Received: 24 October 2019 | Revised: 22 June 2020 | Accepted: 29 June 2020

DOI: 10.1111/1755-0998.13221

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

# Genome scans for selection and introgression based on *k*-nearest neighbour techniques

Bastian Pfeifer[1] | Nikolaos Alachiotis[2] | Pavlos Pavlidis[3] | Michael G. Schimek[1]

[1]Research Unit of Statistical Bioinformatics, Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Graz, Austria

[2]Faculty of EEMCS, University of Twente, Enschede, The Netherlands

[3]Institute of Computer Science, Foundation for Research and Technology-Hellas, Crete, Greece

**Correspondence**
Bastian Pfeifer, Research Unit of Statistical Bioinformatics, Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Graz, Austria.
Email: bastian.pfeifer@medunigraz.at

## Abstract

In recent years, genome-scan methods have been extensively used to detect local signatures of selection and introgression. Most of these methods are either designed for one or the other case, which may impair the study of combined cases. Here, we introduce a series of versatile genome-scan methods applicable for both cases, the detection of selection and introgression. The proposed approaches are based on nonparametric *k*-nearest neighbour (kNN) techniques, while incorporating pairwise Fixation Index ($F_{ST}$) and pairwise nucleotide differences ($d_{xy}$) as features. We benchmark our methods using a wide range of simulation scenarios, with varying parameters, such as recombination rates, population background histories, selection strengths, the proportion of introgression and the time of gene flow. We find that kNN-based methods perform remarkably well compared with the *state-of-the-art*. Finally, we demonstrate how to perform kNN-based genome scans on real-world genomic data using the population genomics R-package POPGENOME.

KEYWORDS

adaptation, genome scans, introgression, *k*-nearest neighbours

## 1 | INTRODUCTION

Recent advances in DNA sequencing technology, which reduced costs and increased throughput and accuracy, have driven subsequent advances in population genomics methods for detecting traces of natural selection in DNA fragments. In a recombining chromosome, a subgenomic region under natural selection typically exhibits different levels of variation and differentiation than the rest of the genome (Li & Ralph, 2019). Hence, it can be considered an anomaly that deviates from the overall population structure (François, Martins, Caye, & Schoville, 2016; Haasl & Payseur, 2016). Identifying such anomalies in molecular data is of great significance since the respective footprints of localized natural selection can provide insight into the adaptation process of a population to its environment through different generations.

One of the most frequently used statistics to detect genomic regions under selection is the Fixation Index ($F_{ST}$), which was introduced to quantify population differentiation based on the Wright–Fisher model (Wright, 1949). Several $F_{ST}$ variants are widely employed in population genomics (Hudson, Slatkin, & Maddison, 1992; Weir & Cockerham, 1984; Weir & Ott, 1997) because high $F_{ST}$ values can be an indication of local adaptation. However, when the population history deviates from the Wright–Fisher model, or when evolutionary history is described by a hierarchical population structure model, hypothesis testing becomes a challenge because the $F_{ST}$ distribution that accounts for the neutral demographic model of the population under study is not known. In this case, $F_{ST}$-based methods that do not take the genome-wide population structure into account may yield unreliable results, as mentioned in several previous studies (Bonhomme et al., 2010; Excoffier, Hofer, & Foll, 2009; Foll & Gaggiotti, 2008; Lotterhos & Whitlock, 2014; de Villemereuil, Frichot, Bazin, François, & Gaggiotti, 2014).

To better account for the genome-wide population structure, $F_{ST}$ was introduced as a model parameter in the Bayesian approaches and inferred via computationally intensive Markov Chain Monte Carlo (MCMC) simulations. In such approaches, a common migrant pool is modelled as a Dirichlet distribution, and the genome-wide neutral signal is captured in a logistic regression model with a specific parameter shared by all populations. One of the most prominent methods is implemented in the BAYESCAN software (Foll & Gaggiotti, 2008), which is built upon the works of Beaumont and Nichols (1996) and Beaumont and Balding (2004). It has been reported, however, that these methods still suffer from a high false discovery rate (FDR; De Villemereuil & Gaggiotti, 2015; Duforet-Frebourg, Bazin, & Blum, 2014; Duforet-Frebourg, Luu, Laval, Bazin, & Blum, 2015). Alternative approaches were developed to address some of these shortcomings. The method implemented in BAYENV2 (Günther & Coop, 2013), for instance, uses correlations to environmental variables to improve the detection of local adaptation. An extension of this method is introduced in BAYPASS (Gautier, 2015), which includes several modifications of the underlying Bayesian model assumptions.

The majority of the aforementioned methods consider population differentiation based on allele frequencies. Methods that detect adaptive evolution using haplotype frequencies are implemented in the *HAPF$_{LK}$* software (Fariello, Boitard, Naya, SanCristobal, & Servin, 2013). The employed models in *HAPF$_{LK}$* account for linkage disequilibrium (LD), while their parameters are inferred via an expectation–maximization (EM) algorithm. More recent and less computationally demanding approaches rely on principal component analyses (PCAs; Duforet-Frebourg et al., 2015; Luu, Bazin, & Blum, 2017). The PCADAPT software, for instance, does not assign individual samples to populations since the overall population structure is captured by the principal components. Pfeifer and Lercher (2018) recently presented BLOCKFEST, a Bayesian approach that enables the grouping of individual SNPs (single nucleotide polymorphisms) into blocks. This approach facilitates the computation of region-wide $F_{ST}$, which can be used to detect signatures of local adaptation that span a genomic region with multiple SNPs. A composite approach is proposed by Lotterhos et al. (2017). The authors combine the outcome of well-established tests for selection by mapping the corresponding scores into a multidimensional space and apply a series of diverse multivariate distance metrics for enhancing a composite selection measure.

Localized deviations from the genome-wide population structure may also be introduced by other evolutionary forces, such as introgressive hybridization. During introgressive hybridization, species exchange genetic information, a process that may serve as the primary source for the adaptation of populations to their environment (Hedrick, 2013). Detecting introgressed regions has great significance since it can shed light on the effects of hybridization among species. Several approaches have been proposed for this purpose. The most widely applied one is the ABBA-BABA family of methods, which is based on a four-taxon system where the fourth taxon serves as the outgroup (Durand, Patterson, Reich, & Slatkin, 2011; Green et al., 2010; Martin, Davey, & Jiggins, 2014; Pfeifer & Kapan, 2019). Since an outgroup is not always available, approaches based on fewer taxa have also been introduced (Geneva, Muirhead, Kingan, & Garrigan, 2015; Hahn & Hibbins, 2019; Hibbins & Hahn, 2019). Hahn and Hibbins (2019), for instance, recently presented the $D_3$ statistic, which relies on pairwise $d_{xy}$ measurements.

Both selection and introgression introduce patterns that can be considered anomalies with respect to the genome-wide population structure. Various machine-learning (ML) methods are particularly designed for detecting anomalies. In this work, we deploy the pairwise $F_{ST}$ and $d_{xy}$ measures as features in kNN-based (*k*-nearest neighbours) ML approaches, and assess their power in detecting selection and introgression in whole-genome data. The ML family of kNN methods are among the most prominent unsupervised techniques and are already widely applied in several areas of data-driven research to detect anomalies. The underlying idea of our approach is influenced by the population branch length statistics (PBS) method (Yi et al., 2010), but supports an arbitrary number of branches. We also employ pairwise $F_{ST}$, which is less prone to confounding factors such as background population histories than the global $F_{ST}$ that is calculated by the mean divergences between multiple populations. Since $F_{ST}$ estimates may lead to false positives when diversities within populations are low, which can potentially mislead the detection of introgression (Cruickshank & Hahn, 2014), we additionally employ the pairwise nucleotide differences ($d_{xy}$) as a feature for detecting introgression.

We perform a wide range of coalescent simulations to evaluate the ability of the kNN-based approaches to detect selection and introgression under different evolutionary scenarios, such as population bottlenecks, recombination rates, population background histories, selection strengths, the proportion of introgression and the time of gene flow. Finally, we showcase the use of the kNN approaches to detect positively selected regions in the human genome by analysing the data made available by the 1,000 Genomes Project (1000 Genomes Project Consortium & others, 2015). We find that the kNN-based methods are highly suited for the detection of local adaptation acting on a region comprising multiple SNPs. In that case, we observe a substantial gain of accuracy compared with the *state-of-the-art*, as measured by the trade-off between precision and recall. We also observe that the kNN-based methods perform well for the detection of introgressed genomic regions, making our approach a promising tool for the detection of adaptive introgression as well.

