KLinterSel: Intersection among candidates of different selective sweep detection methods.

Antonio Carvajal-Rodríguez[1*], Sara Rocha[2], Marina Pampín[3], Paulino Martínez[3], Armando Caballero[1]

[1]Centro de Investigación Mariña (CIM), Departamento de Bioquímica, Genética e Inmunología. Universidade de Vigo, 36310 Vigo, Spain

[2]Institute of Marine Research, Spanish National Research Council (IIM-CSIC), Vigo, Spain

[3]Departamento de Zoología, Genética y Antropología Física, Facultad de Veterinaria, Universidade de Santiago de Compostela, Lugo, Spain

*Correspondence address: Centro de Investigación Mariña (CIM), Departamento de Bioquímica, Genética e Inmunología. Universidade de Vigo, 36310 Vigo, Spain. Tel: +34 986130052. E-mail: acraaj@uvigo.es

keywords: selective sweeps, hypergeometric, Monte Carlo test, genomics, *Cerastoderma edule*.

Abstract

Detecting signals of natural selection in genomic data often involves applying several methods in parallel. Regions identified by multiple approaches are usually considered strong candidates, as agreement among methods is often taken as supporting evidence. However, the extent to which such overlaps exceed random expectations is rarely evaluated formally. When genomic elements are not independent, coincident candidate sites may arise from the underlying structure of the data rather than from genuine methodological concordance. To address this problem, we introduce two complementary statistical tests designed to evaluate whether the observed overlap among candidate sites detected by different methods exceeds what would be expected by chance. The first is a fast parametric test based on a sequentially conditioned hypergeometric framework that evaluates k-way intersections among candidate sets across genomic windows. The second relies on Monte Carlo simulations and compares the observed inter-method distance profiles with those expected under random association, taking into account the empirical distribution of SNPs along the genome. By capturing different aspects of concordance, these approaches allow

agreement among methods to be assessed across multiple spatial scales. Both tests are implemented in the program KLinterSel, which also identifies clusters of candidate sites shared by multiple methods within a user-defined distance threshold. We illustrate its application using candidate loci associated with resistance of the common cockle (*Cerastoderma edule*) to the parasite *Marteilia cochillia*. The software is written in Python and is available on GitHub together with documentation and precompiled executables for major operating systems.

Introduction

At the genomic level, natural selection impacts not only alleles with beneficial mutations but also neutral alleles of linked loci, resulting in selective sweeps and a loss of diversity around selected sites [1]. Numerous methods are designed to detect traces of selective sweeps in genomes [2,3]. However, these methods face challenges because similar genomic patterns can arise from different scenarios [3–5]. Some authors [4] argue that the best way to validate candidates would be to simulate data capturing the spectrum of biological processes that could underlie those patterns. These data should then be analyzed with various statistical approaches to evaluate which offers the most reliable results. However, it is not always possible to model the complexity of scenarios (often unknown) that could have given rise to the data. Moreover, there will always be doubts whether enough scenarios have been explored and if the missing ones could yield spurious results. Therefore, a common strategy is to use two or more methods to select the most robust candidates [6]. The problem with this approach is that there is often little overlap between candidates [7], which has led to questioning the reliability of the results obtained through these methods. Additionally, it can be argued that the mere coincidence of results from different statistical analyses does not necessarily validate the analyses due to the patterns of non-independence underlying the biological datasets themselves [8,9].

In this paper, we introduce the software KLinterSel, which implements two complementary statistical tests: the hypergeometric k-intersection (HGkI) and the Kullback-Leibler-like Monte Carlo test (T_{KL}). These tests evaluate whether the spatial coincidence among candidate SNPs detected by different methods exceeds what would be expected by chance

given the genomic configuration of the data. The HGkI test is a fast parametric approach based on a sequentially conditioned hypergeometric framework that evaluates k-way intersections among candidate sets across genomic windows. By contrast, T_{KL} is a Monte Carlo-based method that operates on inter-method distance patterns: it computes all pairwise distances between candidate SNPs detected by different methods and estimates the expected distance profile under a null model by repeatedly permuting the original SNP positions. The observed distance profile is then compared against this expectation.

It should be noted that the tests proposed are not intended to detect selective sweeps directly, nor to validate the biological correctness of candidate loci. Instead, they evaluate whether the spatial coincidence among candidate SNPs detected by different methods exceeds what would be expected by chance given the genomic configuration of the data. Consequently, the approaches are complementary to existing sweep-detection methods rather than a replacement for them.

