Genome Scans for Selection and Introgression based on k-nearest Neighbor Techniques

Bastian Pfeifer,*,[1] Nikolaos Alachiotis,[2] Pavlos Pavlidis,[2] and Michael G. Schimek*,[1]

[1] Institute for Medical Informatics, Statistics and Documentation, Medical University, Graz, Austria

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

***Corresponding author:** E-mail: bastianxpfeifer@gmail.com, michael.schimek@medunigraz.at

Associate Editor: -

Abstract

In recent years, genome-scan methods have been extensively used to detect signatures of selection and introgression. Here, we compare the latest genome-scan methods with non-parametric k-nearest neighbors (kNN) anomaly detection algorithms, while incorporating pairwise Fixation Index (F_{ST}) estimates and pairwise nucleotide differences (d_{xy}) as features. Simulations were performed for both positive directional selection and introgression, with varying parameters, such as recombination rates, population background histories, the proportion of introgression, and the time of gene flow. We find that kNN-based methods perform remarkably well while yielding stable results almost over the entire range of k. Furthermore, the weighted-kNN algorithm for detecting directional selection, and the INFLO (Influenced Outlierness) algorithm for detecting introgression, outperform recently published methods. We provide a GitHub repository (*pievos101/kNN-Genome-Scans*) containing R source code to demonstrate how to apply the proposed methods to real-world genomic data using the population genomics R-package PopGenome.

Key words: genome scans, selection, introgression, adaptation, SNPs.

Introduction

The last years have seen great advances in whole genome sequencing and population genomics methods to detect DNA fragments affected by natural selection. Genomic regions under selection are assumed to be rare and thus can be considered as anomalies which deviate from the overall population structure. These anomalies are of great interest as they may act as a major force in the adaptation of populations to their environments during evolution. One of the most widely applied statistics to detect such regions is the Fixation Index (F_{ST}), which was originally proposed as a measure of population differentiation under the Wright-Fisher model (Wright, 1949).

Several variations of the F_{ST} are used by the population genomics community (Hudson *et al.*, 1992; Weir, 1996; Weir and Cockerham, 1984). In general, a high F_{ST} can be an indication of positive directional selection. However, in cases where the neutral population background history (the neutral distribution of F_{ST}) is not known, hypothesis testing of neutral evolution is nearly impossible. This especially applies in cases where the population history deviates from the Wright-Fisher model, or when hierarchical structure is introduced to the system. In such cases, results based on F_{ST} are no longer reliable (Excoffier *et al.*, 2009; Foll and Gaggiotti, 2008).

Also, the F_{ST} estimate was introduced as a model parameter in Bayesian approaches, and inferred via computationally intensive Markov-Chain-Monte-Carlo (MCMC) simulations. In such approaches, a common migrant pool is modeled 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 and 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 suffer from a high False Discovery Rate (FDR) (De Villemereuil and Gaggiotti, 2015; Duforet-Frebourg *et al.*, 2014, 2015). More recently, published approaches based on Principal Component Analyses (PCA) address some of these shortcomings (Duforet-Frebourg *et al.*, 2015; Luu *et al.*, 2017), and at the same time they are computationally less demanding.

Another topic of great interest is the investigation of hybridization and the detection of the related introgressed regions in whole genome scans. Hybridization between species is increasingly recognized as an evolutionary force in which species share genetic information across the species boundary. There has been an explosion of available methods in this area in recent years. Currently, the most widely applied methods are the ABBA-BABA family of methods which are based on a four-taxon system (Durand *et al.*, 2011; Green *et al.*, 2010; Martin *et al.*, 2014; Pfeifer and Kapan, 2019), where the fourth taxon acts as the outgroup. Since an outgroup is not always available, several other approaches based on a three-taxon system were also introduced (Hibbins and Hahn, 2019; Rosenzweig *et al.*, 2016).

We believe that the ability of techniques based on k-nearest neighbors (kNN) to detect genomic signatures of selection or introgression is widely underestimated, and detailed investigations on the use of these techniques are yet to be reported. The kNN-based approaches are among the oldest unsupervised machine learning techniques and have been widely applied in almost all areas of data-driven research. In this paper, we make use of kNN-based techniques while incorporating pairwise F_{ST} estimates as features. We study the ability of these approaches to detect local signatures of directional selection and introgression under a wide range of simulation scenarios. In the case of introgression, we also use pairwise nucleotide differences (d_{xy}) as features because it has been reported that F_{ST} estimates produce inflated values when diversities within populations are low, which may lead to false positives when searching for introgressed regions within the genome (Cruickshank and Hahn, 2014).

We compare the accuracy of these approaches with recently published genome-scan methods, and finally showcase the use of the kNN approaches to detect positively selected regions in the human genomics data made available by the 1000 genomes project (Consortium *et al.*, 2015).

New Approaches

kNN Techniques using F_{ST} as Features

The key idea of the kNN approach (Ramaswamy *et al.*, 2000) is to calculate the distances between a given data point and its k nearest neighbors. Data points at high distance from their neighborhood are considered as outliers. In this work, we use pairwise F_{ST} estimates, as proposed by Hudson *et al.* (1992) and recommended by Bhatia *et al.* (2013), and incorporate them as features into kNN-based algorithms. Consequently, the population pairwise F_{ST} estimates define a genomic region as a data point embedded into an m-dimensional numerical space, where m is the total number of possible population pairwise comparisons ($m = np(np-1)/2$), and np is the total number of populations analyzed. Thus, each genomic region is represented by an F_{ST} vector of length m. The kNN score for a given genomic region x is calculated as follows:

$$kNN_k(x) = \frac{\sum_{o \in N_k(x)} d_k(x,o))}{|N_k(x)|}, \qquad (1)$$

where $N_k(x)$ is the k-nearest neighbor set of the genomic region x, and $d_k(x,o)$ defines the reachability distance between the genomic regions x and o. It is calculated as the euclidean distance between the pairwise F_{ST} vectors F_{STx} and F_{STo}:

$$d_k(x,o) = \sum_{i=1}^{m} (F_{STxi} - F_{SToi})^2. \qquad (2)$$

The basic kNN approach was slightly modified by the weighted-kNN approach (Angiulli and Pizzuti, 2002, 2005), which takes into account the overall distance from a data point to its neighborhood by calculating the sum of distances instead of the arithmetic mean. Another way of calculating the distances is implemented in ODIN (Outlier Detection using Indegree Number) (Hautamaki *et al.*, 2004), which infers outliers based on a kNN graph.