## 2 | MATERIALS AND METHODS

### 2.1 | Using $F_{ST}$ as features in kNN methods

The principal idea of the kNN approach (Cover & Hart, 1967; Fix, 1951) is to classify distant data points with respect to their neighbourhood as outliers. This is achieved by relying on distance metrics such as the Euclidean distance. Computed are all distances from any given point, where *k* is a prespecified positive integer. In this work, we use population pairwise $F_{ST}$ to define the location of the data points in a multidimensional space. $F_{ST}$ between two populations is calculated by

$$F_{ST} = 1 - \frac{H_w}{H_b}, \qquad (1)$$

where $H_w$ is the mean of nucleotide differences *within* two populations and $H_b$ is the mean of nucleotide differences *between* two populations. In case of multiple SNPs, we calculate the ratio of average nucleotide differences within and between two populations, as suggested by Hudson et al. (1992) and Bhatia, Patterson, Sankararaman, and Price (2013). In our framework, the population pairwise $F_{ST}$ estimates represent a genomic region as a data point that is embedded into an *m*-dimensional numerical space, where *m* is the total number of possible population pairwise comparisons $(m = n_p(n_p - 1)/2)$, and $n_p$ is the total number of populations being analysed. Thus, the population structure of each genomic region is represented by an $F_{ST}$ vector of length *m*. A kNN outlier score for a genomic region *x* is calculated according to Equation 2

$$\mathrm{kNN}_k(x) = \frac{\sum_{o \in N_k(x)} d_k(x, o)}{|N_k(x)|}, \qquad (2)$$

where $|N_k(x)|$ is the total number of *k* genomic regions in the nearest neighbour set, while $d_k(x, o)$ quantifies the dissimilarity between genomic regions *x* and *o*. We employ the Euclidean distance between feature vectors $F_{ST}(x)$ and $F_{ST}(o)$ as a dissimilarity measure, which is computed using Equation 3,

$$d_k(x, o) = \sqrt{\sum_{i=1}^{m} (F_{ST}(x)_i - F_{ST}(o)_i)^2}. \qquad (3)$$

Note that any feature, for example pairwise $d_{xy}$, could have easily been used to compute the required $d_k(x, o)$ distances. Building upon the basic kNN approach, Angiulli and Pizzuti (2002, 2005) proposed a weighted-kNN approach that considers the overall distance between a data point and its neighbours by computing the sum of distances instead of the arithmetic mean. Hautamaki, Karkkainen, and Franti (2004) introduced ODIN (outlier detection using indegree number), a method that infers outliers based on a kNN graph. To the best of the authors' knowledge, the most widely used method for quantifying local outlierness is LOF (local outlier factor; Breunig, Kriegel, Ng, & Sander, 2000), which is based on the concept of *local reachability density* (lrd) of the *k*-nearest neighbours. In LOF, a data point is considered to be an outlier when its density is much smaller than the densities of its neighbours. The lrd is defined as

$$\mathrm{lrd}_k(x) = \frac{1}{\mathrm{kNN}_k(x)}, \qquad (4)$$

while LOF is calculated using Equation 5

$$\mathrm{LOF}_k(x) = \frac{1}{|N_k(x)|} \sum_{o \in N_k(x)} \frac{\mathrm{lrd}_k(o)}{\mathrm{lrd}_k(x)}. \qquad (5)$$

Various modifications to LOF and the corresponding concept of lrd have been proposed over the years. Schubert, Zimek, and Kriegel (2014) described the simplified LOF, which employs the basic kNN distances instead of local reachability distances. Tang, Chen, Fu, and Cheung (2002) introduced the connectivity-based outlier factor (COF), which adapts the density estimation of simplified LOF to account for the 'connectedness' of the neighbourhood via a minimum spanning tree (MST). Kriegel, Kröger, Schubert, and Zimek (2009) implemented LoOP (local outlier probabilities), a tool that adopts normalized local density scores based on the quadratic mean, leading to scores that are strictly within the [0,1] interval. Zhang, Hutter, and Jin (2009) introduced the local distance-based outlier factor (LDOF), which quantifies the amount of data points that deviate from their scattered neighbourhood based on the relative distance between a data point and its neighbours. The ABOD (angle-based outlier detection) approach by Kriegel, Schubert, and Zimek (2008) addresses the so-called '*curse of dimensionality*' problem by comparing the angles between pairs of distance vectors. Latecki, Lazarevic, and Pokrajac (2007) presented the local density factor (LDF) that replaces the LOF density estimation by variable-width Gaussian kernel density estimation (KDE). INFLO (influenced outlierness) by Jin et al. (2006) also takes into account the reverse nearest neighbourhood set when calculating local density scores.

The provided list of methods of anomaly detection algorithms is by no means comprehensive. Presenting a complete list is beyond the scope of the current article. In this work, however, we employ the majority of the aforementioned kNN approaches, which were also previously studied in Campos et al. (2016). In their study, they analysed the performance of the aforementioned kNN approaches on biomedical data sets available from the UCI repository (Bache & Lichman, 2013). They found that kNN (Cover & Hart, 1967), the weighted kNN (Angiulli & Pizzuti, 2005) and the original LOF (Breunig et al., 2000) are the *state-of-the-art* of outlier detection algorithms.

## 2.2 | Setting parameter *k*

In supervised classification problems, whereby the labels of the classes are known a priori, parameter *k* is typically inferred via cross-validation. It is well known, however, that choosing an appropriate *k* value in a purely unsupervised setting is a challenging task, since it highly depends on the data being analysed. This challenge especially arises in studies that aim to detect local outliers. In the data mining literature, local outliers are discussed as data points whose immediate neighbours are closer together compared with the rest. Our approach in this study deploys kNN-based methods with the aim to detect global outliers to distinguish between signals of neutral evolving genomic regions and outlier regions subject to local selection or introgression. Therefore, the outcome of our analyses is expected to be robust to the choice of $k$.

Nevertheless, any chosen $k$ value should belong to what we refer to as a stable $k$ region, that is a range of $k$ values that yield highly correlated kNN scores. To address this requirement, we propose the following approach. First, calculate the kNN scores $\mathbf{s}_i$ for $n_k$ = 100 sequentially sampled values of $k$ from [2, $n_r$ − 1], where $n_r$ is the total number of genomic regions

$$\mathbf{s}_i = \mathrm{kNN}_{k(i)}(X) \quad \forall i = 1, \ldots, n_k, X \in \{x_1, \ldots, x_{n_r}\}. \tag{6}$$

Second, calculate Kendall's tau correlation coefficients ($\tau$)

$$\mathrm{corr}_i = \mathrm{median}\left[\tau(\mathbf{s}_{i-1}, \mathbf{s}_i), \tau(\mathbf{s}_i, \mathbf{s}_{i+1}), \tau(\mathbf{s}_{i-1}, \mathbf{s}_{i+1})\right] \quad \forall i = 2, \ldots, n_k - 1, \tag{7}$$

with

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{n_r(n_r - 1)/2}, \tag{8}$$

where *concordant/discordant* refers to the elements between two vectors $\mathbf{s}_i$ and $\mathbf{s}_j$ which agree/disagree on their ranks. Third, from the correlation vector corr infer the longest connected $k$ region with corr > 0.90, and define the median of that region as the optimal $k$. We also provide a diagnostic plot, available from our GitHub repository, for a manual inspection of the correlation vector corr.

Thereafter, the final outlier scores per genomic region are computed based on the chosen and henceforth fixed $k$ value. We are now able to detect genomic regions comprising anomalous population structure, yet it remains unclear which population or population pair causes the outlierness. To this end, we propose the following practical approach. Once the outlier regions are detected, based on kNN outlier scores, we calculate the *medoid* of the pairwise $F_{ST}$ vectors. We have empirically determined that the *medoid* is the most informative data point for our purpose because it reflects the overall population structure. Subtracting the *medoid* from the outlier pairwise $F_{ST}$ vectors reveals which population or population pairs are affected by selection or introgression. We refer to the resulting vectors as the $\Delta F_{ST}$ selection effects. Positive $\Delta F_{ST}$ entries suggest local adaptation, whereas negative values point to introgression (reduced divergence due to gene flow) or other types of selection that significantly reduce the divergence between populations, for example balancing selection.

## 2.3 | Experimental set-up

We conducted simulations under local adaptation and introgression using the MSMS software (Ewing & Hermisson, 2010). For the case of local adaptation, we generated 950 neutral regions and 50 regions under selection, resulting in a total number of $n_r$ = 1,000 regions. We have built three populations, each comprising 100 samples. The number of SNPs per region was 50, and the effective population size was $N_e$ = 10,000. In our baseline model, the first coalescent event of populations $P_1$ and $P_2$ was fixed at $t_{12} = 0.1 \times 4N_e$, and the second coalescent event was set to $t_{123} = 0.9 \times 4N_e$ generations ago. The selection strength for homozygotes was $s = 0.1 = 2{,}000/2N_e$, where selection started at $t_s = 0.1 \times 4N_e$ generations ago in population $P_1$. The recombination rate was $r = 0.01$. We varied recombination rates and the time of coalescence with population $P_3$ ($t_{123}$). We also considered differences in selection strength and the time of selection.

