Detecting selection from linked sites using an F-model

Marco Galimberti[*,†], **Christoph Leuenberger**[‡], **Beat Wolf**[§], **Sándor Miklós Szilágyi**[**], **Matthieu Foll**[††,2] **and Daniel Wegmann**[*,†,1]

[*]Department of Biology and Biochemistry, University of Fribourg, Fribourg, Switzerland, [†]Swiss Institute of Bioinformatics, Fribourg, Switzerland, [‡]Department of Mathematics, University of Fribourg, Fribourg, Switzerland, [§]iCoSys, University of Applied Sciences Western Switzerland, Fribourg, Switzerland, [**]Department of Informatics, University of Medicine, Pharmacy, Science and Technology of Târgu Mureş, Târgu Mureş, Romania, [††]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) that characterizes the effect of selection on linked markers through correlations in the locus specific component along the genome. Using extensive simulations we show that our method has up to two-fold the statistical power 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 such statistics is F_{ST} that measures population differentiation, 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 assessing incongruence between a population tree and the genealogy at a locus (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 (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 be identifying 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 expecta-

doi: 10.1534/genetics.XXX.XXXXXX
Manuscript compiled: Thursday 15th August, 2019

[1]Department of Biology, University of Fribourg, Chemin du Musée 10, 1200 Fribourg, Switzerland, daniel.wegmann@unifr.ch

[2] 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.

tions is far from new (e.g. Lewontin and Krakauer 1973), the difficulty to properly parameterize the null model was quickly realized Nei and Maruyama (e.g. 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 heterozygousity 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 (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)

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

where p_l are the frequencies in the ancestral population and θ_{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 effect 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 a population specific components α_l and β_j, respectively:

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

Loci with $\alpha_l \neq 0$ are interpreted to be targets of either balancing ($\alpha_l < 0$) or divergent ($\alpha_l > 0$) selection (Beaumont and Balding 2004). Targets of selection may then be identified by contrasting models with $\alpha_l = 0$ or $\alpha_l \neq 0$ for each locus l, as is for instance done using reversible-jump MCMC 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 evaluation 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 positivs may be avoided as their signals is unlikely shared with linked loci.

Unfortunately, fully accounting for linkage is often statistically challenging as well as computationally very costly. A much more feasible approach is to model linkage through the auto-correlation of hierarchical parameters along the genome. 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 α_l among linked markers. As we show using both simulations and an application to human data, aggregating information across loci results in up to two-fold 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 $j = 1, \ldots, J$ at locus $l = 1, \ldots, L$ by θ_{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 (1). We thus have

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

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

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

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

In the absence of selection, all loci are assumed to experience the same amount of population specific drift. Following Beaumont and Balding (2004), we thus 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, as shown in (2).

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

$$\alpha(S_l) = \frac{s}{s^{max}} \alpha^{max} \tag{6}$$

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

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

with elements $[Q(d_l)]_{ij}$ denoting the probabilities to go from state i at locus $l-1$ to state j at locus l at known distance d_l and given the strength of auto-correlation measured by the positive scaling parameter κ. Here, Λ is the $m \times m$ generating matrix