A wide range of methods have been developed to also account for local outlierness. The best known one is LOF (Local Outlier Factor) (Breunig *et al.*, 2000), which is based on the concept of local reachability density (lrd) of the k-nearest neighbors. In this context, a data point is considered to be an outlier when its density is much smaller than the densities of its neighbors. The lrd is defined as

$$lrd_k(x) = \frac{1}{kNN_k(x)}, \qquad (3)$$

and the LOF can be calculated as

$$LOF_k(x) = \frac{1}{|N_k(x)|} \sum_{o \in N_k(x)} \frac{lrd_k(o)}{lrd_k(x)}. \qquad (4)$$

The LOF algorithm and the corresponding lrd concept was later modified in several ways. For example, the simplified-LOF (Schubert *et al.*, 2014) uses the basic kNN distances instead of the LOFs reachability distance. COF (Connectivity-based Outlier Factor) (Tang *et al.*, 2002) modifies the density estimation of the simplified-LOF to account for the connectedness of a neighborhood via a minimum spanning tree (MST). Another tool is called LoOP (Local Outlier Probabilities) (Kriegel *et al.*, 2009), which adopts normalized local density scores based on the quadratic mean. Therefore, the scores are strictly within the [0,1] interval, and can be interpreted as *p*-values. LDOF (Local Distance-based Outlier Factor) (Zhang *et al.*, 2009) uses the relative distance from a data point to its neighbours, measuring how many data points deviate from their scattered neighbourhood. The ABOD approach addresses the so-called "curse of dimensionality" problem by comparing the angles between pairs of distance vectors. FastABOD (Fast Angle-Based Outlier Detection) is a faster variant of ABOD (Kriegel *et al.*, 2008). LDF (Local Density Factor) (Latecki *et al.*, 2007) replaces LOF's density estimation by a variable-width Gaussian kernel density estimation (KDE). INFLO (Influenced Outlierness) (Jin *et al.*, 2006) takes into account also the reverse nearest neighborhood set when calculating the local density scores.

The Selection of k

In classification problems, the parameter k can be inferred, for instance, by cross-validation. However, it is well known that the inference of an appropriate k in a purely unsupervised setting is a challenging task, and highly depends on the data analyzed. This challenge especially arises in studies where the goal is the detection of local outliers. However, here we use the kNN-based methods for global outlier detection, with the aim to distinguish between signals of neutral evolving genomic regions and outlier regions subject to selection or introgression. Thus, the choice of k may not have a big influence on the outcomes as long as the value is not too small.

There are two main requirements that an adequate k should fulfill. First, the scores of the corresponding kNN methods need to be reasonable, which we think is fulfilled when the ranks of the kNN scores approximately align with the findings of well established methods. Second, the k should be settled in a stable k region. We refer to a stable k region when the ranks of the corresponding kNN scores within that region are highly correlated. To address the former requirement we simply use F_{ST} estimates as features for the kNN-based methods. F_{ST} is successfully applied in genomic scans to detect positive selection as well as introgression. In this work, we use the F_{ST} estimate as proposed by Hudson *et al.* (1992) and recommended by Bhatia *et al.* (2013). To address the latter requirement we propose the following approach:

First, calculate the kNN-scores ($\mathbf{s}_i$) for $n_k = 100$ sequentially sampled ks from $[2, n_r - 1]$, where n_r is total number of genomic regions:

$$\mathbf{s}_i = kNN_{k(i)}(X) \quad \forall i = 1, \ldots, n_k \tag{5}$$

Second, calculate Kendalls tau (τ) correlation coefficients

$$corr_i = median[\tau(\mathbf{s}_{i-1}, \mathbf{s}_i), \tau(\mathbf{s}_i, \mathbf{s}_{i+1}), \tau(\mathbf{s}_{i-1}, \mathbf{s}_{i+1})] \\ \forall i = 2, \ldots, n_k - 1 \tag{6}$$

where

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{n_r(n_r - 1)/2} \tag{7}$$

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.

Inferring the Type of Selection

The kNN-based methods are reporting on anomalies as strong deviations from the overall population structure. Once the outlier are detected from the kNN-based outlier scores we suggest to remove the corresponding pairwise F_{ST} vectors and to calculate the medoid based on the remaining data points. We argue that the medoid is the most informative data point and sufficiently reflects the overall population structure. Subtracting the medoid from the outlier pairwise F_{ST} vectors will indicate which population or population pairs are affected by selection. We are pointing to these resulting vectors as the ΔF_{ST} selection effects. Positive entries of ΔF_{ST} are an indication for positive directional selection, whereas negative values point to introgression (reduced diversity due to gene-flow) or other types of selection which significantly reduce the diversity between populations, such as balancing selection.

Results

On the Power to Detect Positive Directional Selection

Simulations under positive directional selection indicate that the kNN-based methods are almost unaffected by the choice of k (fig. 1). Unstable results are only observed for either small or high values of k, with respect to the total number of genomic regions analyzed. In comparison with established methods, such as pcadapt, the kNN-based methods do remarkably well. As expected, the F_{ST} results are fully comparable for star-like genealogies (fig. 1A). However, as soon as hierarchical structure is introduced to the population history, our proposed competing methods show overall higher AUC values. In fact, the kNN-based techniques remain almost