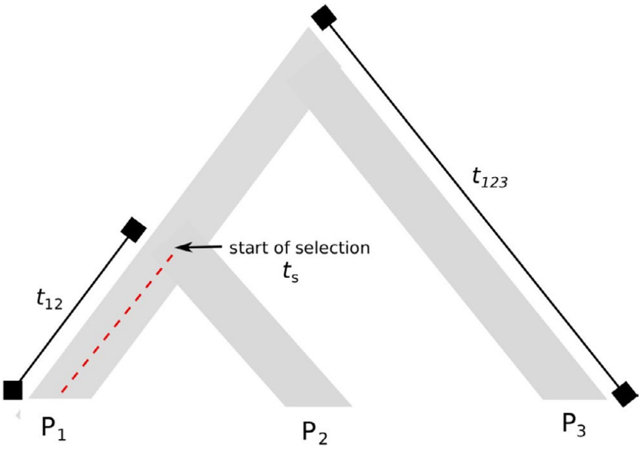

**FIGURE 1** A graphical illustration of local adaptation. A three population genealogy with selection introduced at $t_s \times 4N_e$ generations ago in population $P_1$ [Colour figure can be viewed at wileyonlinelibrary.com]

We benchmarked the proposed methods for their ability to distinguish between local adaptation and genetic footprints left by bottleneck scenarios. A simplified description of our selection model is illustrated in Figure 1. The main calls to the MSMS program can be found in the Supporting Information Section 1.1.

To generate introgression events, we again made use of the MSMS software with the same background population history as in the selection case. We simulated 950 neutral regions and 50 regions under introgression with three populations including 100 samples each. In our baseline model, the coalescent times were $t_{12} = 0.1 \times 4N_e$ and $t_{123} = 0.9 \times 4N_e$ generations ago. The scaled mutation rate was set to $\theta$ = 1. We introduced $P_3 \rightarrow P_2$ introgression $t_{GF} = 0.01 \times 4N_e$ generations ago with a fraction of introgression $f$ = 0.3. The recombination rate was fixed to $r$ = 0.01 in all simulations. We varied the proportion and the time of introgression. A schematic representation of the introgression model is illustrated in Figure 2.

For the selection cases, we compared our kNN approaches with the methods implemented in the R-package PCADAPT (Luu et al., 2017), BLOCKFEST (Pfeifer & Lercher, 2018) and *HAPF*$_{LK}$ (Fariello et al., 2013). To ensure a fair comparison with PCADAPT and $F_{LK}$, when conducting region-specific selection, we simulated the same number of SNPs for all regions and computed the sum of *log-p-values* in order to assign a region-specific score. Keeping the number of SNPs constant ensured that the results are not biased by the size of a particular region. The number of principal components was set to $K$ = 2 for PCADAPT, and the number of clusters was set to $K$ = 1 for the $F_{LK}$ method.

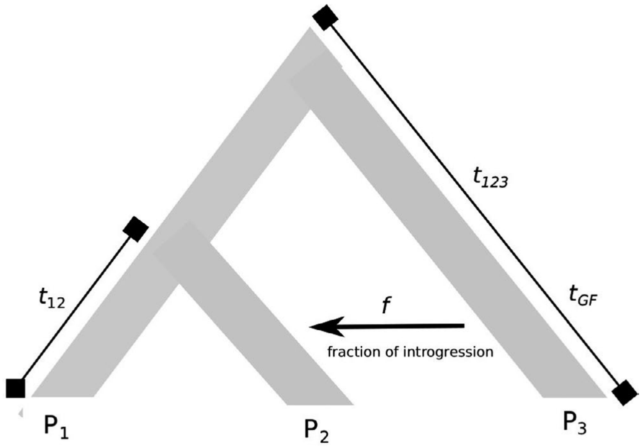

**FIGURE 2** A graphical illustration of introgression. A three population species tree with an unidirectional introgression event from the ancestral population $P_3$ to population $P_2$ introduced $t_{GF} \times 4N_e$ generations ago. The proportion of introgression is indicated by $f$

In the introgression cases, we compare the kNN-based methods to the $D_3$ approach (Hahn & Hibbins, 2019), which follows a similar logic as the $d_f$ method introduced by Pfeifer & Kapan (2019), but without the use of an outgroup. We also show the absolute value of $D_3$, denoted as $|D_3|$. We believe that the absolute value is a more appropriate value to compare with our approaches, because outlier detection methods typically generate positive numbers. The $D_3$ method is compared to the kNN-based techniques using pairwise $d_{xy}$ and pairwise $F_{ST}$ as features. Finally, we relate all of these methods to the global $F_{ST}$ estimate by Hudson et al. (1992) as a baseline approach. Accuracy is measured by the area under the curve (AUC), as implemented in the R-package PROC (Robin et al., 2011), and the precision–recall area under the curve (PR-AUC), as implemented in the R-package PRROC (Grau, Grosse, & Keilwagen, 2015).

# 3 | RESULTS

## 3.1 | On the power to detect selection

We assess the power of the kNN approaches to detect single-locus signatures of selection, where each region comprises a single SNP. When selection is strong, we observe that both the kNN methods and the $F_{LK}$ method yield comparable results, whereas $F_{ST}$ and PCADAPT achieve overall lower accuracy (Figure S1a,b). When selection strength is low, however, the kNN approaches become unstable (Figure S1c,d) and competing methods such as $F_{LK}$ are found to be more accurate (see also Figure S2).

Simulations under region-specific selection indicate that the kNN-based methods remain largely unaffected by the choice of $k$ (Figure 3), outperforming well-established methods such as PCADAPT and $F_{LK}$. Unstable results are only observed for either very small or very high $k$ values with respect to the total number of genomic regions analysed. As expected, the $F_{ST}$ results are fully comparable for star-like genealogies (Figure 3a). However, as soon as hierarchical structure is introduced to the population history, our proposed competing methods show overall higher AUC and PR-AUC values (Figure 4). The methods PCADAPT, BLOCKFEST and the global $F_{ST}$ are most affected when varying the coalescent times to population $P_3$, whereas the kNN-based techniques and $F_{LK}$ remain almost unaffected (Figure 4a,b).

The weighted-kNN and simplified-LOF methods are the strongest kNN-based methods, both outperforming $F_{ST}$, PCADAPT and $F_{LK}$, and are comparable to BLOCKFEST (Figure 3). However, the PR-AUC values in Figure 4b,d show that BLOCKFEST has overall lower precision–recall values indicating a substantially higher false discovery rate. Also, BLOCKFEST is based on computationally intensive MCMC runs and for that reason might not be generally applicable. Overall, the performance of all methods under consideration decreases with increasing recombination rates (Figure 4c,d). This is expected because the signal of selection gets eroded, which makes it harder to detect these patterns. Notably, $F_{LK}$ performs slightly better compared with the competing methods when the recombination rate is high.

Based on the simulations with varying selection strengths and time of selection (Figure S3), we observe that the power to detect selection drops when the selection strength is $s = 0.01$. Overall, $F_{LK}$ competes well with the kNN approaches with only slight lower AUC and PR-AUC values. We also simulated bottleneck scenarios in population $P_1$ occurring $0.05 \times 4N_e$ generations ago. The weighted kNN and the simplified LOF outperform the other methods in the case of moderate bottlenecks (Figures S5 and S6). Increasing the bottleneck strength, however, leads to overall lower performance for all methods. The PR-AUC values of $F_{LK}$ are higher than those obtained by BLOCKFEST (see Figure S6).

Based on our simulations, we observed that the INFLO algorithm is the weakest kNN-based method for the detection of selection. As a matter of fact, INFLO is the most sensitive to background population histories as can be seen in Figure 4a.

## 3.2 | On the power to detect introgression

Simulations under unidirectional introgression from population $P_3$ to an in-group population $P_2$ confirm that the kNN-based family of methods is not greatly affected by the choice of $k$ (Figure 5). We observe overall low accuracy for $F_{ST}$ when varying the time of gene flow and the fraction of introgression (Figures 5 and 6). This is especially true when gene flow is recent and the fraction of introgression is high (Figure 5a,b). $D_3$ is more powerful when varying the time of gene flow (Figure 6c,d) and outperforms the kNN-based methods based on pairwise $F_{ST}$ features. When the gene flow is recent, $D_3$ has overall lower AUC values than the kNN-based approaches (Figure 6a), whereas $D_3$ and $|D_3|$ outperform the kNN techniques when the gene flow occurred further in the past ($t_{GF} > 0.1$). This stability against the time of gene flow was also previously mentioned by Pfeifer and Kapan (2019). These authors used a similar concept to $D_3$, which was realized on a four-taxon system.

