to determine the false-discovery rates (FDR)

$$\begin{aligned} q_d(l) &= 1 - \mathbb{P}(\alpha_l > 0 | \mathbf{n}, \mathbf{N}) \\ q_b(l) &= 1 - \mathbb{P}(\alpha_l < 0 | \mathbf{n}, \mathbf{N}) \end{aligned}$$

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

### 17 Implementation

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

20 Given the heavy computational burden of the proposed  
 21 model, we introduce several approximations. Most impor-  
 22 tantly, we group the distances  $d_l$  into  $E + 1$  ensembles such  
 23 that  $e_l = \lceil \log_2 d_l \rceil$ ,  $e_l = 0, \dots, E$  and use the same transition ma-  
 24 trix  $\mathbf{Q}(2^e)$  for all loci in ensemble  $e$ . We then calculate  $\mathbf{Q}(1)$  for

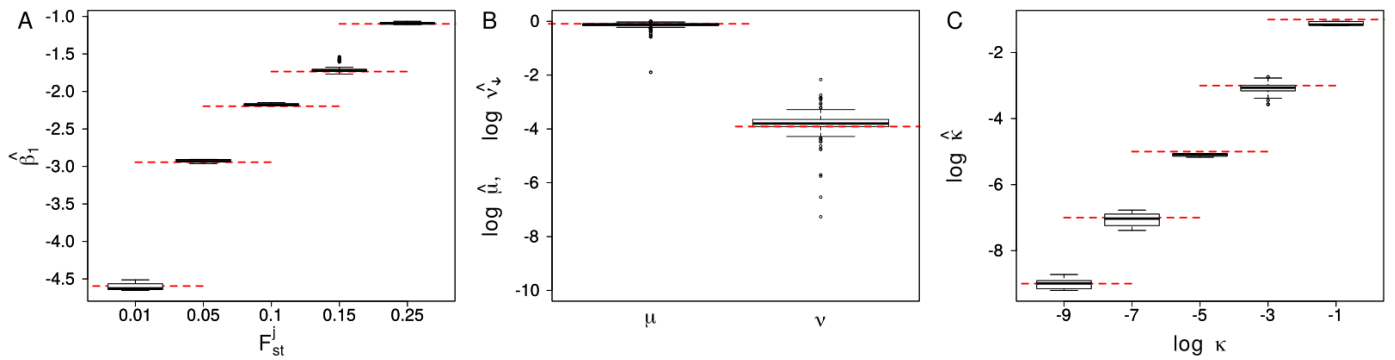
the first ensemble using the computationally cheap yet accurate  
 approximation

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

with  $r = \log_2(D/3) + 10$  where  $D = 2s_{\max} + 1$  is the dimension-  
 27 ality of the transition matrix (Ferrer-Admetlla *et al.* 2016). The  
 28 transition matrices of all other ensembles is obtained through the  
 29 recursion  $\mathbf{Q}(e) = \mathbf{Q}(e-1)^2$ . (See Supplementary Information  
 30 for other details regarding the implementation).  
 31

**Table 1** Parameters used in simulations

Name	J	$F_{ST}$	N	$\log(\kappa)$
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
$\log \kappa$ -1	10	0.15	50	-1
$\log \kappa$ -5	10	0.15	50	-5
$\log \kappa$ -7	10	0.15	50	-7
$\log \kappa$ -9	10	0.15	50	-9



**Figure 3** Boxplot of the parameters  $\beta_1$  (left),  $\nu$  and  $\mu$  (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.

### 1 Data availability

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

### 9 Results

#### 10 Comparison with BayeScan

11 **Simulation parameters** To quantify the benefits of accounting  
 12 for auto-correlation in the locus specific components  $\alpha_l$  among  
 13 linked loci, we first compared our method implemented in Flink  
 14 against the method implemented in BayeScan (Foll and Gaggiotti  
 15 2008) on simulated data. All simulations were conducted  
 16 under the model laid out above for a single group, using rou-  
 17 tines available in Flink and with parameter settings similar  
 18 to those used in (Foll and Gaggiotti 2008). Specifically, we fo-  
 19 cused on a reference simulation in which we sampled  $N = 50$   
 20 haplotypes from  $J = 10$  populations with  $\beta_j$  chosen such that  
 21  $F_{ST}^j = 0.15$  in the neutral case ( $\alpha_l = 0$ ). Following Foll and  
 22 Gaggiotti (2008), we simulated all  $p_l \sim \text{Beta}(0.7, 0.7)$  and about  
 23 20% of sites affected by selection (i.e. with  $\alpha_l \neq 0$ ) by setting  
 24  $\mu = 0.91$  and  $\nu = 0.02$ . We further set  $s_{\max} = 10$  (resulting in  
 25  $m = 21$  states) and  $\alpha_{\max} = 3$ , and simulated  $10^3$  loci for each  
 26 of 10 chromosomes, with a distance of 100 positions between  
 27 adjacent sites and strength of auto-correlation  $\log(\kappa) = -3$ . We  
 28 then varied the number of populations  $J$ , the sample size  $N$ ,  $F_{ST}^j$   
 29 or the strength of auto-correlation  $\kappa$  individually, while keeping  
 30 all other parameters constant (Table 1). We further added a case  
 31 without linkage (i.e.  $\kappa \rightarrow \infty$ ) by simulating each locus on its  
 32 own chromosome.