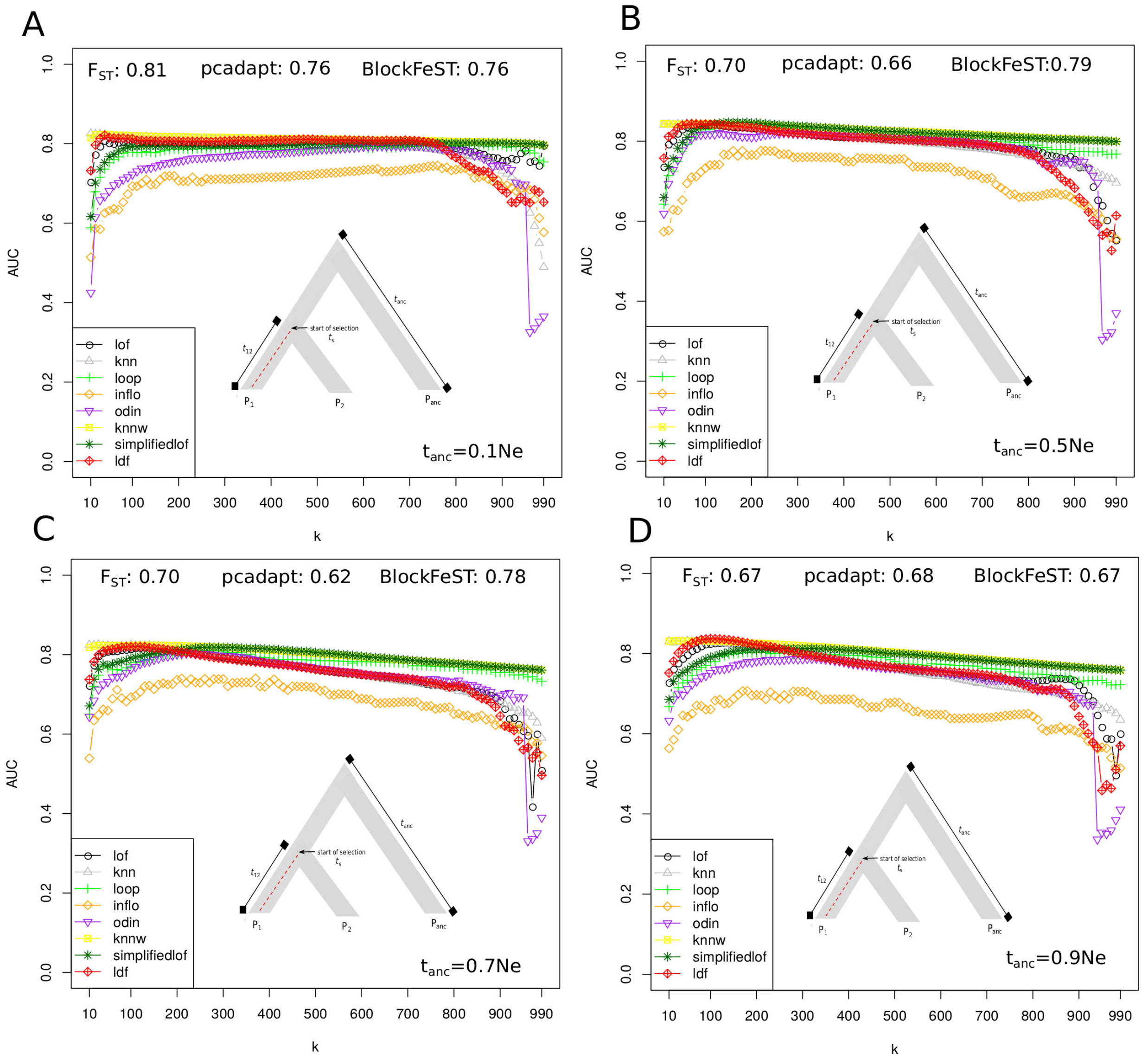

FIG. 1. Positive directional selection: varying the coalescent time to the ancestral population (t_{anc}). The results for the kNN-based methods using F_{ST} as features are shown for 100 sequentially sampled ks ($k=[1,10,\ldots,990,1000]$) and in comparison to the accuracy of F_{ST}, pcadapt and BlockFeST. Recombination rate is set to $r=0.001$. A. The simulations are based on a star formed genealogie ($t_{12}=0.1=t_{anc}$). B. The coalescent time to the ancestral population is $t_{anc}=0.5N_e$. C. The coalescent time to the ancestral population is $t_{anc}=0.7N_e$. D. The coalescent time to the ancestral population is $t_{anc}=0.9N_e$.

unaffected when varying the coalescent times to the ancestral population P_{anc} (fig. 2A).

The weighted-kNN and simplified-LOF methods are the strongest kNN-based methods, both outperforming F_{ST} and pcadapt, and are comparable to BlockFeST (fig. 1, fig. 2). However, 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 (fig. 4B). This is expected because the signal of selection gets eroded, which makes it harder to detect these patterns. Based on our simulations, we observed that the INFLO algorithm is the weakest kNN-based method for the detection of