For both tests, KLinterSel computes p-values that quantify the probability of observing the detected level of spatial coincidence under a null hypothesis of random overlap, thereby providing a statistical framework to evaluate whether the observed concordance among candidate sets is compatible with random expectations.

In addition, the software computes, for any user-specified distance threshold *D*, the size of the intersection among candidate lists restricted to SNPs lying within distance *D* of one another. Both the statistical tests and the intersection analysis can be applied to any number of candidate lists produced by different detection methods. The methodology is detailed below, together with an application example using data from a previous study on selection

for resistance to the common cockle parasite *Marteilia cochillia*. Finally, we evaluate the statistical performance of the proposed tests through analyses of false-positive rates and statistical power under controlled simulation scenarios.

Materials and Methods

Hypergeometric k-way intersection (HGkI) test

Parametric methods provide an explicit probabilistic framework to evaluate whether the overlap among candidate sets exceeds that expected under random sampling. The hypergeometric distribution offers a natural model for this purpose when subsets are drawn without replacement from a finite universe. In genome scans for selection, multiple methods are often applied to the same dataset, each producing a list of candidate loci or genomic regions. Even in the absence of a shared selective signal, some overlap among these candidate sets is expected by chance, particularly when the number of tested loci is large or when individual methods report many candidates. The hypergeometric distribution provides a formal way to assess whether the observed overlap exceeds these expectations. Building on this rationale, we introduce the Hypergeometric k-way Intersection (HGkI) test, which quantifies the probability of the observed overlap among candidates detected by k methods under a null model of random association. As a parametric approach based on a known probability distribution, HGkI is computationally efficient and does not require resampling-based simulations.

Rationale and overview

The HGkI test handles intersections involving an arbitrary number of methods ($k \geq 2$) and accounts for genomic proximity through a window-based construction. For $k > 2$, the null

distribution of the k-way intersection does not have a simple closed-form expression. Instead, it can be constructed through sequential conditioning, yielding a cascade of hypergeometric distributions [10,11]. Similar conditioning-based approaches have been used in other contexts, such as enrichment analyses involving multiple gene sets [12,13], and hypergeometric tests have also been applied to assess genetic repeatability between pairs of independent lineages [14]. Below, we describe the construction of the HGkI statistic for overlapping candidate regions identified by selective sweep scans.

Defining disjoint windows

HGkI operates on a discretized representation of the genome. For each chromosome, the genomic coordinate range under analysis is partitioned into non-overlapping windows of fixed width W (bp), covering the portion of the chromosome represented in the input data. Two cases are distinguished:

(i) SNP-level units ($W = 1$)

When $W = 1$, the test is performed at the SNP level. The universe consists of all SNPs available for analysis in the chromosome, and overlap is defined as exact coincidence of SNP positions across methods. This corresponds to the classical hypergeometric overlap test.

(ii) Window-level units ($W > 1$)

When $W > 1$, chromosomes are partitioned into disjoint genomic windows of size W bp. The window size is user-defined and applied consistently across methods and chromosomes. Only windows containing at least one SNP are retained, as empty windows cannot be selected and would violate the exchangeability assumption of the hypergeometric null model. Accordingly, the total number of units N for a chromosome is defined as the number

of occupied windows.

Mapping candidate lists to genomic windows

Each selection detection method provides, for each chromosome, a list of candidate loci defined by genomic coordinates. In HGkI, each candidate is assigned to the window containing its position. For each method and chromosome, this yields a binary occupancy vector across windows: a window is marked as occupied if at least one candidate from that method falls within it, and multiple candidates within the same window are counted once. Thus, the genomic window becomes the unit of analysis rather than the raw SNP coordinate.

Let $U = \{1, \ldots, N\}$ be the finite universe of units for a given chromosome, where units correspond either to individual SNPs ($W = 1$) or to occupied genomic windows of size W ($W > 1$). For that chromosome, let pos_{min} denote the minimum SNP position in the original dataset. A genomic position pos is assigned to window

$$u = \lfloor \frac{(pos - pos_{min})}{W} \rfloor + 1$$

such that all positions between

$$pos_{min} + (u-1) \cdot W \wedge pos_{min} + u \cdot W - 1$$

belong to window u.