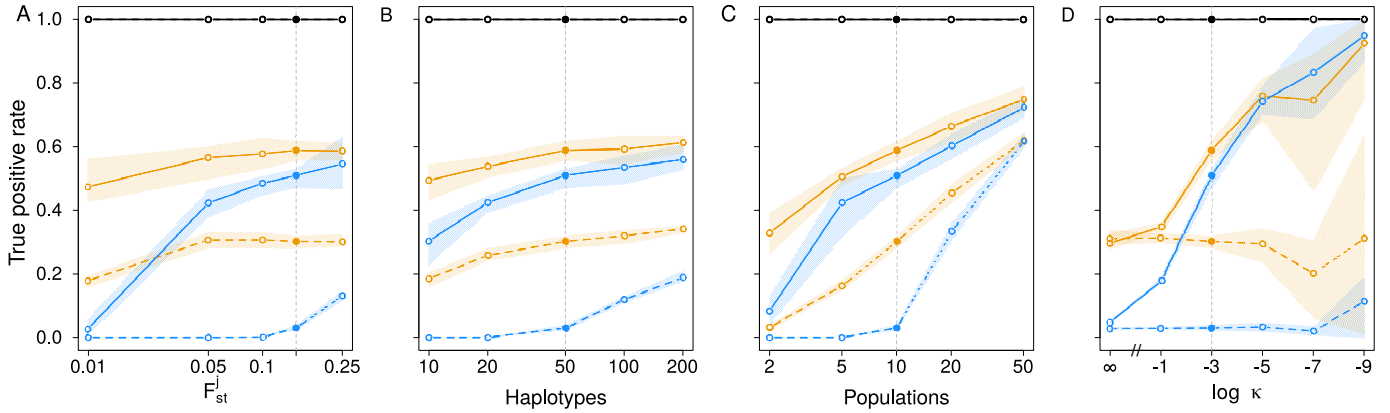
33 To infer parameters with Flink, we set  $s_{\max}$  and  $\alpha_{\max}$  to the  
 34 true values and ran the MCMC for  $7 \cdot 10^5$  iterations, of which  
 35 we discarded the first  $2 \cdot 10^5$  as burnin. During the chain, we  
 36 recorded parameter values every 100 iterations as posterior sam-  
 37 ples. To infer parameters with BayeScan, we used version 2.1  
 38 and set the prior odds for the neutral model to 50, which we  
 39 found to result in the same power as Flink in the reference sim-  
 40 ulation (see below) and in the absence of linkage ( $\kappa \rightarrow \infty$ ). We  
 41 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  $\beta_j$ ,  $\nu$ ,  $\mu$  and  $\kappa$ . As shown  
 through the distributions of posterior means across all simula-  
 tions, these estimates were very accurate and unbiased, re-  
 gardless of the parameter values used in the simulations (Fig-  
 ure 3). This suggests that the power to identify selected loci is  
 not limited by the number of loci we used to infer hierarchical  
 parameters.

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  $F_{ST}^j$ , the number of sam-  
 pled haplotypes and the number of sampled populations (Fig-  
 ure 4A-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 ( $F_{ST}^j = 0.01$ ) 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 (Supplementary 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 4A,B). Importantly, this increase in  
 power described here is fully explained by Flink accounting  
 for auto-correlation among the  $\alpha_l$  values as we chose the prior  
 odds in BayeScan to result in the same power if the strength of  
 auto-correlation vanishes (i.e.  $\kappa \rightarrow \infty$ ). Exploiting informa-  
 tion 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 ( $F_{ST}^j \leq 0.1$ , Figure 4A), in which case even  
 many neutral loci are expected to show virtually no difference in



**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 auto-correlation  $\kappa$  (D). Lines indicate the mean and range of true positive rates obtained with F1ink (solid) and BayeScan (dashed) across 10 replicate simulations. Filled dots and the vertical gray line indicate the reference simulation shown in each plot.

1 allele frequencies and only an aggregation of loci with a subtle  
 2 reduction in  $F_{ST}^j$  can be interpreted as a reliable signal for a target  
 3 of selection in the region (Foll and Gaggiotti 2008).

4 **Runtime** Thanks to careful optimization, there is little to no  
 5 overhead of our implementation compared to that of BayeScan.  
 6 On the reference simulation of  $10^4$  loci from 10 populations,  
 7 for instance, F1ink took on average 130 minutes on a modern  
 8 computer if calculations were spread over 4 CPU cores. On the  
 9 same data, BayeScan took 361 minutes. However, we note that  
 10 comparing the two implementations is difficult due to many  
 11 settings that strongly impact run times such as the number of  
 12 iterations or the use of pilot runs in BayeScan. Without pilot runs,  
 13 the run time of BayeScan reduced to 182 minutes on average  
 14 for the default number of iterations ( $10^5$  including burnin). In  
 15 the same time, F1ink runs for close to  $10^6$  iterations, but also  
 16 requires more to converge.

17 But since computation times scale linearly with the number  
 18 of loci, they remain prohibitively slow for whole genome appli-  
 19 cations in a single run. However, computations are easily spread  
 20 across many computers by analyzing the genome in independ-  
 21 ent chunks such as for each chromosome or chromosome arm  
 22 independently. This is justified because 1) linkage does not per-  
 23 sist across chromosome boundaries and is usually weak across  
 24 the centromere and 2) because our simulations indicate that  
 25  $10^4$  polymorphic loci were sufficient to estimate the hierarchi-  
 26 cal parameters accurately.