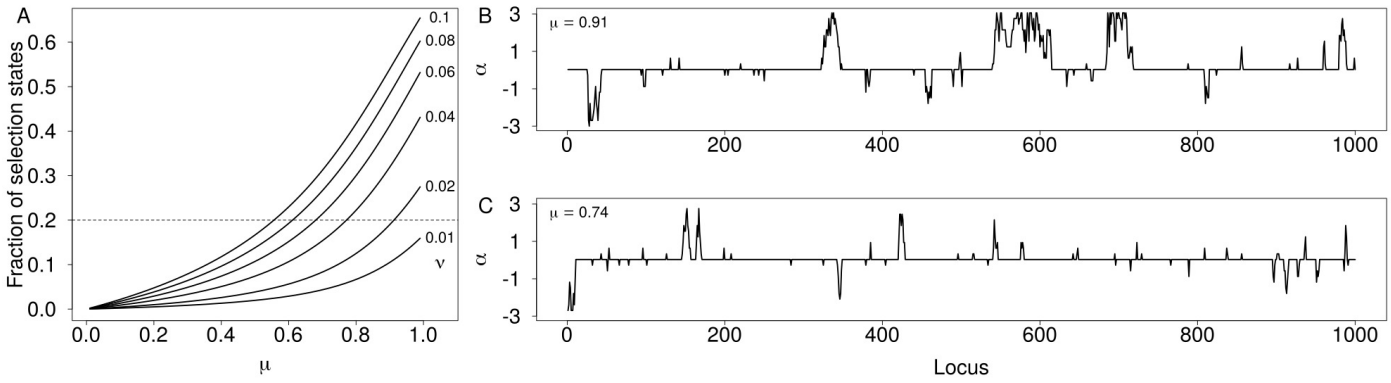

Figure 1 (A) Proportion of neutral sites as a function of μ and ν. The dashed line indicates a fraction of 80%. (B and C) Example trajectories of $alpha_l$ along 1,000 loci simulated with $s_{max} = 10$, $\alpha_{max} = 3.0$, $\log(k) = -3.0$, $d_l = 100$, $\nu = 0.02$ and $\mu = 0.91$ (B) and $\mu = 0.74$ (C), respectively.

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

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

$$\begin{pmatrix} 0 & \dots & \nu\mu & -2\nu\mu & \nu\mu & \dots & 0 \end{pmatrix}.$$

As exemplified in Figure 1, the two parameters μ and ν control the distribution of sites under selection in the genome with large ν affecting the number of selected regions and μ their extent and selection strength, with higher values leading to more sites under selection. The stationary distribution of this Markov chain is given by

$$\Pi = c \cdot \begin{pmatrix} 1 & \frac{1}{\mu} & \frac{1}{\mu^2} & \dots & \frac{1}{\mu^{s-1}} & \frac{1}{\mu^s\nu} & \frac{1}{\mu^{s-1}} & \dots & 1 \end{pmatrix},$$

with

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

Note that as $\kappa \to \infty$, 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 the 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 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

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

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, $S_l = -s^{max}, -s^{max}+1, \dots, s^{max}$ are the states of a Markov model with m states and transition matrix $Q_l = e^{\kappa d_l \Lambda}$ 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 all levels of the hierarchy, as outlined in Figure 2.

Inference

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

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

Following Beaumont and Balding (2004), we used $\mu_b = 0$ and $\sigma_b^2 = 1.8$ throughout. We further set $a_p = b_p = 1$.

To identify candidate regions under selection, we used our MCMC samples to determine the false-discovery rates

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

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

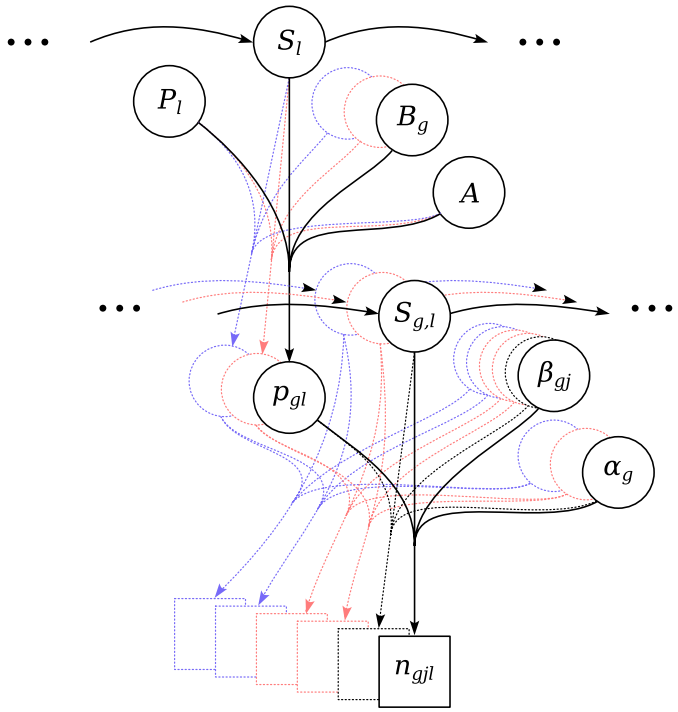

Figure 2 A directed acyclic graph (DAG) of the proposed hierarchical model with three groups (black, red and blue) of two populations each.