Definition of k-way intersection window

For any combination of k methods, a k-way intersection window is defined as a window occupied by all k methods. The observed number of such windows is denoted by K_{obs}. The HGkI test evaluates whether this observed overlap exceeds that expected by chance.

Hypergeometric statistic: k-way intersection in windows

Under the null hypothesis that candidate genomic units from k different methods are independently distributed across N units (windows, or SNPs when $W = 1$), conditional on their respective numbers $n_1, \ldots, n_k$ (i.e. the number of candidate SNPs for $W = 1$, or the number of windows containing at least one candidate SNP for $W > 1$), the expected overlap is modeled using the hypergeometric distribution.

Methods are internally reordered by decreasing list size to ensure a deterministic and numerically stable evaluation; the resulting null distribution is invariant to the labeling (or ordering) of methods.

Define a sequence of random variables ($K_1, \ldots, K_j, \ldots, K_k$) describing the progressive intersection of methods, with $K_1 = n_1$. For two methods ($j = 2$), the probability of observing exactly $K_2 = x$ overlapping windows follows the classical hypergeometric distribution:

$$P(K_2 = x \mid K_1 = n_1) = \frac{\binom{n_1}{x}\binom{N - n_1}{n_2 - x}}{\binom{N}{n_2}}$$

with $x = 0, \ldots, min(n_1, n_2)$.

For $j > 2$ methods, the construction is extended iteratively. The overlap between the first two methods defines an intermediate set of size K_2, and each additional method is incorporated by conditioning on this intermediate overlap. This yields the joint distribution of the k-way intersection size K_k.

The corresponding probability mass function is:

$$P(K_j = x \mid K_{j-1} = t) = \frac{\binom{t}{x}\binom{N-t}{n_j - x}}{\binom{N}{n_j}}$$

with

$$x = 0, \ldots, min(t, n_j)$$

Global p-value (right-tail test)

The distribution of K_j is obtained recursively by convolution:

$$P(K_j = x) = \sum_{t=0}^{min(n_{j-1}, n_j)} P(K_{j-1} = t)\, P(K_j = x \mid K_{j-1} = t)$$

The recursion is initialized by:

$$P(K_1 = n_1) = 1$$

and uniquely defines the distribution of K_k.

The global p-value for the observed k-way overlap among methods is defined as the right tail of the distribution of K_k:

$$p = P(K_k \geq K_{obs}) = \sum_{x = K_{obs}}^{n_k} P(K_k = x)$$

This provides an exact analytical p-value under the hypergeometric null hypothesis.

Kullback-Leibler-like (T_{KL}) Monte Carlo test

Monte Carlo approaches derive null expectations through empirical resampling rather than from an explicit probabilistic model of overlap. Instead of counting coincident SNPs or windows, T_{KL} characterizes concordance among methods through the distribution of

genomic distances between their reported candidates. The observed distance profile is defined as the set of pairwise distances between candidate loci identified by different methods. For each chromosome, all pairwise distances between candidate sets from different methods are computed, yielding a vector of observed distances, denoted D_o. This vector summarizes spatial proximity among candidates without requiring exact positional coincidence.

To estimate the distance pattern expected by chance, candidate positions are resampled under a null model based on the original SNP dataset. For each method and chromosome, the same number of candidate positions is sampled without replacement from the available SNP positions, preserving both chromosome-specific structure and differences in list sizes among methods. Pairwise distances are then computed for each resampled dataset, producing a vector of simulated distances. This procedure is repeated R times (default R = 10,000, adjustable via the *--perm* argument). For each replicate, distances are sorted in ascending order, and the expected distance profile, denoted D_e, is obtained by averaging these ordered vectors across replicates. Specifically, the mean of each order statistic (smallest to largest distance) is computed across the R replicates. The entire procedure is performed independently for each chromosome.

The computational complexity of estimating D_e is $O(Rk^2n^2)$, where k is the number of methods and n the maximum number of candidate positions per method. To manage computational demands, the program checks available memory and processes permutations sequentially or in parallel as needed. In practice, when n is large, reducing the number of permutations to 1,000 may be necessary; benchmark tests show that results remain virtually unchanged relative to 10,000 permutations.