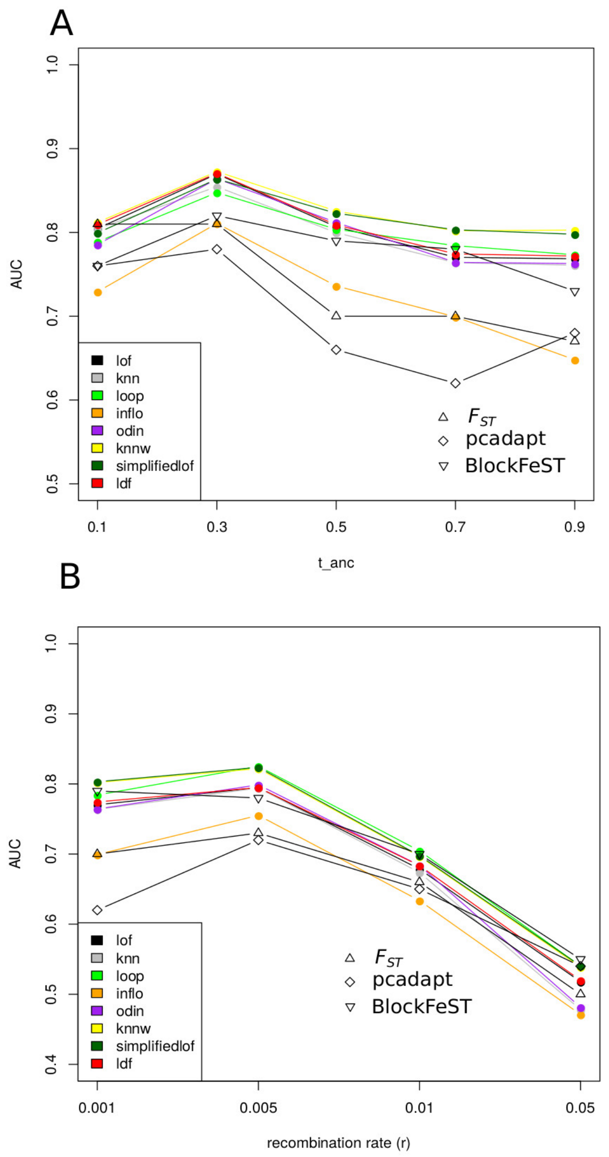

FIG. 2. Detecting positive directional selection with a computed k. The kNN methods using F_{ST} as features compared to F_{ST}, pcadapt and BlockFeST.
A. Varying the coalescent time with the ancestral population (t_{anc}=[0.1,0.3,0.5,0.7,0.9]N_e generations ago). The recombination rate is set to r=0.001. B. Varying the recombination rates (r=[0.001,0.005,0.01,0.05]). The coalescent time with the ancestral population is t_{anc}= 0.7N_e generations ago.

positive directional selection, especially when the recombination rates are high. Finally, INFLO is the most sensitive to background population histories as seen from fig. 2A.

On the Power to Detect Introgression

Simulations under uni-directional introgression from the archaic population P_{anc} to population P_2 confirm that the kNN-based family of methods is almost unaffected by the choice of k. Surprisingly, we observed that in some cases F_{ST} outperforms the other more specialized methods (fig. 3A). However, we also report unstable results for F_{ST} when varying the time of gene-flow (fig. 4B). In fact, F_{ST} has been previously reported to potentially lead to false positives when scanning the genome for introgressed regions (Cruickshank and Hahn, 2014; Rosenzweig *et al.*, 2016). The reason is that F_{ST} also takes into account the within-population diversity and thus might has inflated values at genomic locations with overall low diversity. RNDmin is more stable in these cases but not as powerful as the kNN-based methods (fig. 4).

Interestingly, while INFLO is the weakest kNN-based method for the detection of directional selection, our simulations point at INFLO as a potentially powerful algorithm to detect local signatures of introgression (fig. 3). Increasing the proportion of introgression has the same effect on all kNN-based methods: the accuracy increases and is nearly at 100% when f=0.5. RNDmin has a much lower AUC value in that case. Overall, similar as reported for the positive directional selection cases, the weighted kNN method and simplified-LOF show high accuracy also in the introgression cases and provide stable