### 27 **Effect of model misspecification**

28 The  $F$ -model makes the explicit assumption that the allele fre-  
 29 quencies in a structured population can be characterized by  
 30 a multinomial Dirichlet distribution. This distribution is ap-  
 31 propriate for a wide range of demographic models, but not if  
 32 some pairs of populations share a more recent ancestry than  
 33 others (Beaumont and Balding 2004; Excoffier et al. 2009). Un-  
 34 surprisingly, several previous studies found high false-positive  
 35 rates when challenging BayeScan with models of isolation-by-  
 36 distance (IBD), recent range expansions, recent admixture or  
 37 asymmetric divergence (e.g. Lotterhos and Whitlock 2014; Luu

et al. 2017). These high false-positive rates are partially mitigated  
 by choosing higher prior odds (e.g. 50 as used here, Lotterhos  
 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 F1ink 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 quant iNemo (Neuenschwan-  
 der et al. 2018), we simulated genomic data from 11 popula-  
 tions 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 popula-  
 tions through symmetric dispersal between neighboring popula-  
 tions 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  $f_i \sim \text{Beta}(0.7, 0.7)$ . 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 centimor-  
 gans. We then inferred selection on all loci still polymorphic at  
 the end of the simulations with both BayeScan and F1ink for 10  
 replicates per set.

Across all simulations, BayeScan identified no locus as af-  
 fected as balancing selection and only 0.16% as affected by di-  
 vergent selection. This low false positive rate 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 F1ink  
 on the simulations with high recombination (genetic map of 100  
 centimorgans), in which case no linkage information could be  
 exploited. Across these simulations, F1ink identified no locus  
 as affected by balancing selection and only 0.14% as affected

1 by divergent selection. The number of false positives, however,  
2 was rising sharply with decreasing recombination rate. At a  
3 genetic map of 10 centimorgan, 5.0% and 2.8% of all loci were  
4 wrongly inferred as affected by balancing and divergent selec-  
5 tion, respectively. At a genetic map of only 1 centimorgan and  
6 hence tight linkage, the corresponding false positive rates were  
7 22.7% and 7.5%, respectively. These results thus highlight that  
8 the power gained by F1ink in exploiting linkage information  
9 also translates into a higher false positive rate in case the model  
10 is misspecified. Under such scenarios, other methods such as  
11 hapFLK (Fariello *et al.* 2013) or PCAdapt (Luu *et al.* 2017) are thus  
12 more appropriate.

### 13 Application to Humans

14 To illustrate the usefulness of F1ink we applied it to SNP data of  
15 46 populations analyzed as part of the Human Genome Diver-  
16 sity Project (HGDP) (Rosenberg N.A. *et al.* 2002; Rosenberg *et al.*  
17 2005) and available at <https://www.hagsc.org/hgdp/files.html>. We  
18 then used P1ink v1.90 (Chang *et al.* 2015) to transpose the data  
19 into vcf files and used the liftOver tool of the UCSC Genome  
20 Browser (Kent *et al.* 2002) to convert the coordinates to the hu-  
21 man reference GRCh38.

22 We divided the 46 populations into 6 groups (Table 2) of be-  
23 tween 4 and 15 populations each according to genetic landscapes  
24 proposed by Peter *et al.* (2017). We then inferred divergent and  
25 balancing selection using the hierarchical version of F1ink on all  
26 22 autosomes, but excluded 5 Mb on each side of the centromere  
27 and adjacent to the telomeres. The final data set consisted of  
28 563,589 SNPs. We analyzed each chromosome arm individually  
29 with  $\alpha_{\max} = 4.0$ ,  $s_{\max} = 10$  and using an MCMC chain with  
30  $7 \cdot 10^5$  iterations, of which we discarded the first  $2 \cdot 10^5$  as burnin.  
31 Estimates of hierarchical parameters are shown in Supplemen-  
32 tary Figure S2 and the locus-specific FDRs  $q_d(l)$  and  $q_b(l)$  are  
33 shown for all loci, all groups as well as the higher hierarchy in  
34 Supplementary Figures S4-S42. All regions identified as poten-  
35 tial targets for selection are further detailed in Supplementary  
36 Files. As summarized in Table 2, we discovered between 759 and  
37 1,889 and between 433 and 1,735 candidate regions for divergent  
38 and balancing selection, respectively, spanning together about  
39 10% of the genome.

40 **Comparison with BayeScan** We first validated our results by run-  
41 ning BayeScan on the same data. We then identified divergent  
42 regions as continuous sets of SNP markers that passed an FDR  
43 threshold of 0.01 or 0.05 for each method and determined the  
44 FDR threshold necessary to identify at least one locus within  
45 these regions by the other method. To ensure the observed dif-  
46 ferences between methods is due to accounting for linkage only,  
47 we used the hierarchical version BayeScanH (Foll *et al.* 2014) that  
48 also implements the same hierarchical island model as F1ink.