The test statistic T_{KL} is defined as a Kullback-Leibler-like discrepancy measure between the observed and expected ordered distance profiles. Before computing the statistic, both D_o and D_e are normalized to sum to one, yielding ordered vectors d_0 *and* d_e. This normalization allows the KL-like measure to be computed consistently while capturing differences in the relative weighting of short versus long distances. The test is, therefore:

$$T_{KL}(D_o \| D_e) = \sum d_o \log(d_o / d_e), \quad d = D / \sum D.$$

Note that this is not a Kullback-Leibler divergence in the strict sense, as these vectors do not represent probability distributions, but it provides a consistent basis for comparing distance profiles. Statistical significance is assessed by Monte Carlo resampling, using the same procedure. Specifically, the statistic is evaluated only when the median of the observed distances (M_o) is less than or equal to that of the expected profile (M_e); otherwise, the result is classified as non-significant. This condition ensures that only deviations toward smaller distances, indicating increased spatial coincidence among candidates, are considered relevant, while larger-than-expected distances are not interpreted as evidence against the null hypothesis.

The p-value is obtained by Monte Carlo resampling, comparing the observed discrepancy with those from permuted datasets and applying the Davison and Hinkley [15] correction to account for finite resampling:

$$p = \frac{1 + \sum_{i=1}^{R} I\left(T_{KL}(D_i \| D_e) \geq T_{KL}(D_o \| D_e)\right)}{R+1}$$

where R is the number of permutations and $I(\cdot)$ is the indicator function. For consistency with the test definition, only resampled discrepancies satisfying $M_i \leq M_e$ are considered.

Input data and software usage

Input files

Both HGkI and T_{KL} tests are implemented in the software KLinterSel. When running the program, the user must specify the input files to be analyzed. The first file contains the genomic positions of all analyzed SNPs, whereas the remaining files contain the candidate sites identified by different selection-detection methods. One file is required per method, with a minimum of two methods; thus, at least three input files are needed: one with all SNP positions and two with the candidate SNPs to be compared. The basic format for both genomic positions and candidate files is a simple text file (.csv, .tsv, or .txt) with a header row and two columns: chromosome identifier and genomic position. The program also accepts genomic position files in .map format (without header) and candidate files in .norm format. The number of input files is not limited, except by the available computational resources.

Intersection lists

The results of the HGkI or T_{KL} tests identify chromosomes where the overlap or distance patterns between candidates from different methods are significantly smaller than expected by chance, given the SNP locations in the original dataset. Additionally, we can be interested in identifying groups of candidates from different methods at a specific physical distance $\leq D$ (default D = 10 kb, can be changed with the --*dist* argument). We refer to these sites detected by different methods and lying at a distance $\leq D$, as intersections. Note that if we have k methods and we focus on their intersections, we will have a maximum distance of $(k - 1)D$ within the k-tuple of each intersection. Therefore, to maintain a maximum distance D within the k-tuples, we must define the --*dist* parameter as $D / (k - 1)$.

Software usage

Precompiled binaries are provided for Windows, Linux, and macOS (arm64) and are expected to run on most versions of these operating systems. Alternatively, the program can be installed from the source files available in the GitHub repository. By default, the Monte Carlo T_{KL} test is executed; to run the hypergeometric test, the --*HG* flag can be specified. Detailed information on installation, usage, program options, and file formats, as well as an example dataset (test_dataset.zip), is provided in the program manual and on the project's GitHub repository.

Application to real data

In the results section, we will present an example using published population genomics and transcriptomics data of the common cockle (*Cerastoderma edule*), associated with divergent selection against the parasite *Marteilia cochillia* [16]. For each dataset, we will compare the previously identified set of candidates with three new ones: two methods included in the latest version of selscan [17] and, the J_{HAC} method included in the iHDSel program [18].

Data

We will work with two types of data analysed by Pampín et al. [16]: 6,077 SNPs from RAD-seq (restriction site-associated DNA sequencing) data and 13,004 SNPs from differentially expressed genes (DEGs). The number of SNPs comes from filtering for invariant sites, missing values, and minimum allele frequency (MAF: see Supplementary Files for details on data filtering). Samples exposed and unexposed to the marteiliosis outbreak are compared for RAD-seq and infected and non-infected for DEGs data (see Supplementary Files). The J_{Hac} and selscan methods detect selective sweeps and work best with haplotypes, thus SHAPEIT [19]

with default parameters for a low density array was used to phase both datasets.