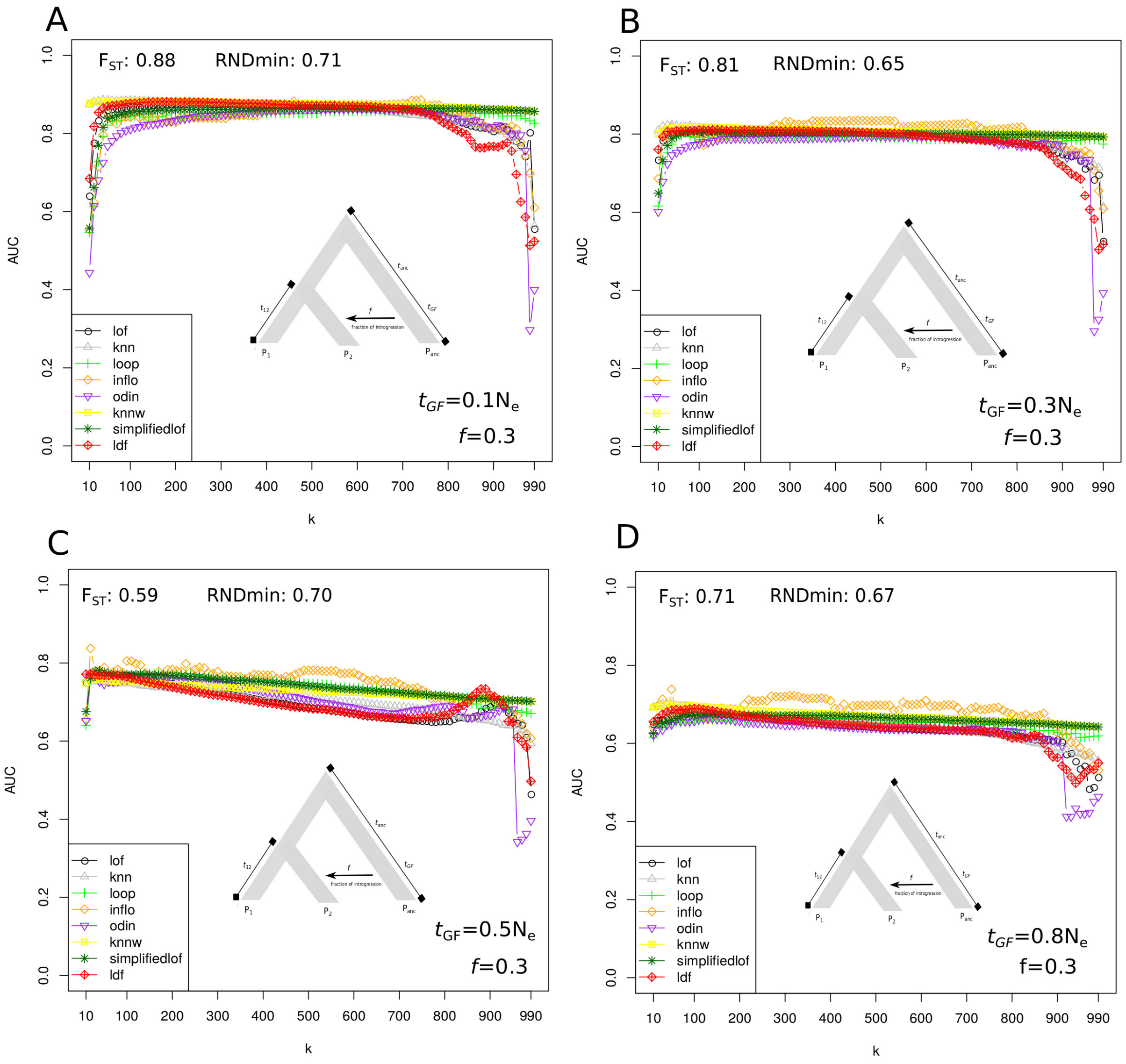

FIG. 3. Varying the time of gene-flow (t_{GF}). The results for the kNN-based methods using F_{ST} as features shown for 100 sequentially sampled kś (k=[1,10,…,990,1000]). Coalescent times are $t_{12}=1N_e$ and $t_{anc}=2N_e$ generations ago. Recombination rate is set to $r=0.01$ in all simulations. The outcome of the kNN-based methods are compared to F_{ST} and RNDmin. The time of gene-flow is set to A. $t_{GF}=0.1N_e$ B. $t_{GF}=0.3N_e$ C. $t_{GF}=0.5N_e$ and D. $t_{GF}=0.8N_e$ generations ago.

score rankings almost across the full range of k. Also, in comparison to the other kNN approaches, low and very high k values have the most negative effect on ODIN and LDF. Using d_{xy} as features also give stable results for almost all choices of k (Supplementary Fig. S1). However, in this situation results are not as good as those of kNN techniques with incorporated pairwise F_{ST} estimates used as features. This is especially true when the time of gene-flow is recent, and a low fraction of introgression is shared by P_2 and the archaic population P_{anc} (Supplementary Fig. S2).

Application to the 1000 Genomes Data

We also analyzed the 1000 Genomes Data (Consortium *et al.*, 2015) to demonstrate the

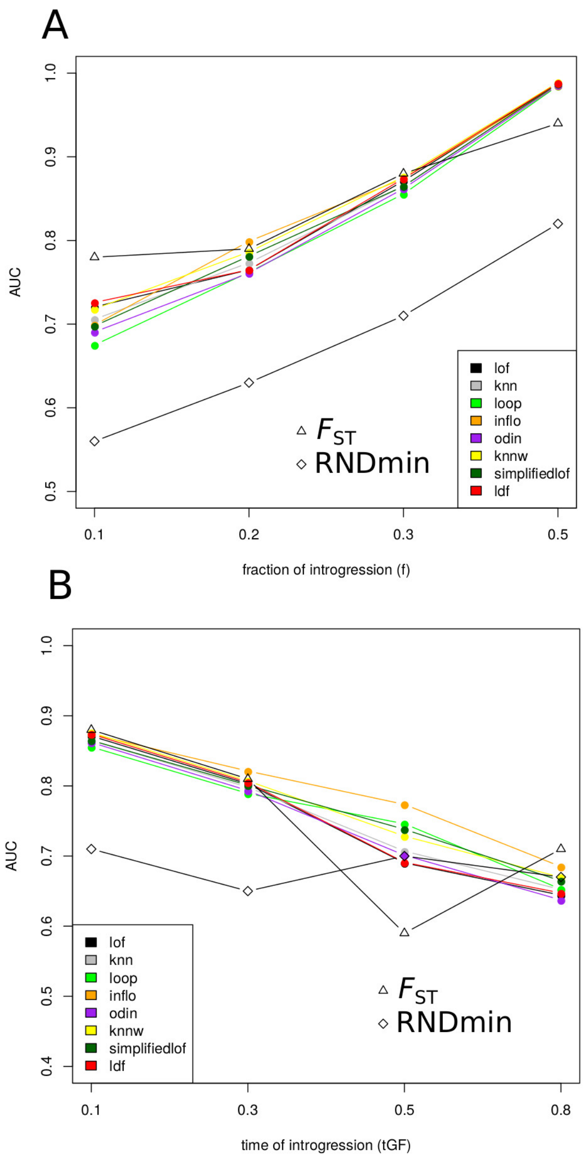

FIG. 4. Detecting introgression with a computed k. The accuracy of the kNN-methods using F_{ST} as features compared to F_{ST} and RNDmin. Recombination rate is set to $r=0.01$ in all simulations. A. Varying the fraction of introgression ($f=[0.1,0.2,0.3,0.5]$) B. Varying the time of gene-flow ($t_{GF}=[0.1,0.3,0.5,0.8]$).

efficacy of our proposed kNN-based approaches when processing real data. The employed dataset is currently one of the largest publicly available datasets, 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 on the samples of the populations CEU, CHB, and YRI, evaluating non-overlapping sliding windows of size 100kb. Here, we summarize the results for chromosome 2 by reporting the windows all kNN tools agree on to be outlier windows. We report on the nearest genes to these outlier windows (Table 1). For each tool we consider kNN scores within a conservative 0.005-quantile to define the outlier candidates.

Describing the properties and attributes of all these genes may lead to the story-telling fallacy (Pavlidis *et al.*, 2012). We therefore report for some of them what has been reported in literature. The top-2 candidate genes for directional positive selection are the protein coding genes EXOC6B and EDAR (table 1, fig. 5). Baye *et al.* (2009) report 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: 126.1-126.2Mb) and is confirmed by all tools but ODIN. The selection effect is $\Delta F_{ST}=$ [CEU/CHB=0.39, CEU/YRI=0.01,