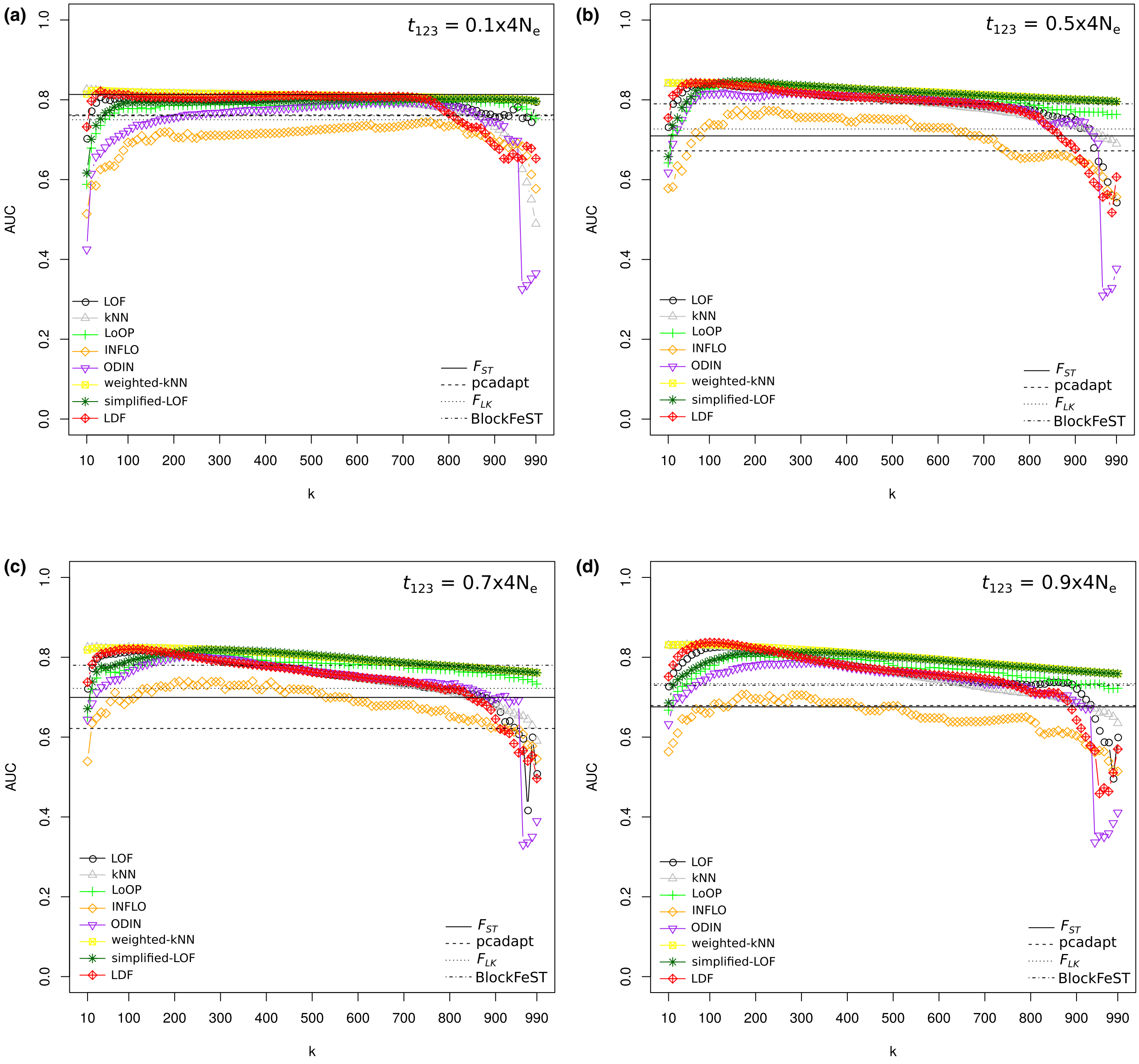

**FIGURE 3** Local adaptation: varying the coalescent time to population $P_3$ ($t_{123}$). The result for the kNN-based methods using pairwise $F_{ST}$ as features, for 100 sequentially sampled $k$ values and in comparison with the accuracy of $F_{ST}$, PCADAPT, $F_{LK}$ and BLOCKFEST. The recombination rate is $r = 0.001$, and the number of SNPs per region is 50. (a) The simulations are based on a star formed genealogie ($t_{12} = 0.1 \times 4N_e = t_{123}$). The coalescent time to population $P_3$ is (b) $t_{123} = 0.5 \times 4N_e$, (c) $t_{123} = 0.7 \times 4N_e$ and (d) $t_{123} = 0.9 \times 4N_e$ generations ago. The expected AUC value of a random classifier is AUC = 0.5

Increasing the proportion of introgression has a similar effect on all kNN-based methods: the accuracy increases and almost reaches 100% when $f$ = 0.7. INFLO again is the weakest kNN-based method (Figures 5 and 6). Overall, the weighted-kNN method and the simplified LOF show high accuracy also in the introgression cases and provide stable score rankings almost across the full range of $k$.

Using $d_{xy}$ as feature also provides stable results for almost all choices of $k$ (Figure S7). However, in this situation results are not as reliable as those of kNN techniques with incorporated pairwise $F_{ST}$ estimates. This is especially true when the time of gene flow is recent (Figure S7a). Using a combined feature vector of $F_{ST}$ and $d_{xy}$ has no benefit over applying pairwise $F_{ST}$ vectors only (Figure S8). We learned that when the divergence between population $P_1$ and $P_2$ is excluded from the feature set qualitatively superior results (see Figure S9) are obtained.

We also test the ability of the proposed methods to detect introgressed genomic regions when varying the signal-to-noise ratio, specified by the number of introgressed regions divided by the number of neutral regions (Figure S10). We set the number of introgressed regions to a constant size of 50 and successively decrease the number of neutral regions. Our simulations indicate that the kNN