Selection detection methods

These methods exploit different statistical signals of selection, including population differentiation and haplotype structure. Pampín et al. [16] inferred candidate SNPs for the RAD-seq data using FDIST implemented in ARLEQUIN v3.5 [20], while for DEGs, F_{ST} was estimated using Genepop 4.7.5 [21], and SNPs with the most extreme F_{ST} values were selected. Both approaches are based on population differentiation, identifying SNPs that show unusually high F_{ST} values compared with neutral expectations. Thus, for each of the two types of data, we have a candidate file, which we will term as Pampín23, derived from these previous analyses (PGCAND and TCAND sites in chromosomes, respectively for RAD-seq and DEGs; see Table 3 in Ref. [16]).

We additionally included three groups of candidates obtained with haplotype-based selective sweep detection methods: XP-EHH [22] and XP-nSL [23], implemented in the selscan program, and J_{HAC} in the iHDSel program. XP-EHH detects regions where one population shows unusually long, high-frequency haplotypes relative to another population, indicative of recent positive selection, whereas XP-nSL uses a related approach based on the number of segregating sites within haplotypes rather than haplotype length. J_{HAC} compares the distribution of haplotype allelic classes (HAC) between populations, defined by their Hamming distance from a reference haplotype; regions under positive selection tend to show reduced haplotypic diversity and an excess of low-distance haplotype classes.

The selscan methods were applied mostly with default parameters (see Supplementary Files). To normalize the obtained scores, the selscan companion program norm was used,

while for J_{HAC}, the automatic block selection with the default parameters was applied.

KLinterSel analysis options

For the HGkI test, rather than selecting a single window size, we interpret results across multiple spatial scales. In addition to exact SNP-level coincidences, which are preserved across all window sizes and are inherently informative when the test indicates a non-random signal, the most robust candidates are those showing consistency across several window sizes, particularly when significance at smaller scales is retained. Accordingly, the HGkI test is applied across a range of window sizes, including windows of size 1 (individual SNPs), 3×10^3, 3×10^4, 10^5, 3×10^5,10^6, 1.5×10^6 and 2×10^6 base pairs.

For the T_{KL} test, we use 10,000 permutations (the default setting) to estimate the expected distance profile. The program provides a *--paint* option to generate histograms of the observed and expected-by-chance distance profiles. To define intersections between the candidates, we use a maximum pairwise distance of 1 kb (*--dist* 1000).

False-Positive Rate Under Controlled Null Scenarios

To investigate the behavior of the HGkI and T_{KL} statistics under the null hypothesis and to assess the effect of SNP spatial distribution along chromosomes, we generated synthetic chromosomes spanning the same genomic range and containing the same number of SNPs as in the empirical dataset described above, but with controlled SNP placement patterns. For a given chromosome length and total number of SNPs, we considered several spatial distributions: (i) uniformly distributed SNPs, (ii) SNPs clustered toward one or both chromosomal extremes, (iii) SNPs clustered toward the center of the chromosome, and (iv) SNPs following the empirical distribution observed in the original dataset. In the clustered

scenarios, SNP positions were generated from skewed distributions that concentrate probability mass toward the corresponding chromosomal regions (left, right, both extremes, or center).

Simulation framework

For each generated SNP universe, candidate positions were randomly sampled in numbers matching those reported by each method, under the null hypothesis of random association among methods. Importantly, the null model is constructed independently for each chromosome using its empirical SNP distribution, both in real-data analyses and in simulations, so chromosome-specific genomic features (e.g., SNP density or spatial clustering) are inherently incorporated into the test, thereby preserving chromosome-specific genomic structure.

In all simulations, explicit genomic positions were retained. For the T_{KL} test, these positions were used directly to compute inter-method distance profiles, whereas for HGkI only the number and genomic allocation of candidates within windows were required.

For the HGkI test, false-positive behavior was evaluated independently for each SNP distribution scenario. For a given scenario, R = 100 independent realizations of the SNP universe were generated. For each universe realization, S = 100 independent sets of candidate loci were sampled for the *k* methods under the null hypothesis. This yielded a total of 10,000 replicates per scenario, from which the distribution of HGkI p-values was obtained and used to assess calibration and false-positive rates. This procedure was repeated across a range of window sizes to evaluate how different spatial definitions of genomic coincidence affect test behavior, including window sizes of 1 (individual SNPs), 3 ×

10^3, 3×10^4, 10^5 , 3×10^5, 10^6, 1.5×10^6 and 2×10^6 base pairs.