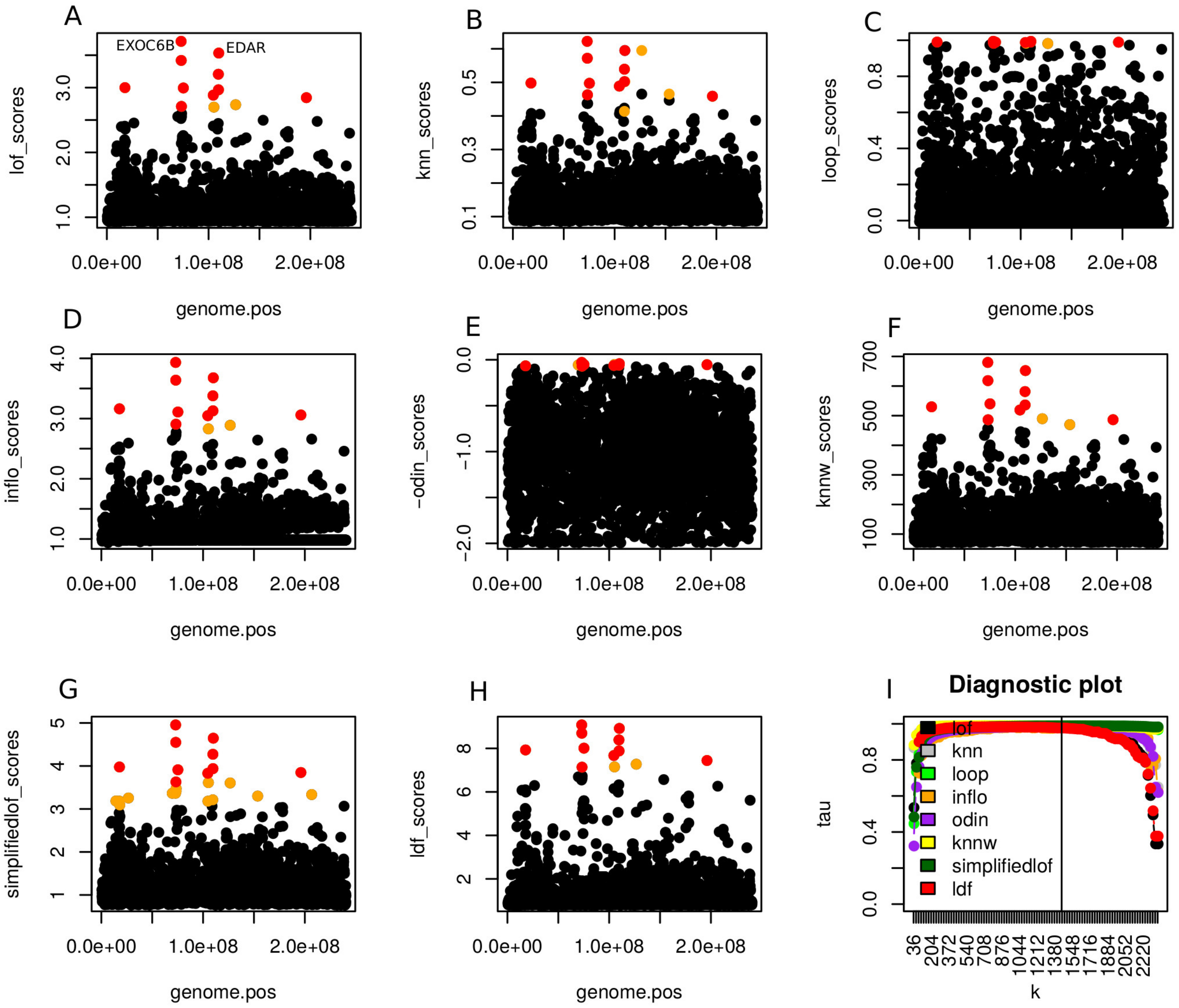

FIG. 5. Genome scan plots of human chromosome 2. A-H. The kNN scores are shown along human chromosome 2 based on 100kb consecutive sliding windows. Red and Orange dots are the outliers found by that specific kNN tool (0.005-quantile of the scores). Red dots indicate that all kNN tools agree on these outliers. I. A diagnostic plot is shown with the pairwise rank correlations of the kNN scores while varying the parameter k.

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

Table 1. Human chromosome 2 outlier windows

Mb (start)	Mb (end)	Nearest genes	F_{ST} [CEU/CHB,CEU/YRI,CHB/YRI]	ΔF_{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***	[0.15,0.45,0.69]	[0.06,0.31,0.54]
72.6	**72.7**	**EXOC6B***	[0.13,0.43,0.65]	[0.03,0.29,0.50]
72.8	72.9	EXOC6B*	[0.13,0.34,0.57]	[0.03,0.20,0.42]
74.7	74.8	CCDC142*, M1AP*	[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*,LIMS1*	[0.43,0.09,0.58]	[0.33,-0.05,0.43]
109.2	109.3	LIMS1*	[0.40,0.09,0.55]	[0.30,-0.04,0.40]
109.5	109.6	**EDAR***	[0.63,0.32,0.34]	[0.53,0.18,0.19]
195.6	195.7	LINC01790*	[0.03,0.50,0.42]	[-0.07,0.36,0.27]

NOTE.—Shown are the 0.005-quantile outlier 100kb windows all kNN-based methods agree with and the nearest genes to these windows. The medoid F_{ST} vector is FST-medoid= [CEU/CHB=0.09, CEU/YRI=0.14, CEU/YRI=0.15]. The top-3 outlier windows are highlighted in bold. *The outlier window overlaps with the gene

YRI.

Discussion

In this paper, we have investigated the usage of the kNN-based algorithms to detect signatures of selection and introgression in whole-genome scans. Coalescent simulations under positive directional selection and introgression show that the kNN-based methods using F_{ST} as features perform remarkably well, and are almost unaffected by the choice of k. Furthermore, 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 populations. We have demonstrated that the evaluated kNN-based methods achieve qualitatively comparable performance with the Bayesian approach implemented in the R-package BlockFeST when detecting positive directional selection, while being considerably less compute-intensive. We showcased the capacity of the kNN-based methods to analyze real-world data by scanning the second chromosome of the human genome (data available by the 1000 Genomes project). We confirm known genes under positive selection, like EDAR and EXOC6B, but also report a set of new candidate genes, like LIMS1 and CNTNAP5. Outlier loci with significantly reduced diversity, and thus potentially pointing to gene-flow or balancing selection cannot be reported for human chromosome 2. The only candidate genes showing that type of signal are the LINC01127 and ANTXR1 genes, with slightly reduced diversity 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