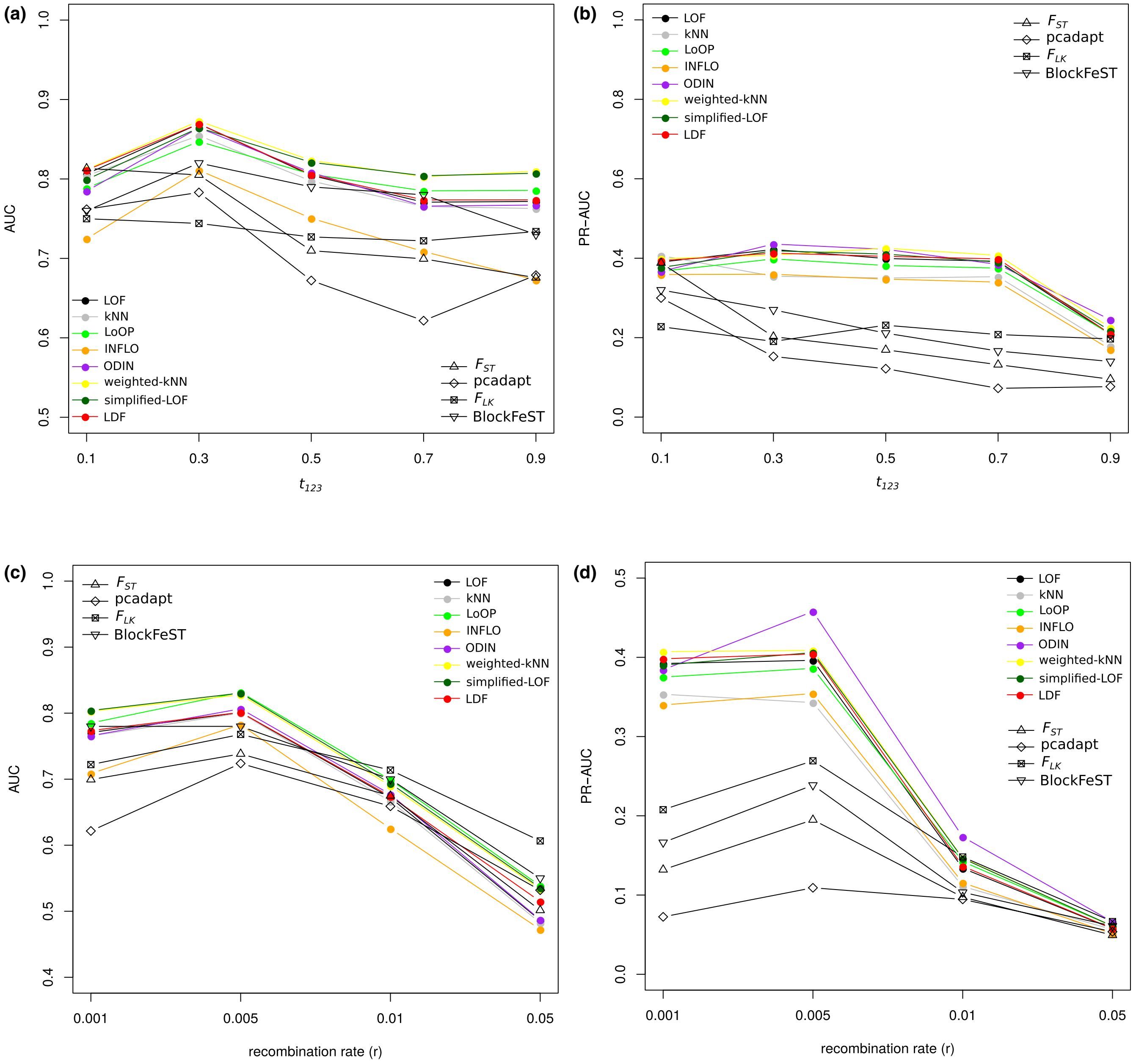

**FIGURE 4** Detecting selection with a computed *k*. The kNN methods with pairwise $F_{ST}$ as features, compared with $F_{ST}$, PCADAPT, $F_{LK}$ and BLOCKFEST. The recombination rate is $r = 0.001$, and the number of SNPs per region is 50. (a and b). Varying the coalescent time with population $P_3$ ($t_{123} = [0.1, 0.3, 0.5, 0.7, 0.9] \times 4N_e$ generations ago). The realized mean $F_{ST}$ over all regions is $F_{ST} = [0.17, 0.31, 0.42, 0.50, 0.55]$. (c and d). Varying the recombination rate ($r = [0.001, 0.005, 0.01, 0.05]$). The coalescent time with population $P_3$ is $t_{123} = 0.7 \times 4N_e$ generations ago. The realized mean $F_{ST}$ over all regions is $F_{ST} = [0.31, 0.32, 0.32, 0.31]$. The expected value of a random classifier is AUC = 0.5 and PR-AUC = 50/1,000 = 0.05

approaches achieve high accuracy when the signal-to-noise ratio is under 0.2, and become unstable otherwise (Figure S10d).

## 3.3 | Application to the 1,000 Genomes data

We also analysed the 1,000 Genomes data (1000 Genomes Project Consortium & others, 2015) to demonstrate the efficacy of our proposed kNN-based approaches when processing real data. The employed data set is currently one of the largest publicly available data sets, both in terms of number of samples and number of SNPs, with 2,504 human samples from 26 populations, and 77,832,252 SNPs in the entire set of autosomes (phase 3). We applied all implemented kNN-based techniques on a per-autosome basis to the samples of the population with Northern and Western European ancestry (CEU), the population with East Asian ancestry (CHB) and the population with African ancestry (YRI), evaluating nonoverlapping sliding windows of size 100 kb, in a total number of $n_r$ = 2,400 windows (Figure 7). The window size of 100 kb approximately maps a human recombination rate of $r$ = 0.001, for which we observed good kNN

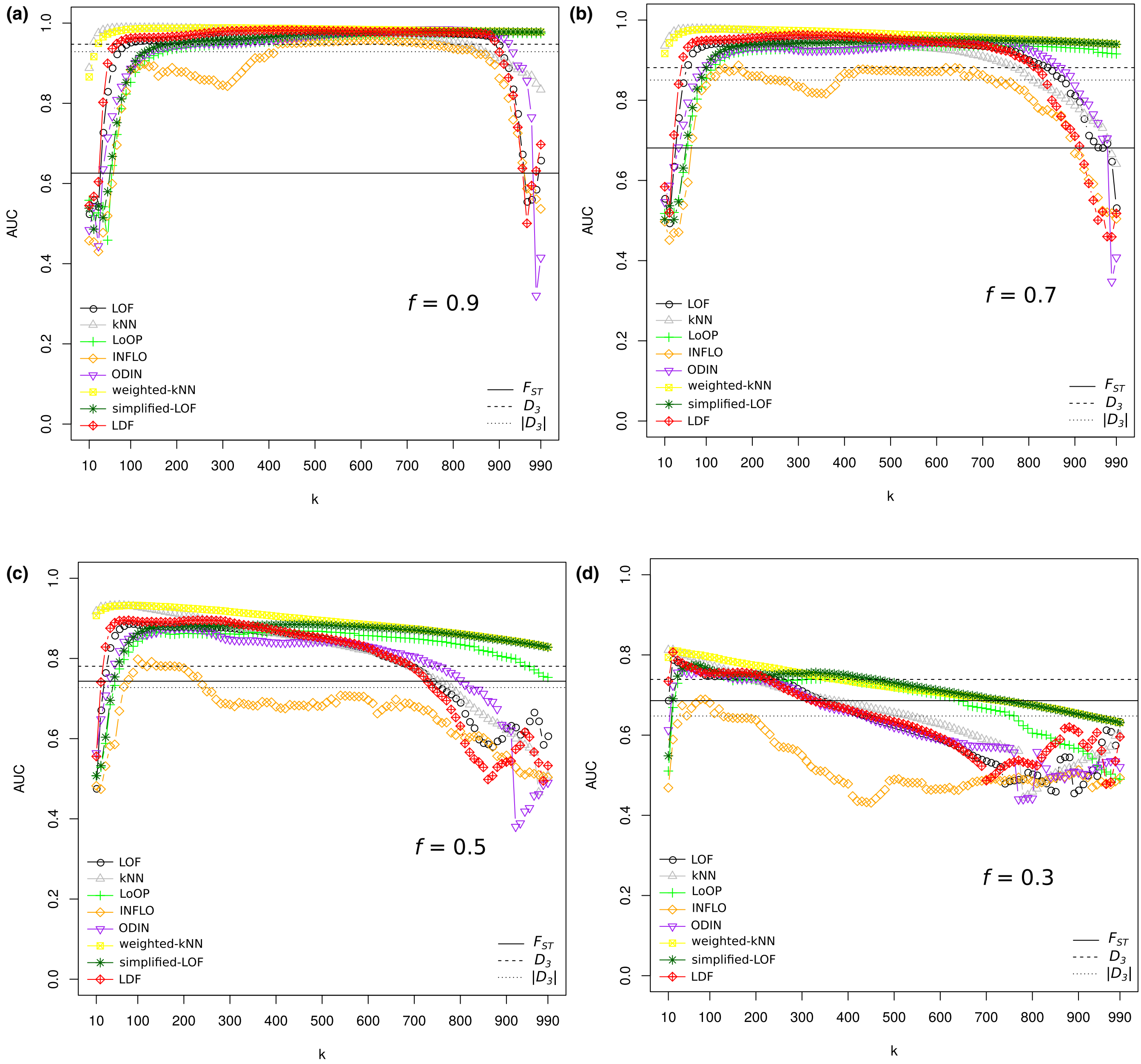

**FIGURE 5** Varying the fraction of introgression ($f$). The result for the kNN-based methods using pairwise $F_{ST}$ as features, for 100 sequentially sampled $k$ values. Coalescent times are $t_{12} = 0.1 \times 4N_e$ and $t_{123} = 0.9 \times 4N_e$ generations ago. Recent introgression is introduced $t_{GF} = 0.01 \times 4N_e$ generations ago, and the recombination rate is set to $r = 0.01$ in all simulations. The outcome of the kNN-based methods is compared to $F_{ST}$, $D_3$ and $|D_3|$. The fraction of introgression is (a) $f = 0.9$, (b) $f = 0.7$, (c) $f = 0.5$ and (d) $f = 0.3$. The expected AUC value of a random classifier is AUC = 0.5

results in the simulations (Figure 4d). As seen in the diagnostic plot in Figure 7i, the kNN scores are highly correlated for almost all sampled $k$ values.

Here, we summarize the results for chromosome 2, showing the genomic windows that are discovered as outliers by the kNN methods. We report the nearest genes to these outlier windows (Table 1) when the outlier window does not overlap with a gene. For each tool, we consider kNN scores within a conservative 0.005-quantile to define the outlier candidates. The top-2 candidate genes for adaptive evolution are the protein-coding genes EXOC6B and EDAR (Table 1 and Figure 7). Baye, Wilke, and Olivier (2009) name EXOC6B as a positively selected gene. Intellectual disability and developmental delay are associated with this gene. Our kNN approaches suggest directional selection between the YRI population and both the CEU and CHB populations (Table 1). Bryk et al. (2008) report EDAR, which is a gene involved in ectodermal development, increased in frequency in East Asia due to positive selection 10,000 years ago. The kNN-based approaches suggest the strongest effect between CEU and CHB (Table 1).