For the T_{KL} test, the expected distance profile was first estimated for each scenario using a single realization of the SNP universe (R = 1) and the Monte Carlo procedure described above, based on 10,000 resampling iterations. Once the expected profile had been established, candidate positions for each method were randomly sampled under the null hypothesis, and the corresponding observed distance profile was computed. The probability of this observed profile was then evaluated relative to the expected distribution. This entire procedure was repeated S = 5,000 times per scenario to characterize the false-positive behavior of the T_{KL} statistic under the null hypothesis.

Estimation modes

False-positive rates were estimated using two complementary modes, referred to as *per-chromosome* and *any-chromosome*, which capture different aspects of statistical behavior in multi-chromosome analyses.

Per-chromosome estimation mode

In the per-chromosome mode, the false-positive rate (FPR) was estimated independently for each chromosome: a replicate was counted as a false positive for a given chromosome if the corresponding test statistic yielded a p-value below the chosen significance threshold. This mode measures the probability that a single chromosome produces a significant result under the null hypothesis and therefore provides a chromosome-wise assessment of calibration. The significance threshold was set to $\alpha = 0.05$, independently of the test and spatial scenario.

Any-chromosome estimation mode

In contrast, the any-chromosome mode evaluates false positives at the experiment-wide

level. A replicate was classified as a false positive if at least one chromosome among all those analyzed yielded a significant result. This mode captures the effect of multiple testing across chromosomes and provides an empirical estimate of the family-wise error rate (FWER), reflecting the probability of observing at least one spurious signal anywhere in the genome under the null hypothesis.

Power Analysis Under Controlled Alternative Scenarios

To evaluate the power of the HGkI and T_{KL} statistics under controlled alternative scenarios with non-random spatial concordance among methods, we performed a power analysis using the same simulation architecture adopted for false-positive rate estimation (*R* independent SNP universes × *S* replicate candidate sets per universe). In contrast to the false-positive study, which evaluates calibration under the null hypothesis, the present analysis introduces controlled alternative models that generate structured spatial concordance among methods. These alternative models are applied on top of the different underlying SNP distributions considered in the null scenarios (e.g. uniform, empirical, and related variants). Specifically, we considered a hotspot model, where candidate loci are preferentially concentrated within a given genomic region, and a distance compression model, where candidate loci are accumulated into clusters.

(A) Shared hotspot model (localized concordance)

A genomic hotspot of fixed length L_h (e.g., 1 Mb) was randomly positioned along the chromosome. For each method i with n_i candidates ($i \in [1, k]$), a fraction f of candidates was sampled from SNPs located within the hotspot, whereas the remaining 1 - f candidates were drawn uniformly from the chromosomal background SNP universe. The parameter $f \in [0,1]$

controls the strength of the effect.

This model induces localized regional concordance among methods by concentrating candidates within a shared genomic interval. Consequently, it is expected to increase window-based intersection counts (HGkI) and, to a lesser extent, the frequency of short inter-method distances (T_{KL}). In the present study, we set $f = 0.6$, meaning that 60% of candidates per method were drawn from the shared hotspot, representing a moderate degree of spatial concordance.

Because candidates are always drawn from the empirical SNP universe, the genomic SNP density structure is preserved under both null and alternative models. Thus, detected signals reflect concordance among methods rather than variation in underlying SNP density.

(B) Distance compression model (global distance shift)

For each chromosome, several shared clusters were defined by randomly selecting real SNPs as cluster centers. Around each center, a local pool of SNPs was defined within ±δ base pairs. For each method i with n_i candidates, a fraction φ (= 0.9) of candidates was sampled from the union of these local pools, whereas the remaining candidates were drawn from the chromosomal background. If the local pools contained insufficient SNPs to satisfy the shared fraction, additional candidates were sampled from the chromosome to complete the set.

This model increases spatial clustering and reduces inter-method distances without enforcing exact overlaps among methods. In the present study we used 2 and 8 clusters with δ = 0.25 Mb and δ = 1.8 Mb, respectively.