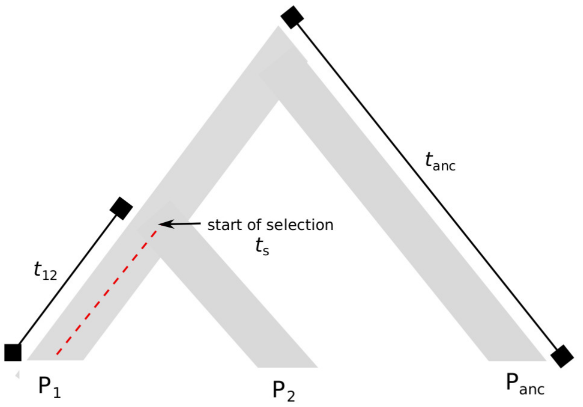

FIG. 6. A sketched graphical illustration of positive directional selection. A three population genealogy with positive directional selection introduced at $t_S N_e$ generations ago in population P_1.

methods is the choice of k, for which the optimal value highly depends on the data. This problem is not fully solved with our approaches. We have shown, however, that under coalescent simulations and a convincing set of population models, the parameter k does not greatly affect the accuracy of our approaches. Future investigations will analyze the power of the kNN techniques, both analytically as well as through additional simulations over a wide range of population models.

Materials and Methods

Simulation of Positive Directional Selection

We generated 950 neutral regions and 50 regions under positive directional selection (fig. 6) with the MSMS software tool (Ewing and Hermisson, 2010). The main calls to the MSMS program are as follows:

Neutral model:

```
msms 300 950 -s 50 -N 10000 -I 3 100
100 100 0 -ej 0.1 2 1 -ej 0.3 3 1 -r 100
10000
```

Alternative model:

```
msms 300 50 -s 50 -N 10000 -I 3 100 100
100 0 -ej 0.1 2 1 -ej 0.3 3 1 -r 100
10000 -SAA 2000 -SaA 0 -SI 0.1 3 0.01 0 0
-SFC
```

The above calls generate three populations, each comprising 100 samples (`-I`). The number of SNPs per each region is 50 (`-s`), and the effective population size is $N_e = 10000$ (`-N`). The first coalescent event of population P_1 and P_2 is fixed at $t_{12} = 0.1 N_e$, and the second coalescent event is set to $t_{13} = 0.2 N_e$ generations ago. The selection strength for homozygotes is $s = 0.1$ (`-SAA`), where selection starts at $t_s = 0.1 N_e$ generations ago (`-SI`) in population P_1. The recombination rate is $r = 0.01$ (`-r`). We varied recombination rates ($r = [0, 0.001, 0.005, 0.01, 0.05]$) and the time of coalescence with the ancestral population ($t_{anc} = [0.1, 0.3, 0.5, 0.7, 0.9]$). In each of these simulations, we made use of the `SFC` parameter in order to drop simulations when the selected allele gets lost.

Simulation of Introgression

To generate introgression events (fig. 7) we follow the simulation set-up of Martin *et al.* (2014). The below calls to the MS software (Hudson, 2002) generate the topologies:

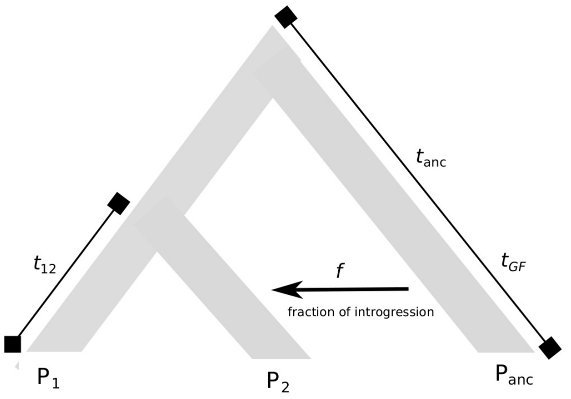

FIG. 7. A sketched graphical illustration of introgression. A three population species tree with an uni-directional introgression event from the ancestral population P_{anc} to population P_2 introduced $t_{GF}N_e$ generations ago. The proportion of introgression is indicated by f.

Neutral model:

```
ms 32 950 -I 4 8 8 8 8 -ej 1 2 1 -ej 2 3
1 -ej 3 4 1 -r 10 1000
```

Alternative model:

```
ms 32 50 -I 4 8 8 8 8 -ej 1 2 1 -ej 2 3 1
-ej 3 4 1 -es 0.1 2 0.7 -ej 0.1 5 3 -r 10
1000
```

These calls simulate 950 neutral regions and 50 regions under introgression with four populations including 8 samples each (`-I`). The coalescent times are $t_{12}=1N_e$, $t_{13}=2N_e$ and $t_{14}=3N_e$ generations ago. We introduced $P_3 \rightarrow P_2$ introgression $t_{GF}=0.1N_e$ generations ago (`-es`) with a fraction of introgression $f=0.3$ (`-es 0.1 2 [1-f]`). The recombination rate was set to $r=0.01$ in all simulations.

Finally, the nucleotide sequences were generated using the seq-gen (Rambaut and Grass, 1997) software with the following call:

```
seq-gen -mHKY -I 1000 -s 0.01
```

This generates 1kB sequences under the Hasegawa-Kishino-Yano (`-mHKY`) substitution model with a branch scaling factor of $s=0.01$ (`-s`). We varied the proportion of introgression ($f=[0.1,0.2,0.3,0.5]$) and the time of gene-flow ($t_{GF}=[0.1,0.3,0.5,0.8]$).

Comparison with Other Methods

We selected a set of kNN-based techniques similar to Campos *et al.* (2016), and compare them with well established genome-scan methods. For the selection cases, we contrast the kNN-based algorithms to the method implemented in the R-package pcadapt (Luu *et al.*, 2017). We computed the sum of *log-p-values* to label a region to make it comparable to the other methods under consideration. The number of principal components was set to K=2. In addition, we report the accuracy of the recently published method implemented in the R-package BlockFeST (Pfeifer and Lercher, 2018) using the default parameters. In the introgression cases we compare the kNN-based methods to the RNDmin approach (Rosenzweig *et al.*, 2016). In addition, this method is compared to the kNN-based techniques using d_{xy} as features. Finally, we relate all of these methods to the F_{ST} estimate (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).