49 As shown in Figure 5A for selected regions among Europeans,  
50 the majority of regions identified by BayeScanH were replicated  
51 by F1ink at small FDR thresholds. In contrast, most of the  
52 regions identified by F1ink were not replicated by BayeScanH,  
53 in line with a higher statistical power for the former. Visual  
54 inspection indeed revealed that for most regions identified by  
55 F1ink but not BayeScanH, the latter also showed a signal of  
56 selection at multiple markers, each of which not passing the  
57 FDR threshold individually (see Figure 5B for examples). In  
58 contrast, sites identified by BayeScanH but not F1ink usually  
59 consisted of a signal at a single site, suggesting many of those  
60 are likely false positives (Figure 5C).

61 Results were similar for the other groups (Supplementary Fig-  
62 ure S3), but the correspondence between the methods was higher  
63 for the African group and considerably lower for the American  
64 group, likely due to the different patterns of divergence among  
65 populations (Supplementary Figure S2).

66 **Comparison with a recent scan for selective sweeps** Positive se-  
67 lection acting in a subset of populations may also lead to an  
68 increase in population differentiation (Nielsen 2005). We there-  
69 fore compared our outlier regions also to those of a recent scan  
70 for positive selection that combined multiple test for selection  
71 using a machine learning approach (Sugden *et al.* 2018). Among  
72 the 593 candidate loci reported for the CEU population of the  
73 1000 Genomes Project (1000 Genomes Project Consortium *et al.*  
74 2015) and overlapping the chromosomal segments studied here,  
75 293 loci (49.4%) fall within a region we identified as under diver-  
76 gent selection either among European populations (154 loci), at  
77 the higher hierarchy (132 loci), or both (7 loci).

78 To test if this overlap exceeds random expectations, we gener-  
79 ated 10,000 bootstrapped data sets by randomly sampling the  
80 same amount of loci among all those found polymorphic in the  
81 1000 Genome Project CEU samples and within the chromoso-  
82 mal segments studied here. We then determined the overlap  
83 with our outlier regions for each data set. On average, 46.6 loci  
84 overlapped with our regions identified among European popula-  
85 tions or at the higher hierarchy. Importantly, the largest overlap  
86 observed among the bootstrapped data set (72 loci) was much  
87 smaller than that observed (293 loci,  $p < 10^{-4}$ ).

88 **Example: The LCT region** As illustration, we show the FDRs  
89  $q_d(l)$  and  $q_b(l)$  for 30 Mb around the *LCT* gene in Figure 6 for  
90 the higher hierarchy as well as the European, Middle Eastern  
91 and East Asian group. The *LCT* gene is a well studied target  
92 of positive selection which has acted to increase lactase per-  
93 sistence in several human populations, including Europeans  
94 (Bersaglieri *et al.* 2004; Burger *et al.* 2020). Lactase persistence  
95 varies among Europeans and decreases on a roughly north-south  
96 cline (Bersaglieri *et al.* 2004; Leonardi *et al.* 2012; Burger *et al.*  
97 2007), consistent with the signal of divergent selection we de-  
98 tected among European populations (Figure 6). In line with  
99 previous findings (e.g. Grossman *et al.* 2013), we detected a  
100 signal of divergent selection among Europeans also in various  
101 genes around *LCT*, most notably in *R3HDM1* but also *MIR128-1*,  
102 *UBXN4* and *DARS*. In contrasts, we detected no such signal for  
103 the other groups.

## 104 Discussion

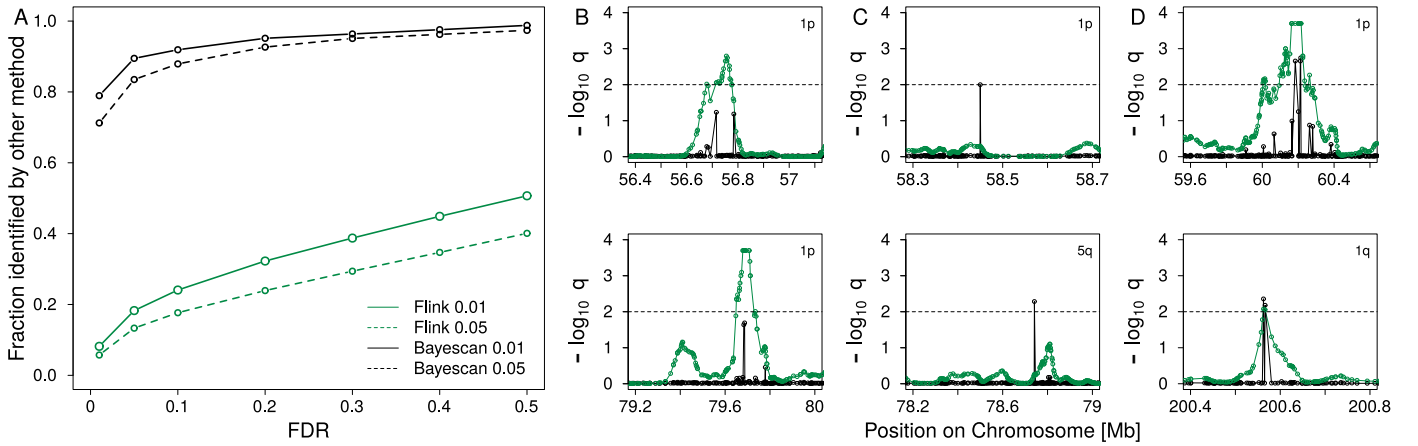
105 Genome scans are common methods to identifying loci that con-  
106 tribute to local adaptation among populations. Here we extend  
107 the particularly powerful method implemented in BayeScan  
108 (Foll and Gaggiotti 2008) to linked sites.