Comparative structure of the alternative models

The hotspot model induces localized clustering by concentrating a fraction of candidates

within a shared genomic region. In contrast, the distance compression model defines one or more clusters centered on real SNPs and samples candidates from local SNP pools around these centers, with the remaining candidates drawn from the chromosomal background. This approach increases spatial clustering and reduces inter-method distances without enforcing exact overlaps or restricting candidates to a single interval. Because the two models introduce spatial structure through different mechanisms, their parameters are not directly comparable but instead reflect the strength of spatial concentration within each framework. The hotspot model modifies the spatial structure of the genomic background, whereas the compression model acts directly on candidate sets by placing them around shared cluster centers. As a result, HGkI is expected to be more sensitive to hotspot-like alternatives that generate concentrated overlap patterns, while T_{KL} may be more sensitive to compression-like alternatives that shift the distribution of inter-method distances. Together, these models capture distinct forms of spatial concordance among candidate sets and are intended to generate different dependence structures rather than to represent specific biological scenarios.

Simulation framework

For each chromosome scenario, we generated *R* independent SNP universes with predefined spatial distributions (uniform, extreme clustering, central clustering, and empirical), as previously described. For each universe, *S* independent replicate candidate datasets were generated for *k* methods, matching the observed per-method candidate counts for that chromosome. Because the simulation framework used to estimate statistical power is computationally demanding, $R = 20$ realizations of the SNP universe were used. For each realization, $S = 50$ independent sets of candidate loci were sampled for the *k* methods in the

HGkI test.

For the T_{KL} test, the expected distance profile was first estimated for each scenario using a single realization of the SNP universe (R = 1) together with the Monte Carlo procedure described above, based on 10,000 resampling iterations. This procedure was then repeated S = 50 times per scenario to estimate the statistical power of the T_{KL} statistic.

In all cases, explicit genomic positions were retained. For the T_{KL} test, these positions were used directly to compute inter-method distance profiles, whereas for HGkI they were discretized into genomic windows to quantify k-way overlaps.

For each replicate dataset, the statistical test (HGkI or T_{KL}) was applied independently to every chromosome (per-chromosome mode), producing a p-value for each chromosome. A chromosome was counted as successfully detected if its p-value was less than or equal to the significance level ($\alpha = 0.05$). Power was then computed for each chromosome as the proportion of significant chromosome-level tests across all replicates. In this way, power represents the probability that a chromosome under the alternative model is detected as significant in a single chromosome-level test.

Results

Application to real data

RAD-seq data

For the RAD-seq data, a total of 6077 SNPs were studied across 19 chromosomes. Pampín23 identified 153 candidate sites, whereas XP-EHH detected 129, XP-nSL 135, and J_{HAC} 141 candidates. The chromosomes with significant overlapping in any of the two tests are given

in Table 1. Significant overlappings under HGkI are considered if the SNP coincidence is exact or if they appear in at least three windows. Only two chromosomes hold this criteria, chromosome 5, that appeared in all windows beginning from $W = 3 \times 10^3$, and chromosome 18 that has an exact coincidence and correspondingly appeared in every window. For each significant chromosome, the reported p-value corresponds to the smallest window size at which significance was first observed. Significant overlaps under T_{KL} involved chromosomes 1, 2, 6, 7, 12, 15, 18, and 19 (Table 1).

Table 1. Chromosomes from RAD-seq data with significant overlaps between candidates of the four selection detection methods. The significance level is 0.05.

Chromosomes	HGkI p-value	T_{KL} p-value
1	ns	$<10^{-4}$
2	ns	$<10^{-4}$
5	0.0001	ns
6	ns	$<10^{-4}$
7	ns	$<10^{-4}$
12	ns	$<10^{-4}$
15	ns	0.0468
18	0.0008	$<10^{-4}$
19	ns	$<10^{-4}$

ns: non-significant

Chromosome 18 is the only chromosome detected as significant by both tests. Notably, chromosome 5 is consistently detected as significant by the parametric HGkI test but not by T_{KL}. This discrepancy arises because the observed median distance is larger than expected under the null model, leading T_{KL} to classify the chromosome as non-significant.

DEGs data

For the DEGs dataset, a total of 13004 SNPs were analyzed. Pampín23 identified 123 candidate sites, whereas XP-EHH detected 400, XP-nSL 356, and J_{HAC} 339 candidates.

Chromosomes showing significant overlap among methods are reported in Table 2.