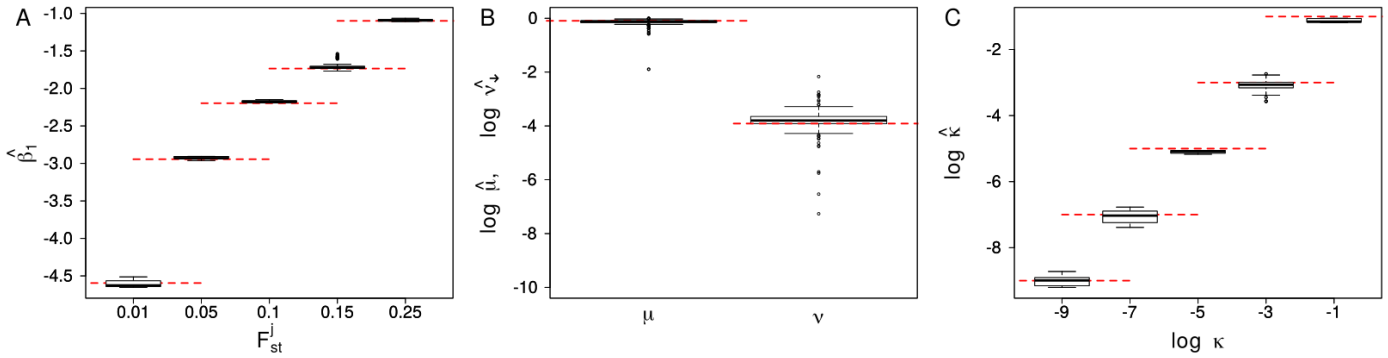

Figure 3 Boxplot of the parameters β_1 (left), ν and μ (center) and $\log(\kappa)$ (right). The values are obtained from the mean of the posterior distributions obtained using `Flink` on the 10 simulations run for each of the set of parameters reported in Table 1. The red dotted lines show the true values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

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 $e_l = \lceil \log_2 d_l \rceil$, $e_l = 0, \ldots, E$ and use the same transition matrix $Q(2^e)$ for all loci in ensemble e. We then calculate $Q(1)$ for the first ensemble using the computationally cheap yet accurate approximation

$$Q_0 = e^{\kappa d_0 \Lambda} \approx \left(\mathcal{I} + \frac{1}{2^m} \kappa d_0 \Lambda \right)^{2^m}$$

with $m = \log_2(D/3) + 10$ where $D = 2s^{max} + 1$ is the dimensionality of the transition matrix (Ferrer-Admetlla *et al.* 2016). The transition matrices of all other ensembles can then be obtained through the recursion $Q(e) = Q(e-1)^2$. (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.

Simulation study

Simulation parameters

To quantify the benefits of accounting for auto-correlation in the locus specific components α_l among linked loci, we used

simulations to compare the power to identify loci under selection of our method implemented in `Flink` against the method implemented in `BayeScan` (Foll and Gaggiotti 2008). 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 $F_{ST}^{lj} = 0.15$ in the neutral case $\alpha_l = 0$. We then varied the number of populations J, the sample size N, F_{ST}^{lj} or the strength of auto-correlation κ individually, while keeping all other parameters constant able 1). Following Foll and Gaggiotti (2008), we simulated all $p_l \sim \text{Beta}(0.7, 0.7)$ and 20% of sites under selection by setting $\mu = 0.91$ and $\nu = 0.02$. We further set $s^{max} = 10$ (resulting in $m = 21$ states) and $\alpha_g^{max} = 3$ for all simulations. We simulated 10^3 loci for each of 10 chromosomes, with a distance of 100 between adjacent sites.

To infer parameters with `Flink`, we set s^{max} and α^{max} to the true values and ran the MCMC for $7 \cdot 10^5$ iterations, of which we discarded the first $2 \cdot 10^5$ as burnin. During the chain, we recorded all parameter values every 100 iterations as posterior samples. To infer parameters with `Bayescan`, we used version 2.1 with default settings. We identified loci under selection at a False-Discovery-Rate (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 accurat 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 used.

We next studied the impact of the sample size and the strength of population differentiation on power. In line with findings reported by (Foll and Gaggiotti 2008), power generally increased with F_{ST}^j, the number of sampled haplotypes and the number of sampled populations (Figure 4A-C). Importantly, 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 ($J = 2$).

We finally compared the power of `Flink` to that of `BayeScan` on the same set of simulations. As shown in Figure 4, `Flink` had a higher power at the same FDR across all simulations, and often considerably so, unless the number of populations sampled was large. 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 is fully explained by `Flink` accounting for auto-correlation among the α_l values. As shown in Figure 4D, the power of both methods converges as soon as the strength of auto-correlation vanishes (i.e. κ is large). 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 ($F_{ST}^j \leq 0.1$, Figure 4A), in which case even many neutral loci are expected to show virtually no difference in allele frequency and only an aggregation of such loci can be interpreted as a reliable signal for selection (Foll and Gaggiotti 2008).