109 Accounting for linkage in population genetic methods, while  
110 desirable, is often computationally hard. We propose to alleviate  
111 this problem by modeling the dependence among linked sites  
112 through auto-correlation among hierarchical parameters, rather  
113 than the population allele frequencies or haplotypes themselves.  
114 In the context of genome scans, this has been previously used  
115 successfully to classify each locus as selected or neutral using  
116 Hidden-Markov Models (Boitard *et al.* 2009; Kern and Haussler  
117 2010). Here, we extend this idea by modeling auto-correlation  
118 among the degree of spatial differentiation acting at individ-  
119 ual loci. While ignoring auto-correlation at the genetic level

**Table 2 Population groups analyzed**

Group	Populations	Divergent			Balancing		
		SNPs (%)	Regions	Length <sup>a</sup>	SNPs (%)	Regions	Length <sup>a</sup>
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)	1,137	20.6	18,432 (3.27)	848	41.2
Europe	Adygei, French, French Basque, North Italian, Orcadian, Russian, Sardinian, Tuscan	19,128 (3.39)	1,466	22.0	37,736 (6.7)	1,382	48.3
America	Colombians, Karitiana, Maya, Pima, Surui	33,062 (5.87)	1,889	29.8	34,499 (6.12)	1,735	39.4
Central Asia	Balochi, Brahui, Burusho, Hazara, Kalash, Makrani, Pathan, Sindhi	16,663 (2.96)	1,290	22.6	25,473 (4.52)	1,132	44.5
East Asia	Uyгур, Dai, Daur, Han, Hezhen, Lahu, Miao zu, Mongola, Naxi, Oroqen, She, Tu, Tujia, Xibo, Yizu	20,528 (3.64)	1,832	17.3	33,678 (5.98)	1,656	35.2
Higher hierarchy	N/A	24,595 (4.36)	1,692	26.8	20,156 (3.58)	1,074	31.2

<sup>a</sup> Median length of the regions in kb.



**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.

1 certainly leads to a loss of information, the resulting method remains computationally tractable. And as we showed here with  
 2 simulations and an application to human data, the resulting  
 3 method features much improved statistical power compared to  
 4 BayeScan, a similar method that ignores linkage completely.  
 5

6 This is particularly evident for loci with more similar allele  
 7 frequencies among populations than expected by the genome-  
 8 wide divergence. These loci are generally interpreted as being  
 9 under balancing selection (Foll and Gaggiotti 2008; Beaumont  
 10 and Balding 2004), but may also be the result of purifying se-  
 11 lection restricting alleles from reaching high allele frequencies.  
 12 Given the large number of loci we inferred in this class from the  
 13 HGDP data (about 5% of the genome), we speculate that bal-

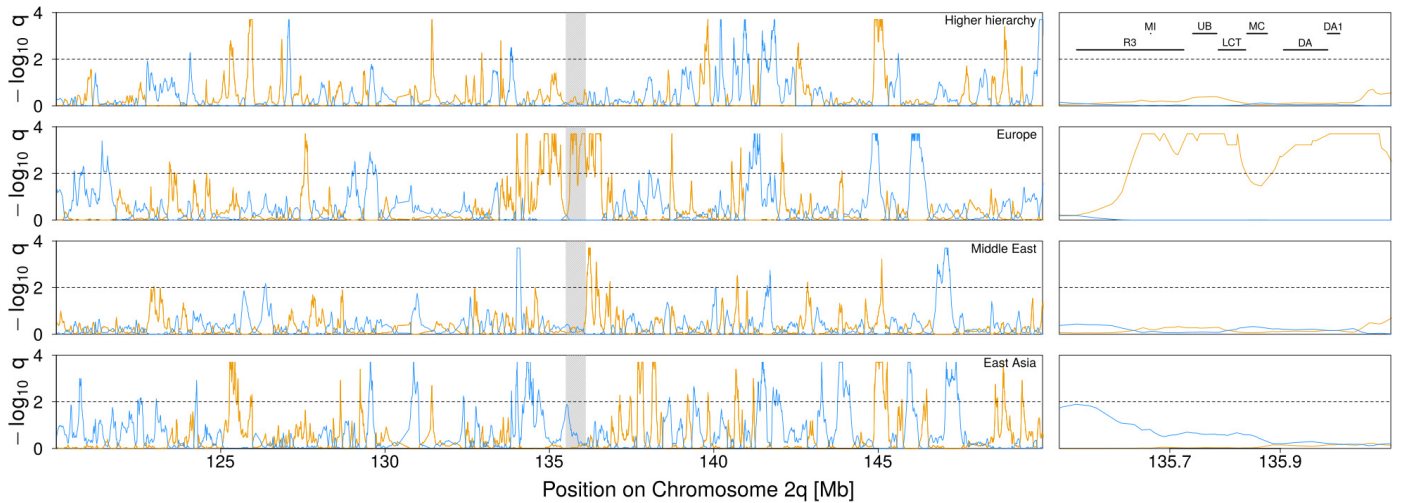
ancing selection is unlikely the main driver, and caution against  
 14 over-interpreting these results. But we note that the empirical  
 15 false discovery rate (FDR) for loci under balancing selection  
 16 was extremely low in our simulations, except if the assumptions  
 17 underlying the  $F$ -model was violated.  
 18