Another candidate gene is CNTNAP5 (outlier window: 126.1–126.2 Mb) and is confirmed by all tools but ODIN. The selection effect is $\Delta F_{ST}$ = [CEU/CHB = 0.39, CEU/YRI = 0.01, CHB/

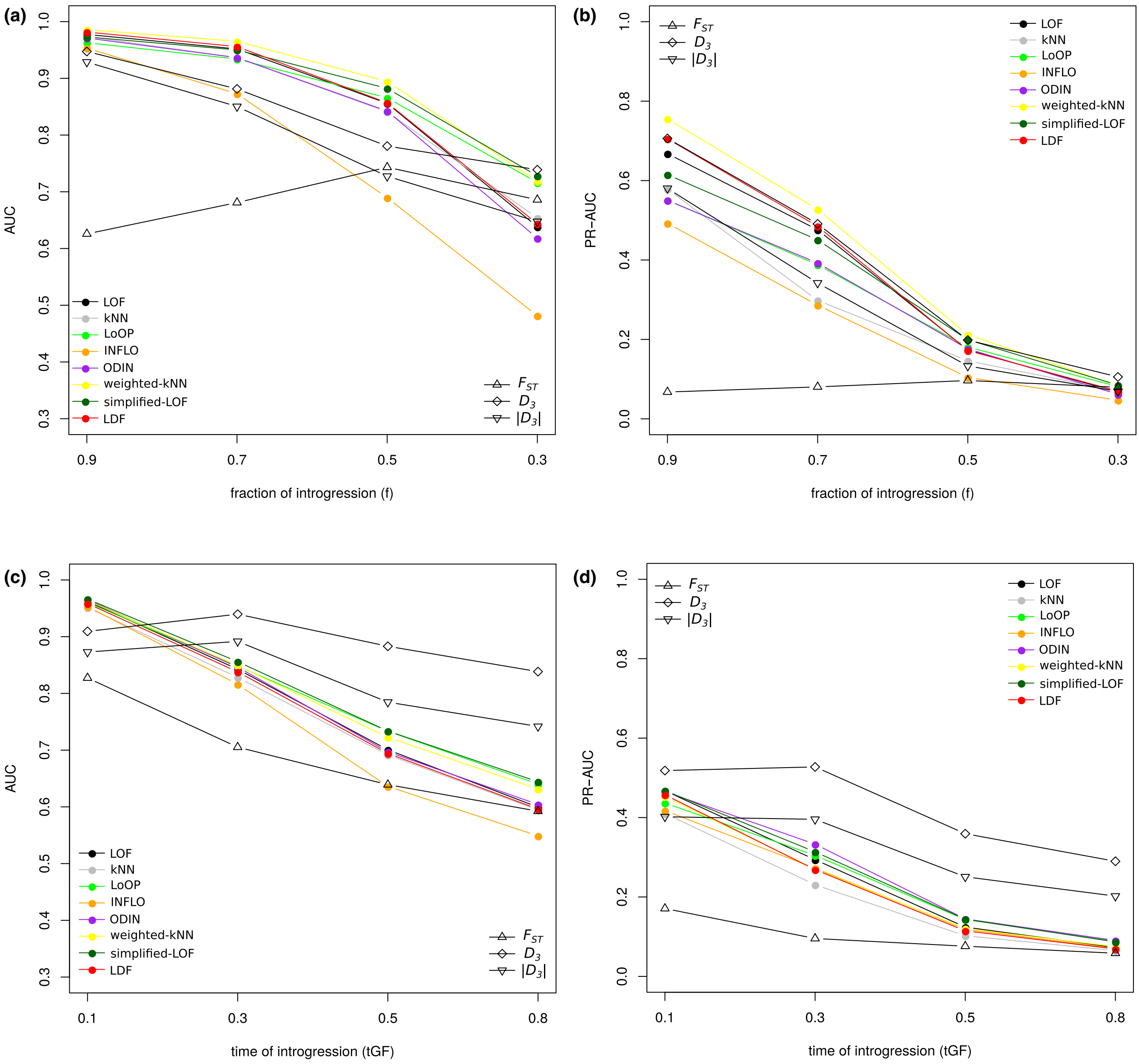

**FIGURE 6** Detecting introgression with a computed *k*. The accuracy of the kNN methods with pairwise $F_{ST}$ as features, compared with $F_{ST}$, $D_3$ and $|D_3|$. The recombination rate is $r = 0.01$ in all simulations. (a, b). Varying the fraction of introgression ($f=[0.9, 0.7, 0.5, 0.3]$). Coalescent times are $t_{12}=0.1\times 4N_e$ and $t_{123}=0.9\times 4N_e$ generations ago, and recent introgression is introduced $t_{GF}=0.01\times 4N_e$ generations ago. The realized mean $F_{ST}$ over all regions is $F_{ST}=[0.50, 0.50, 0.50, 0.50]$. (c, d). Varying the time of gene flow ($t_{GF}=[0.1, 0.3, 0.5, 0.8]\times 4N_e$) with an fixed fraction of introgression of $f = 0.7$. Coalescent times are $t_{12}=1\times 4N_e$ and $t_{123}=2\times 4N_e$ generations ago. The realized mean $F_{ST}$ over all regions is $F_{ST}=[0.75, 0.75, 0.75, 0.75]$. The expected value of a random classifier is AUC = 0.5 and PR-AUC = 50/1,000 = 0.05

YRI = 0.25] suggesting directional selection in the Asian (CHB) population. An additional candidate gene is FMNL2 (outlier window: 153.1–153.2 Mb) and is exclusively identified by the weighted-kNN and kNN algorithms with a selection effect of $\Delta F_{ST}$ = [CEU/CHB = 0.29, CEU/YRI = 0.11, CHB/YRI = 0.33]). The genomic region 104.7–104.8 Mb is identified by all tools but the weighted kNN and kNN. The nearest gene is LINC01127, and the selection effect is $\Delta F_{ST}$ = [CEU/CHB = 0.32, CEU/YRI = 0.21, CHB/YRI = −0.10]. The ANTXR1 gene (outlier window: 69.2–69.3 Mb) is only reported by the ODIN method as a candidate for selection with a selection effect of $\Delta F_{ST}$ = [CEU/CHB = 0.22, CEU/YRI = 0.31, CHB/YRI = −0.05]. This observation is slightly pointing to positive directional selection in the CEU population and a reduced diversity between the CHB and YRI populations.

## 4 | DISCUSSION

In this study, we have investigated the efficacy of kNN-based algorithms, that is a family of unsupervised machine-learning

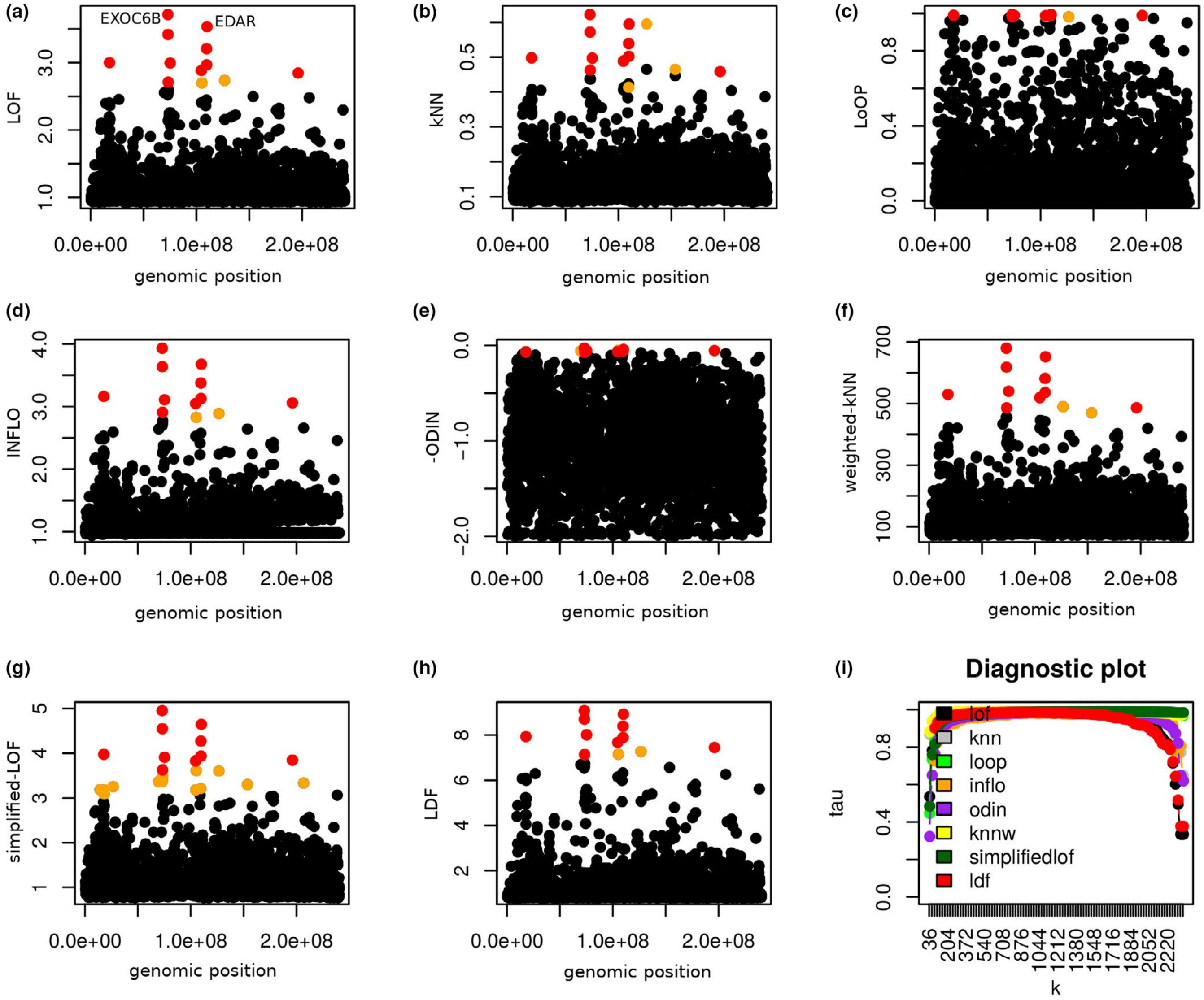

**FIGURE 7** Genome-scan plots of human chromosome 2. (a–h). The kNN scores are shown along human chromosome 2 based on 100-kb consecutive sliding windows. Red and orange dots are outliers identified by the kNN methods (0.005-quantile of the scores). Red dots indicate that all methods agree on these outliers and orange dots otherwise. (i). A diagnostic plot is shown with the pairwise rank correlations of the kNN scores while varying the parameter *k*