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

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 minutes on a modern computer if calculations were spread over 4 CPU cores. On the same data, `BayeScan` took 361 minutes. 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 minutes on average for the default number of iterations (10^5 including burnin). In the same time, `Flink` runs for close to 10^6 iterations, but also requires more to converge.

But since computation times scale linerly with the number of loci, they remain prohibitively slow for whole genome applications in a single run. However, they 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 also weak across the centromere and 2) because our simulations indicate that 10^4 polymorphic loci were sufficient to estimate the hierarchical parameters accurately.

Application to Humans

To illustrate the usefulness of `Flink` we applied it to SNP data of 46 populations analyzed as part of the Human Genome Diversity Project (HGDP) (Rosenberg N.A. *et al.* 2002; Rosenberg *et al.* 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 (James Kent *et al.* 2002) to convert the coordinates to the human reference GRCh38.

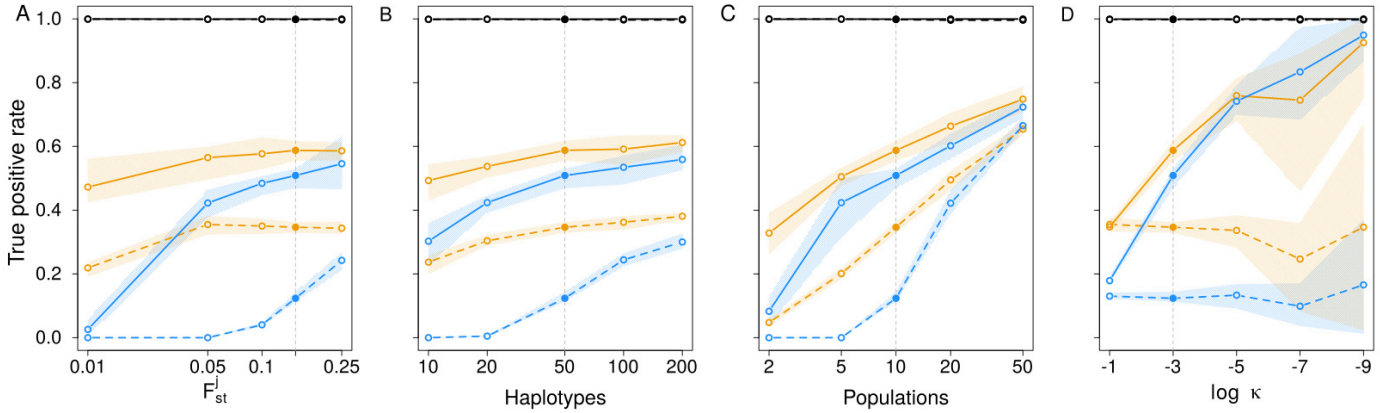

Figure 4 The true positive rate in classifying loci as neutral (black) or under divergent (orange) or balancing selection (blue) as a function of the F_{st} between populations (A), the number of haplotypes N (B), the number of populations J (C) and the strength of auto-correlation κ (D). Lines indicate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

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)	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	Uygur, Dai, Daur, Han, Hezhen, Lahu, Miaozu, 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

[a] Median length of the regions in kb.

We divided the 46 populations into 6 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 centromer and adjacent to the telomeres. The final data set consists in total 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 burnin. 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 Supplementary Figures S4-S42. All regions identified as potential targets for selection are further detailed in Supplementary Files. As summarized in Table 2, we discovered between 759 and 1,889 and between 433 and 1,735 candidate regions for divergent and balancing selection, respectively, spanning together about 10% of the genome.