19 A benefit of accounting for auto-correlation among locus-  
 20 specific effects was previously postulated by Guo *et al.* (2009),  
 21 who proposed a conditional autoregressive (CAR) prior on  $\alpha_l$   
 22 such that

$$\alpha_l | \alpha_{-l} \sim \mathcal{N} \left( \frac{1}{\bar{w}_l} \sum_{m \neq l} w_{lm} \alpha_m, \frac{1}{\tau \bar{w}_l} \right),$$

where  $\alpha_{-l}$  denotes the collection of all other  $\alpha_m, m \neq l, w_{lm}$  23





**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 Supplementary Figure S7.

1 indicates the covariance between loci  $l$  and  $m$ , which is assumed  
 2 to decrease exponentially with distance, and  $\bar{w}_i = \sum_{m \neq l} w_{ml}$ .  
 3 While Guo *et al.* (2009) did not evaluate the benefit of their CAR  
 4 implementation on the power of selection inference, they found  
 5 that it was a better fit to high resolution data. Here we show that  
 6 the power increase by exploiting auto-correlation among loci is  
 7 substantial: of all regions identified as under divergent selection  
 8 by F1ink, less than half were also identified by BayeScan,  
 9 despite evidence that these consist mostly of true outliers.

10 In this context, it is important to note that due to computa-  
 11 tional challenges, Guo *et al.* (2009) suggested to run their method  
 12 on low-resolution data with few markers first, and then to apply  
 13 the CAR version on inferred candidate regions only. As our  
 14 analysis suggests, such an approach would likely fail to har-  
 15 vest the full benefit of accounting for auto-correlation among  
 16 locus-specific parameters. Running F1ink on high-resolution  
 17 data is possible because the first-order Markov assumption on  
 18 locus-specific effects  $\alpha_l$  allows for cheap MCMC updates at a  
 19 single locus that does not require a recalculation of the prior  
 20 on the full vector  $\alpha = \{\alpha_1, \dots, \alpha_L\}$ . Unfortunately, however, no  
 21 implementation of the method by Guo *et al.* (2009) is available  
 22 for a direct comparison.

23 Our proposed model has yet another computational advan-  
 24 tage: while the hierarchical parameters of the exponential decay  
 25 in the model by Guo *et al.* (2009) need to be fixed upfront due  
 26 to numerical instabilities, the hierarchical parameters of the dis-  
 27 crete Markov model proposed here are all estimated well if  
 28 sufficient sites are provided. Our simulations indicated that  $10^4$   
 29 polymorphic loci were sufficient, based on which we decided  
 30 to parallelize the analysis of the human data by chromosome  
 31 arm. Smaller windows may be considered, but the model may  
 32 struggle to differentiate between population-specific and locus-  
 33 specific components if too few consecutive loci are used. The  
 34 window analyzed should therefore span significantly more loci  
 35 than are expected to be affected by selection within an outlier  
 36 region. But note that the model does not make any assumption  
 37 regarding the spacing of loci within the analyzed window, nor  
 38 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, F1ink 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 F1ink and the CAR  
 method of Guo *et al.* (2009) is that the former discretizes the locus-  
 specific effects  $\alpha_l$ . While such a discretization leads to a loss of  
 precision in estimating locus specific effects, it allows to directly  
 calculate a false-discovery rate to identify outlier loci at any de-  
 sired 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 dis-  
 tributions on  $\theta_l$  are significantly different from the distribution  
 of  $\theta_l$  values under the inferred hyper-parameters. Importantly,  
 the discretization seems to come at no cost on power: in our  
 simulations, F1ink and BayeScan had virtually identical power  
 if we simulated unlinked data.

An obvious draw-back of modeling the locus-specific selec-  
 tion 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 vari-  
 ant. 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 ( $x$ )  
 (Sugden *et al.* 2018) may further be used on the identified regions  
 to infer locus-specific selection coefficients or other statistics in-  
 formative about the targets of selection.

We finally note that the implementation provided through  
 F1ink 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 as-  
 sociation 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 ac-

1 curately describe population structure and loci associated with  
2 the phenotype will be identified as those highly divergent be-  
3 tween the two “populations”. A natural assumption would then  
4 be that the locus-specific coefficients  $\alpha_i$  are shared among all  
5 groups, i.e. that they are governed by a single HMM. While  
6 we have not made use of such a setting here, we note that it is  
7 readily available as an option in F1 ink.

## 8 Acknowledgments

9 We are grateful for the very constructive feedback of two anony-  
10 mous reviewers as the editor Nick Barton on an earlier version  
11 of this manuscript. This study was supported by two Swiss Na-  
12 tional Foundation grants to DW with numbers 31003A\_149920  
13 and 31003A\_173062.