techniques, for the detection of local adaptation and introgression in whole-genome scans. Coalescent simulations under region-specific selection and introgression show that kNN-based methods that employ pairwise $F_{ST}$ as features perform remarkably well and are not greatly affected by the choice of *k*. Based on our simulations, we conclude that the weighted kNN is the most powerful approach for genome-wide scans to detect region-specific selection. The detection of selection on a single-SNP basis, however, is only recommended for strong selection and when our proposed diagnostic plot yields stable rank correlations of the kNN scores.

We found that the kNN approaches using pairwise $F_{ST}$ as features are also suited for the detection of recent introgression, but are not as robust as other methods with regard to the time of gene flow. However, the approaches presented in this work are highly flexible with regard to the choice of the feature set and researchers are not limited to using $F_{ST}$ or $d_{xy}$. A different set of features or a combination of selection/introgression sensitive features may further improve accuracy. In a feature experiment for the introgression cases, for instance, we observed that excluding the divergence between the in-group taxa ($P_1$ and $P_2$) results in a framework which is more robust to the time of gene flow.

In contrast to other genome-scan approaches, the kNN-based approaches are based on simple concepts while at the same time do not depend on specific assumptions about the distributions of the underlying data. The algorithm implemented in the R-package PCADAPT, for example, uses a principal component transformation of the data in combination with a linear regression model, and thus assumes linear relationships between the variant sites of the underlying data.

We have demonstrated that the evaluated kNN-based methods achieve qualitatively comparable performance with the Bayesian

**TABLE 1** Human chromosome 2 outlier windows

| Mb (start) | Mb (end) | Nearest genes | $F_{ST}$ [CEU/CHB, CEU/YRI, CHB/YRI] | $\Delta F_{ST}$ [CEU/CHB, CEU/YRI, CHB/YRI] |
|---|---|---|---|---|
| 17.3 | 17.4 | VSNL1, AC010880.1 | [0.55,0.16,0.34] | [0.46,0.02,0.19] |
| **72.5** | **72.6** | **EXOC6B**[a] | [0.15,0.45,0.69] | [0.06,0.31,0.54] |
| **72.6** | **72.7** | **EXOC6B**[a] | [0.13,0.43,0.65] | [0.03,0.29,0.50] |
| 72.8 | 72.9 | EXOC6B[a] | [0.13,0.34,0.57] | [0.03,0.20,0.42] |
| 74.7 | 74.8 | CCDC142[a], M1AP[a] | [0.53,0.37,0.22] | [0.43,0.23,0.07] |
| 104.1 | 104.2 | LINC01127 | [0.42,0.10,0.52] | [0.37,−0.04,0.36] |
| 109.1 | 109.2 | GCC2[a], LIMS1[a] | [0.43,0.09,0.58] | [0.33,−0.05,0.43] |
| 109.2 | 109.3 | LIMS1[a] | [0.40,0.09,0.55] | [0.30,−0.04,0.40] |
| 109.5 | 109.6 | **EDAR**[a] | [0.63,0.32,0.34] | [0.53,0.18,0.19] |
| 195.6 | 195.7 | LINC01790[a] | [0.03,0.50,0.42] | [−0.07,0.36,0.27] |

*Note:* Displayed are the 0.005-quantile outlier 100-kb windows all the kNN-based methods agree on, and the nearest genes to these windows. $F_{ST}$-medoid = [CEU/CHB = 0.09, CEU/YRI = 0.14, CEU/YRI = 0.15]. The top-3 outlier windows are highlighted in bold.

[a]The outlier window overlaps with the gene.

approach implemented in the R-package blockfest when detecting region-specific selection, while being considerably less computer-intensive. We showcased the capacity of the kNN-based methods to analyse real-world data by scanning the second chromosome of the human genome (data available by the 1,000 Genomes Project). We confirm known genes under positive selection, such as EDAR and EXOC6B, but also report a set of new candidate genes, such as LIMS1 and CNTNAP5. Outlier loci with significantly reduced divergence and thus potentially pointing to gene flow or balancing selection cannot be reported for human chromosome 2. The only candidate genes showing a weak signal of that type are the LINC01127 and ANTXR1 genes, with slightly reduced divergence between the CHB and YRI populations.

We have also discussed certain challenges that arise when employing kNN-based techniques. A widely known complication with the kNN-based methods, which merits further investigation, is the choice of $k$, for which the optimal value highly depends on the data. Using coalescent simulations with a wide range of varying parameters, we showed that the chosen $k$ value does not greatly affect the accuracy of our approaches. A genomic scan of human chromosome 2 yield stable kNN-score rank correlations. However, we highly recommend to conduct the provided diagnostic plot prior to the final analysis. Other data sets may behave differently. Also, the window sizes in genome scans for selection may substantially influence the stability of the kNN techniques and our proposed approaches to infer $k$. A visual inspection of the diagnostic plot should clearly indicate ranges of $k$ for which high kNN rank correlations can be reported. Future investigations will analyse the power of the kNN techniques, both analytically as well as through additional simulations over a wider range of population models and feature sets.

## ACKNOWLEDGEMENTS

We thank Fernando Racimo and Ben Rosenzweig for their valuable comments on our manuscript.

## AUTHOR CONTRIBUTIONS

B.P. designed the project. B.P. and M.G.S. developed the methods. B.P. performed the simulations and created the R scripts. B.P., N.A. and P.P. analysed the data from the 1,000 Genomes Project. B.P., N.A. and M.G.S. wrote the manuscript.

## DATA AVAILABILITY STATEMENT

We provide R scripts to perform kNN-based whole-genome scans, available at the GitHub repository *pievos101/kNN-Genome-Scans*. The code interfaces with the powerful genomics R-package popgenome (Pfeifer et al., 2014) and enables flexible genomic scans with sliding windows as well as genomic scans based on genomic features such as genes, UTRs or exons. We have made use of the kNN implementations of the elki software (Schubert and Zimek, 2019).

## ORCID

*Bastian Pfeifer* https://orcid.org/0000-0001-7035-9535

## REFERENCES

1000 Genomes Project Consortium and others (2015). A global reference for human genetic variation. *Nature*, *526*, 68.

Angiulli, F., & Pizzuti, C. (2002). Fast outlier detection in high dimensional spaces. In *European Conference on Principles of Data Mining and Knowledge Discovery. Proceedings*, *15–27*. Springer.

Angiulli, F., & Pizzuti, C. (2005). Outlier mining in large high-dimensional data sets. *IEEE Transactions on Knowledge and Data Engineering*, *17*, 203–215. https://doi.org/10.1109/TKDE.2005.31

Bache, K., & Lichman, M. (2013). UCI machine learning repository. Irvine, CA: University of California. School of information and Computer Science, 28. Retrieved from http://archive.ics.uci.edu/ml

Baye, T. M., Wilke, R. A., & Olivier, M. (2009). Genomic and geographic distribution of private SNPs and pathways in human populations. *Personalized Medicine*, *6*, 623–641. https://doi.org/10.2217/pme.09.54

Beaumont, M. A., & Balding, D. J. (2004). Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, *13*, 969–980. https://doi.org/10.1111/j.1365-294X.2004.02125.x

Beaumont, M. A., & Nichols, R. A. (1996). Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *263*, 1619–1626.

Bhatia, G., Patterson, N., Sankararaman, S., & Price, A. L. (2013). Estimating and interpreting FST: The impact of rare variants. *Genome Research*, *23*, 1514–1521. https://doi.org/10.1101/gr.154831.113

Bonhomme, M., Chevalet, C., Servin, B., Boitard, S., Abdallah, J., Blott, S., & SanCristobal, M. (2010). Detecting selection in population trees: The lewontin and krakauer test extended. *Genetics*, *186*, 241–262. https://doi.org/10.1534/genetics.110.117275

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. *ACM Sigmod Record*, *29*, 93–104. ACM.