Comparison to BayeScan

We first validated our results by running `BayeScan` on the same data but for each group individually. We then identified divergent regions as continuous sets of SNP markers that passed an FDR threshold of 0.01 or 0.01 for each method and determined the FDR threshold necessary to identify at least one locus within these regions by the other method. As shown in Figure 5A

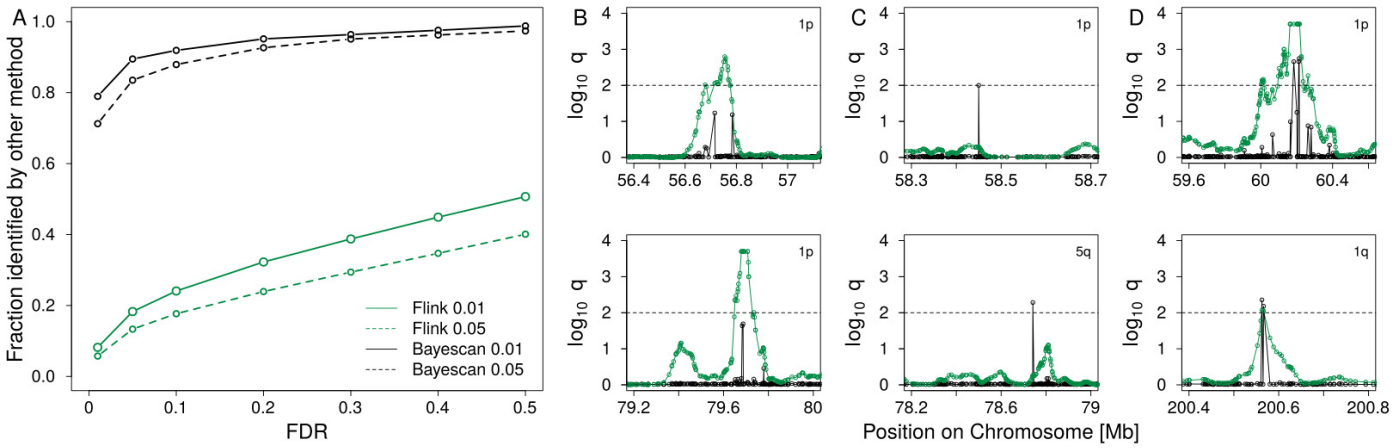

Figure 5 (A) The fraction of regions identified as divergent among Europeans by `Flink` (green) and `Bayescan` (black) at a false discovery rate (FDR) of 0.01 (solid) and 0.05 (dashed) also identified by the other method at different FDR. (B-D) Example of regions found under divergent selection by `Flink` (B), `BayeScan` (C) or both (D). The dashed line represents the FDR threshold of 0.01.

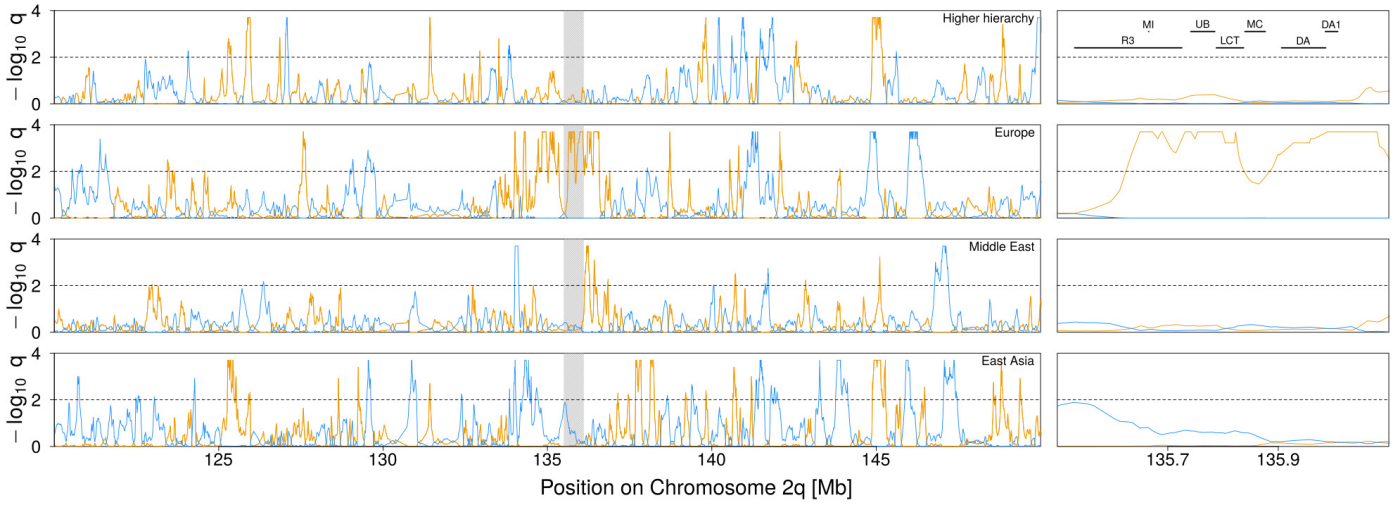

Figure 6 Signal of selection around the *LCT* gene on Chromosome 2q. The orange and blue lines indicate the locus-specific FDR for divergent (orange) and balancing (blue) selection, respectively. The black dashed line shows the 1% FDR threshold. A zoom of the highlighted region is shown on the right 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.