## 14 Literature Cited

15 1000 Genomes Project Consortium, A. Auton, L. D. Brooks, R. M.  
16 Durbin, E. P. Garrison, *et al.*, 2015 A global reference for human  
17 genetic variation. *Nature* **526**: 68–74.  
18 Andrew, R. L. and L. H. Rieseberg, 2013 Divergence is focused on  
19 few genomic regions early in speciation: Incipient speciation  
20 of sunflower ecotypes. *Evolution* **67**: 2468–2482.  
21 Balding, D. J., 2003 Likelihood-based inference for genetic cor-  
22 relation coefficients. *Theoretical Population Biology* **63**: 221  
23 – 230, Uses of DNA and genetic markers for forensics and  
24 population studies.  
25 Beaumont, M. and R. A. Nichols, 1996 Evaluating loci for use in  
26 the genetic analysis of population structure. *P.Roy.Soc.Lond.B*  
27 **263**: 1619–1626.  
28 Beaumont, M. A. and D. J. Balding, 2004 Identifying adaptive  
29 genetic divergence among populations from genome scans.  
30 *Molecular Ecology* **13**: 969–980.  
31 Bersaglieri, T., P. C. Sabeti, N. Patterson, T. Vanderploeg, S. F.  
32 Schaffner, *et al.*, 2004 Genetic Signatures of Strong Recent  
33 Positive Selection at the Lactase Gene. *The American Journal*  
34 *of Human Genetics* **74**: 1111–1120.  
35 Boitard, S., C. Schlötterer, and A. Futschik, 2009 Detecting se-  
36 lective sweeps: A new approach based on hidden Markov  
37 models. *Genetics* **181**: 1567–1578.  
38 Bonin, A., P. Taberlet, C. Miaud, and F. Pompanon, 2006 Explo-  
39 rative genome scan to detect candidate loci for adaptation  
40 along a gradient of altitude in the common frog (*Rana tempo-*  
41 *raria*). *Mol.Biol.Evol.* **23**: 773–783.  
42 Burger, J., M. Kirchner, B. Bramanti, W. Haak, and M. G.  
43 Thomas, 2007 Absence of the lactase-persistence-associated al-  
44 lele in early Neolithic Europeans. *Proceedings of the National*  
45 *Academy of Sciences* **104**: 3736–3741.  
46 Burger, J., V. Link, J. Blöcher, A. Schulz, C. Sell, *et al.*, 2020 Low  
47 prevalence of lactase persistence in bronze age europe indi-  
48 cates ongoing strong selection over the last 3,000 years. *Current*  
49 *Biology* .  
50 Chang, C. C., C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell,  
51 *et al.*, 2015 Second-generation PLINK: Rising to the challenge  
52 of larger and richer datasets. *GigaScience* **4**: 1–16.  
53 Cruickshank, T. E. and M. W. Hahn, 2014 Reanalysis suggests  
54 that genomic islands of speciation are due to reduced diversity,  
55 not reduced gene flow. *Mol.Ecol.* **23**: 3133–3157.  
56 Durand, E. Y., N. Patterson, D. Reich, and M. Slatkin, 2011 Test-  
57 ing for ancient admixture between closely related populations.  
58 *Mol.Biol.Evol.* **28**: 2239–2252.  
59 Eriksson, A. and A. Manica, 2012 Effect of ancient population  
60 structure on the degree of polymorphism shared between

61 modern human populations and ancient hominins. *Proceed-*  
62 *ings of the National Academy of Sciences* **109**: 13956–13960.  
63 Excoffier, L., T. Hofer, and M. Foll, 2009 Detecting loci under  
64 selection in a hierarchically structured population. *Heredity*  
65 **103**: 285–298.  
66 Falush, D., M. Stephens, and J. K. Pritchard, 2003 Inference of  
67 population structure using multilocus genotype data: Linked  
68 loci and correlated allele frequencies. *Genetics* **164**: 1567–1587.  
69 Fariello, M. I., S. Boitard, H. Naya, M. SanCristobal, and  
70 B. Servin, 2013 Detecting signatures of selection through hap-  
71 lotype differentiation among hierarchically structured popula-  
72 tions. *Genetics* **193**: 929–941.  
73 Feder, J. L., S. P. Egan, and P. Nosil, 2012 The genomics of  
74 speciation-with-gene-flow. *Trends in Genetics* **28**: 342–350.  
75 Ferrer-Admetlla, A., C. Leuenberger, J. D. Jensen, and D. Weg-  
76 mann, 2016 An Approximate Markov Model for the Wright-  
77 Fisher Diffusion. *Genetics* **203**: 831–846.  
78 Foll, M. and O. Gaggiotti, 2008 A genome-scan method to iden-  
79 tify selected loci appropriate for both dominant and codominant  
80 markers: A Bayesian perspective. *Genetics* **180**: 977–993.  
81 Foll, M., O. E. Gaggiotti, J. T. Daub, A. Vatsiou, and L. Excoffier,  
82 2014 Widespread signals of convergent adaptation to high  
83 altitude in Asia and America. *American Journal of Human*  
84 *Genetics* **95**: 394–407.  
85 Fournier-Level, A., A. Korte, M. D. Cooper, M. Nordborg,  
86 J. Schmitt, *et al.*, 2011 A map of local adaptation in arabidopsis  
87 thaliana. *Science* **334**: 86–89.  
88 Gaggiotti, O. E. and M. Foll, 2010 Quantifying population struc-  
89 ture using the F-model. *Molecular Ecology Resources* **10**: 821–  
90 830.  
91 Grossman, S. R., K. G. Andersen, I. Shlyakhter, S. Tabrizi, S. Win-  
92 nicki, *et al.*, 2013 Identifying recent adaptations in large-scale  
93 genomic data. *Cell* **152**: 703–13.  
94 Guo, F., D. K. Dey, and K. E. Holsinger, 2009 A Bayesian hier-  
95 archical model for analysis of SNP diversity in multilocus,  
96 multipopulation samples. *Journal of the American Statistical*  
97 *Association* **104**: 142–154.  
98 Jones, F. C., M. G. Grabherr, Y. F. Chan, P. Russell, E. Mauceli,  
99 *et al.*, 2012 The genomic basis of adaptive evolution in three-  
100 spine sticklebacks. *Nature* **484**: 55–61.  
101 Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle,  
102 *et al.*, 2002 The human genome browser at UCSC. *Genome*  
103 *Research* **12**: 996–1006.  
104 Kern, A. D. and D. Haussler, 2010 A population genetic hidden  
105 markov model for detecting genomic regions under selection.  
106 *Molecular Biology and Evolution* **27**: 1673–1685.  
107 Leonardi, M., P. Gerbault, M. G. Thomas, and J. Burger, 2012  
108 The evolution of lactase persistence in Europe. A synthesis  
109 of archaeological and genetic evidence. *International Dairy*  
110 *Journal* **22**: 88–97.  
111 Lewontin, R. C. and J. Krakauer, 1973 Distribution of gene fre-  
112 quency as a test of the theory of the selective neutrality of  
113 polymorphisms. *Genetics* **74**: 175–195.  
114 Lotterhos, K. E. and M. C. Whitlock, 2014 Evaluation of demo-  
115 graphic history and neutral parameterization on the perfor-  
116 mance of *fst* outlier tests. *Molecular Ecology* **23**: 2178–2192.  
117 Luu, K., E. Bazin, and M. G. B. Blum, 2017 pcadapt: an r package  
118 to perform genome scans for selection based on principal  
119 component analysis. *Molecular Ecology Resources* **17**: 67–77.  
120 Nei, M. and T. Maruyama, 1975 Lewontin-krakauer test for neu-  
121 tral genes. *Genetics* **80**: 395–395.  
122 Neuenschwander, S., F. Michaud, and J. Goudet, 2018 Quan-