Bryk, J., Hardouin, E., Pugach, I., Hughes, D., Strotmann, R., Stoneking, M., & Myles, S. (2008). Positive selection in East Asians for an edar allele that enhances nf-κb activation. *PLoS One*, *3*, e2209. https://doi.org/10.1371/journal.pone.0002209

Campos, G. O., Zimek, A., Sander, J., Campello, R. J., Micenková, B., Schubert, E., … Houle, M. E. (2016). On the evaluation of unsupervised outlier detection: Measures, datasets, and an empirical study. *Data Mining and Knowledge Discovery*, *30*, 891–927. https://doi.org/10.1007/s10618-015-0444-8

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*, 21–27. https://doi.org/10.1109/TIT.1967.1053964

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

De Villemereuil, P., & Gaggiotti, O. E. (2015). A new FST-based method to uncover local adaptation using environmental variables. *Methods in Ecology and Evolution*, *6*, 1248–1258.

Duforet-Frebourg, N., Bazin, E., & Blum, M. G. (2014). Genome scans for detecting footprints of local adaptation using a Bayesian factor model. *Molecular Biology and Evolution*, *31*, 2483–2495. https://doi.org/10.1093/molbev/msu182

Duforet-Frebourg, N., Luu, K., Laval, G., Bazin, E., & Blum, M. G. (2015). Detecting genomic signatures of natural selection with principal component analysis: Application to the 1000 genomes data. *Molecular Biology and Evolution*, *33*, 1082–1093. https://doi.org/10.1093/molbev/msv334

Durand, E. Y., Patterson, N., Reich, D., & Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, *28*, 2239–2252. https://doi.org/10.1093/molbev/msr048

Ewing, G., & Hermisson, J. (2010). MSMS: A coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, *26*, 2064–2065. https://doi.org/10.1093/bioinformatics/btq322

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

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

Fix, E. (1951). *Discriminatory analysis: nonparametric discrimination, consistency properties*. USAF school of Aviation Medicine.

Foll, M., & Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. *Genetics*, *180*, 977–993. https://doi.org/10.1534/genetics.108.092221

François, O., Martins, H., Caye, K., & Schoville, S. D. (2016). Controlling false discoveries in genome scans for selection. *Molecular Ecology*, *25*, 454–469. https://doi.org/10.1111/mec.13513

Gautier, M. (2015). Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics*, *201*, 1555–1579. https://doi.org/10.1534/genetics.115.181453

Geneva, A. J., Muirhead, C. A., Kingan, S. B., & Garrigan, D. (2015). A new method to scan genomes for introgression in a secondary contact model. *PLoS One*, *10*, e0118621. https://doi.org/10.1371/journal.pone.0118621

Grau, J., Grosse, I., & Keilwagen, J. (2015). PRROC: Computing and visualizing precision-recall and receiver operating characteristic curves in r. *Bioinformatics*, *31*, 2595–2597. https://doi.org/10.1093/bioinformatics/btv153

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., … Paabo, S. (2010). A draft sequence of the neandertal genome. *Science*, *328*, 710–722. https://doi.org/10.1126/science.1188021

Günther, T., & Coop, G. (2013). Robust identification of local adaptation from allele frequencies. *Genetics*, *195*, 205–220. https://doi.org/10.1534/genetics.113.152462

Haasl, R. J., & Payseur, B. A. (2016). Detecting selection in natural populations: Making sense of genome scans and towards alternative solutions: Fifteen years of genomewide scans for selection: Trends, lessons and unaddressed genetic sources of complication. *Molecular Ecology*, *25*, 5. https://doi.org/10.1111/mec.13339

Hahn, M. W., & Hibbins, M. S. (2019). A three-sample test for introgression. *Molecular Biology and Evolution*, *36*, 2878–2882. https://doi.org/10.1093/molbev/msz178

Hautamaki, V., Karkkainen, I., & Franti, P. (2004). Outlier detection using k-nearest neighbour graph. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. *3*, 430–433. IEEE.

Hedrick, P. W. (2013). Adaptive introgression in animals: Examples and comparison to new mutation and standing variation as sources of adaptive variation. *Molecular Ecology*, *22*, 4606–4618. https://doi.org/10.1111/mec.12415

Hibbins, M. S., & Hahn, M. W. (2019). The timing and direction of introgression under the multispecies network coalescent. *Genetics*, *211*, 1059–1073. https://doi.org/10.1534/genetics.118.301831

Hudson, R. R., Slatkin, M., & Maddison, W. P. (1992). Estimation of levels of gene flow from DNA sequence data. *Genetics*, *132*, 583–589.

Jin, W., Tung, A. K., Han, J., & Wang, W. (2006). Ranking outliers using symmetric neighborhood relationship. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining. Proceedings.*, 577–593. Springer.

Kriegel, H.-P., Kröger, P., Schubert, E., & Zimek, A. (2009). LoOP: Local outlier probabilities. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, 1649–1652. ACM.

Kriegel, H.-P., Schubert, M., & Zimek, A. (2008). Angle-based outlier detection in high-dimensional data. In *Proceedings of the 14th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining*, 444–452. ACM.

Latecki, L. J., Lazarevic, A., & Pokrajac, D. (2007). Outlier detection with kernel density functions. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition. Proceedings.*, 61–75. Springer.

Li, H., & Ralph, P. (2019). Local PCA shows how the effect of population structure differs along the genome. *Genetics*, *211*, 289–304. https://doi.org/10.1534/genetics.118.301747

Lotterhos, K. E., Card, D. C., Schaal, S. M., Wang, L., Collins, C., & Verity, B. (2017). Composite measures of selection can improve the signal-to-noise ratio in genome scans. *Methods in Ecology and Evolution*, *8*, 717–727. https://doi.org/10.1111/2041-210X.12774

Lotterhos, K. E., & Whitlock, M. C. (2014). Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Molecular Ecology*, *23*, 2178–2192.

Luu, K., Bazin, E., & Blum, M. G. (2017). PCADAPT: An R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources*, *17*, 67–77.

Martin, S. H., Davey, J. W., & Jiggins, C. D. (2014). Evaluating the use of ABBA–BABA statistics to locate introgressed loci. *Molecular Biology and Evolution*, *32*, 244–257. https://doi.org/10.1093/molbev/msu269

Pfeifer, B., & Kapan, D. D. (2019). Estimates of introgression as a function of pairwise distances. *BMC Bioinformatics*, *20*, 207. https://doi.org/10.1186/s12859-019-2747-z

Pfeifer, B., & Lercher, M. J. (2018). BLOCKFEST: Bayesian calculation of region-specific $F_{ST}$ to detect local adaptation. *Bioinformatics*, *34*, 3205–3207. https://doi.org/10.1093/bioinformatics/bty299

Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S. E., & Lercher, M. J. (2014). POPGENOME: An efficient Swiss army knife for population genomic analyses in R. *Molecular Biology and Evolution*, *31*, 1929–1936. https://doi.org/10.1093/molbev/msu136

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). PROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 77. https://doi.org/10.1186/1471-2105-12-77

Schubert, E., & Zimek, A. (2019). ELKI: A large open-source library for data analysis-ELKI Release 0.7. 5" heidelberg". *arXiv preprint arXiv:1902.03616*.

Schubert, E., Zimek, A., & Kriegel, H.-P. (2014). Local outlier detection reconsidered: A generalized view on locality with applications to spatial, video, and network outlier detection. *Data Mining and Knowledge Discovery*, *28*, 190–237. https://doi.org/10.1007/s10618-012-0300-z

Tang, J., Chen, Z., Fu, A.-W.-C., & Cheung, D. W. (2002). Enhancing effectiveness of outlier detections for low density patterns. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining. Proceedings*., 535–548. Springer.

de Villemereuil, P., Frichot, É., Bazin, É., François, O., & Gaggiotti, O. E. (2014). Genome scan methods against more complex models: When and how much should we trust them? *Molecular Ecology*, *23*, 2006–2019. https://doi.org/10.1111/mec.12705

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, *38*, 1358–1370.

Weir, B. S., & Ott, J. (1997). Genetic data analysis II. *Trends in Genetics*, *13*, 379. https://doi.org/10.1016/S0168-9525(97)81169-9

Wright, S. (1949). The genetical structure of populations. *Annals of Eugenics*, *15*, 323–354. https://doi.org/10.1111/j.1469-1809.1949.tb02451.x

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z. X. P., Pool, J. E., … Wang, J. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. *Science*, *329*, 75–78. https://doi.org/10.1126/science.1190371

Zhang, K., Hutter, M., & Jin, H. (2009). A new local distance-based outlier detection approach for scattered real-world data. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining. Proceedings*., 813–822. Springer.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

**How to cite this article:** PfeiferB, Alachiotis N, Pavlidis P, Schimek MG. Genome scans for selection and introgression based on *k*-nearest neighbour techniques. *Mol Ecol Resour.* 2020;20:1597–1609. https://doi.org/10.1111/1755-0998.13221