for selected regions among Europeans, the majority of regions identified by `BayeScan` were replicated by `Flink` at small FDR thresholds. In contrast, most of the regions identified by `Flink` were not replicated by `BayeScan`, in line with a higher statistical power for the former. Visual inspection indeed revealed that for most regions identified by `Flink` but not `BayeScan`, 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 `BayeScan` 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 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

Since positive selection might affect a subset of populations only and hence lead to an increase in population differentiation (Nielsen 2005), we 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 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 which has acted to increase lactase persistence in several human populations, including Europeans (Nielsen *et al.* 2007). Lactase persistence varies among Europeans and decreases on a roughly north-south cline (Bersaglieri *et al.* 2004; Leonardi *et al.* 2012; Burger *et al.* 2007; Itan *et al.* 2009), 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 contrasts, we detected no such signal for the other groups.

Discussion

Genome scans are common methods to identifying 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 auto-correlation among hierarchical parameters, rather than the population allele frequencies or haplotypes themselves. In the context of genome scans, this has been previously successfully by classifying each locus as selected or neutral (Boitard *et al.* 2009; Kern and Haussler 2010). Here, we extend this idea by modeling auto-correlation among the strength of selection acting at individual loci. While ignoring auto-correlation at the genetic level certainly leads to a loss of information, the resulting method remains computationally tractable. And as we show 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.

Accounting for partial linkage particularly improved the power to identify 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 Foll and Gaggiotti (2008); Beaumont and Balding (2004), 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 in unlikely the main driver, and caution against over-interpreting these results. But we note that the empirical false discovery rate for loci under balancing selection was extremely low in our simulations.

An obvious draw-back 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 selections coefficients can be obtained for each locus individually to identify the locus with the strongest signal, for which one might then also use complementary methods.

We finally 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 be 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

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.

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.

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

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

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. The American Journal of Human Genetics **74**: 1111–1120.

Boitard, S., C. Schlötterer, and A. Futschik, 2009 Detecting selective sweeps: A new approach based on hidden Markov models. Genetics **181**: 1567–1578.

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.

Burger, J., M. Kirchner, B. Bramanti, W. Haak, and M. G. Thomas, 2007 Absence of the lactase-persistence-associated allele in early Neolithic Europeans. Proceedings of the National Academy of Sciences **104**: 3736–3741.

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**: 1–16.

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.

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.

Eriksson, A. and A. Manica, 2012 Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins **109**: 13956–13960.

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

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.

Feder, J. L., S. P. Egan, and P. Nosil, 2012 The genomics of speciation-with-gene-flow. Trends in Genetics **28**: 342–350.

Ferrer-Admetlla, A., C. Leuenberger, J. D. Jensen, and D. Wegmann, 2016 An Approximate Markov Model for the Wright-Fisher Diffusion. Genetics **203**: 831–846.

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.

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. American Journal of Human Genetics **95**: 394–407.

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.

Gaggiotti, O. E. and M. Foll, 2010 Quantifying population structure using the F-model. Molecular Ecology Resources **10**: 821–830.

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**: 703–13.

Itan, Y., A. Powell, M. A. Beaumont, J. Burger, and M. G. Thomas, 2009 The origins of lactase persistence in Europe. PLoS Computational Biology **5**: 17–19.

James Kent, W., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, *et al.*, 2002 The human genome browser at UCSC. Genome Research **12**: 996–1006.

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.

Kern, A. D. and D. Haussler, 2010 A population genetic hidden markov model for detecting genomic regions under selection. Molecular Biology and Evolution **27**: 1673–1685.

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. International Dairy Journal **22**: 88–97.

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.

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

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

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

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

Peter, B. M., D. Petkova, and J. Novembre, 2017 Genetic landscapes reveal how human genetic diversity aligns with geography. bioRxiv pp. 1–24.

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

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**: 0660–0671.

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

Slatkin, M. and L. Voelm, 1991 FST in a hierarchical island model. Genetics **127**: 627–9.

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.

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. Nature Communications **9**: 1–14.

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