1 tiNemo 2: a Swiss knife to simulate complex demographic  
2 and genetic scenarios, forward and backward in time. *Bioin-*  
3 *formatics* **35**: 886–888.

4 Nielsen, R., 2005 Molecular signatures of natural selection. *Annual*  
5 *Review of Genetics* **39**: 197–218, PMID: 16285858.

6 Peter, B. M., 2016 Admixture, population structure, and f-  
7 statistics. *Genetics* **202**: 1485–1501.

8 Peter, B. M., D. Petkova, and J. Novembre, 2017 Genetic land-  
9 scapes reveal how human genetic diversity aligns with geog-  
10 raphy. *bioRxiv* pp. 1–24.

11 Rannala, B. H. and J. A. Hartigan, 1996 Estimating gene flow in  
12 island populations. *Genetical Research* **67**: 147–158.

13 Riebler, A., L. Held, and W. Stephan, 2008 Bayesian variable  
14 selection for detecting adaptive genomic differences among  
15 populations. *Genetics* **178**: 1817–1829.

16 Rosenberg, N. A., S. Mahajan, S. Ramachandran, C. Zhao, J. K.  
17 Pritchard, *et al.*, 2005 Clines, clusters, and the effect of study  
18 design on the inference of human population structure. *PLoS*  
19 *Genetics* **1**: 0660–0671.

20 Rosenberg N.A., Pritchard J.K., Weber J.L., Cann H.M., Kidd  
21 K.K., *et al.*, 2002 Genetic structure of human populations. *Sci-*  
22 *ence* **298**: 2981–2985.

23 Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J.  
24 Richter, *et al.*, 2002 Detecting recent positive selection in the  
25 human genome from haplotype structure. *Nature* **419**: 832–  
26 837.

27 Sabeti, P. C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter, *et al.*,  
28 2007 Genome-wide detection and characterization of positive  
29 selection in human populations. *Nature* **449**: 913–918.

30 Slatkin, M. and L. Voelm, 1991  $F_{ST}$  in a hierarchical island model.  
31 *Genetics* **127**: 627–9.

32 Stölting, K. N., R. Nipper, D. Lindtke, C. Caseys, S. Waeber,  
33 *et al.*, 2013 Genomic scan for single nucleotide polymorphisms  
34 reveals patterns of divergence and gene flow between ecologi-  
35 cally divergent species. *Mol.Ecol.* **22**: 842–855.

36 Sugden, L. A., E. G. Atkinson, A. P. Fischer, S. Rong, B. M.  
37 Henn, *et al.*, 2018 Localization of adaptive variants in human  
38 genomes using averaged one-dependence estimation. *Nature*  
39 *Communications* **9**: 1–14.

40 Tang, K., K. R. Thornton, and M. Stoneking, 2007 A new ap-  
41 proach for using genome scans to detect recent positive selec-  
42 tion in the human genome. *PLOS Biology* **5**: 1–16.

43 Voight, B. F., S. Kudravalli, X. Wen, and J. K. Pritchard, 2006 A  
44 map of recent positive selection in the human genome. *PLOS*  
45 *Biology* **4**.

46 Wu, 2001 The genic view of the process of speciation. *J.Evol.Biol.*  
47 **14**: 851–865.