Code Availability

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.

References

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

Angiulli, F. and Pizzuti, C. 2005. Outlier mining in large high-dimensional data sets. *IEEE transactions on Knowledge and Data engineering*, 17(2): 203–215.

Baye, T. M., Wilke, R. A., and Olivier, M. 2009. Genomic and geographic distribution of private snps and pathways in human populations. *Personalized medicine*, 6(6): 623–641.

Beaumont, M. A. and Balding, D. J. 2004. Identifying adaptive genetic divergence among populations from genome scans. *Molecular ecology*, 13(4): 969–980.

Beaumont, M. A. and 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(1377): 1619–1626.

Bhatia, G., Patterson, N., Sankararaman, S., and Price, A. L. 2013. Estimating and interpreting fst: the impact of rare variants. *Genome research*, 23(9): 1514–1521.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. 2000. Lof: identifying density-based local outliers. In *ACM sigmod record*, volume 29, pages 93–104. ACM.

Bryk, J., Hardouin, E., Pugach, I., Hughes, D., Strotmann, R., Stoneking, M., and Myles, S. 2008. Positive selection in east asians for an edar allele that enhances nf-κb activation. *PLoS One*, 3(5): e2209.

Campos, G. O., Zimek, A., Sander, J., Campello, R. J., Micenková, B., Schubert, E., Assent, I., and Houle, M. E. 2016. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data Mining and Knowledge Discovery*, 30(4): 891–927.

Consortium, . G. P. *et al.* 2015. A global reference for human genetic variation. *Nature*, 526(7571): 68.

Cruickshank, T. E. and Hahn, M. W. 2014. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular ecology*, 23(13): 3133–3157.

De Villemereuil, P. and Gaggiotti, O. E. 2015. A new fst-based method to uncover local adaptation using environmental variables. *Methods in Ecology and Evolution*, 6(11): 1248–1258.

Duforet-Frebourg, N., Bazin, E., and Blum, M. G. 2014. Genome scans for detecting footprints of local adaptation using a bayesian factor model. *Molecular biology and evolution*, 31(9): 2483–2495.

Duforet-Frebourg, N., Luu, K., Laval, G., Bazin, E., and 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(4): 1082–1093.

Durand, E. Y., Patterson, N., Reich, D., and Slatkin, M. 2011. Testing for ancient admixture between closely related populations. *Molecular biology and evolution*, 28(8): 2239–2252.

Ewing, G. and Hermisson, J. 2010. Msms: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, 26(16): 2064–2065.

Excoffier, L., Hofer, T., and Foll, M. 2009. Detecting loci under selection in a hierarchically structured population. *Heredity*, 103(4): 285.

Foll, M. and Gaggiotti, O. 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a bayesian perspective. *Genetics*, 180(2): 977–993.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y., *et al.* 2010. A draft sequence of the neandertal genome. *science*, 328(5979): 710–722.

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

Hibbins, M. S. and Hahn, M. W. 2019. The timing and direction of introgression under the multispecies network coalescent. *Genetics*, 211(3): 1059–1073.

Hudson, R. R. 2002. Generating samples under a wright–fisher neutral model of genetic variation. *Bioinformatics*, 18(2): 337–338.

Hudson, R. R., Slatkin, M., and Maddison, W. P. 1992. Estimation of levels of gene flow from dna sequence data. *Genetics*, 132(2): 583–589.

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

Kriegel, H.-P., Schubert, M., and 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*, pages 444–452. ACM.

Kriegel, H.-P., Kröger, P., Schubert, E., and Zimek, A. 2009. Loop: local outlier probabilities. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1649–1652. ACM.

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

Luu, K., Bazin, E., and Blum, M. G. 2017. pcadapt: an r package to perform genome scans for selection based on principal component analysis. *Molecular ecology resources*, 17(1): 67–77.

Martin, S. H., Davey, J. W., and Jiggins, C. D. 2014. Evaluating the use of abba–baba statistics to locate introgressed loci. *Molecular biology and evolution*, 32(1): 244–257.

Pavlidis, P., Jensen, J. D., Stephan, W., and Stamatakis, A. 2012. A critical assessment of storytelling: gene ontology categories and the importance of validating genomic scans. *Molecular biology and evolution*, 29(10): 3237–3248.

Pfeifer, B. and Kapan, D. D. 2019. Estimates of introgression as a function of pairwise distances. *BMC bioinformatics*, 20(1): 207.

Pfeifer, B. and Lercher, M. J. 2018. Blockfest: Bayesian calculation of region-specific fst to detect local adaptation. *Bioinformatics*, 34(18): 3205–3207.

Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S. E., and Lercher, M. J. 2014. Popgenome: an efficient swiss army knife for population genomic analyses in r. *Molecular biology and evolution*, 31(7): 1929–1936.

Ramaswamy, S., Rastogi, R., and Shim, K. 2000. Efficient algorithms for mining outliers from large data sets. In *ACM Sigmod Record*, volume 29, pages 427–438. ACM.

Rambaut, A. and Grass, N. C. 1997. Seq-gen: an application for the monte carlo simulation of dna sequence evolution along phylogenetic trees. *Bioinformatics*, 13(3): 235–238.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. 2011. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC bioinformatics*, 12(1): 77.

Rosenzweig, B. K., Pease, J. B., Besansky, N. J., and Hahn, M. W. 2016. Powerful methods for detecting introgressed regions from population genomic data. *Molecular ecology*, 25(11): 2387–2397.

Schubert, E., Zimek, A., and 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(1): 190–237.

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

Weir, B. 1996. Genetic data analysis ii. sunderland. *MA: Sinauer Associates*, pages 161–173.

Weir, B. S. and Cockerham, C. C. 1984. Estimating f-statistics for the analysis of population structure. *evolution*, 38(6): 1358–1370.

Wright, S. 1949. The genetical structure of populations. *Annals of eugenics*, 15(1): 323–354.

Zhang, K., Hutter, M., and 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*, pages 813–822. Springer.