No exact SNP-level coincidences were observed. Significant overlaps under the HGkI test were therefore defined as those appearing in at least three windows. Under this criterion, chromosomes 9, 14, 15, 16, and 18 were identified as significant. Chromosomes 9, 14, and 16 were detected consistently from window size $W = 3 \times 10^3$ up to the largest window considered (W = 2 Mb), whereas chromosomes 15 and 18 were detected from $W = 3 \times 10^4$ onward. For each significant chromosome, the reported p-value corresponds to the smallest window size at which significance was first observed. Significants overlappings under T_{KL} involved chromosomes 3, 5, 11, 12, 13, 14, 15, 16, and 18 (Table 2). Chromosomes 14, 15, 16, and 18 were detected as significant by both tests. Chromosome 9 was detected by HGkI but not by T_{KL}. As in the case of chromosome 5 in the RAD-seq dataset, the observed median inter-method distance for chromosome 9 was larger than expected under the null model, leading T_{KL} to classify the chromosome as non-significant.

Table 2. Chromosomes from DEGs data with significant overlaps between candidates of the four selection detection methods. The significance level is 0.05.

Chromosomes	HGkI p-value	T_{KL} p-value
3	ns	0.0006
5	ns	0.0029
9	0.0023	ns
11	ns	0.0013
12	ns	0.0078
13	ns	$<10^{-4}$
14	0.0014	0.0395
15	0.0237	0.0217
16	0.0002	$<10^{-4}$
18	0.0075	$<10^{-4}$

ns: non-significant

False-positive analyses (see next section) showed that HGkI exhibits conservative behavior, with experiment-wide false-positive rates substantially below the nominal expectation across scenarios and window sizes. In this setting, applying an additional multiple-testing correction would be overly conservative and would unnecessarily reduce statistical power. Accordingly, HGkI results are interpreted without further multitest adjustment beyond the intrinsic conservativeness demonstrated by the null simulations.

By contrast, T_{KL} is well calibrated at the experiment-wide level, with the observed family-wise error rate closely matching its theoretical expectation. This indicates that extreme values arise at the expected frequency when scanning multiple chromosomes, thereby supporting the need for explicit multiple-testing control when interpreting genome-wide significance for T_{KL}. For multiple-testing correction of T_{KL} results, any of several available methods and software tools may be applied [24]. In our case, we used the program Myriads1.3 [25]. After correction, either by controlling the false discovery rate (FDR) at 5% or by retaining only *q*-values (estimates of the minimum positive false discovery rate, pFDR) below 5%, all significant results were retained except chromosome 15 for the RAD-seq data and chromosome 14 for the DEGs dataset.

Candidate intersections between methods

For chromosomes showing a significant statistical pattern, the program computes all pairwise intersections between candidate sets from different methods, then intersections involving three methods, and so on, up to the combination of all methods. An intersection distance of 1 kb (--*dist* 1000) with four methods implies that candidate sites may be up to 3 kb apart in some method pairs, while remaining within 1 kb in others, depending on their

relative configuration.

To illustrate the intersection between candidates identified by different methods, we focus on chromosome 18, as it is the only chromosome consistently detected as significant by both statistical tests and across both datasets, providing a robust and unambiguous proof of concept. Complete results, including candidate counts per chromosome and method, distance histograms, and HGkI and T_{KL} test outputs for each chromosome, are provided in the Supplementary Files.

As expected, most intersections occur between XP-EHH and XP-nSL (10 intersections with RAD-seq data and 23 with DEGs). In contrast, the overlap of XP-EHH and XP-nSL with Pampín23 and J_{HAC} is limited, with a maximum of three shared sites (see the DEGs and RAD-seq intersection files in the Supplementary Files). We therefore focus exclusively on candidates present in the intersection of all four methods.

For the RAD-seq dataset, a single SNP located at position 17,284,544 on chromosome 18 is identified. For the DEGs dataset, four sites located at positions 16,983,304, 16,984,143, 16,985,090, and 16,985,155 on chromosome 18 are detected. Chromosome 18 thus shows a significant pattern under both tests and in both datasets and is the only chromosome exhibiting an intersection among all four methods in both datasets. The distance between the RAD-seq and DEGs candidates is approximately 0.3 Mb.

Figure 1 shows the histograms of distance profiles between candidates of the four methods for this chromosome; the left panel shows observed profiles, while the right panel those expected by chance.

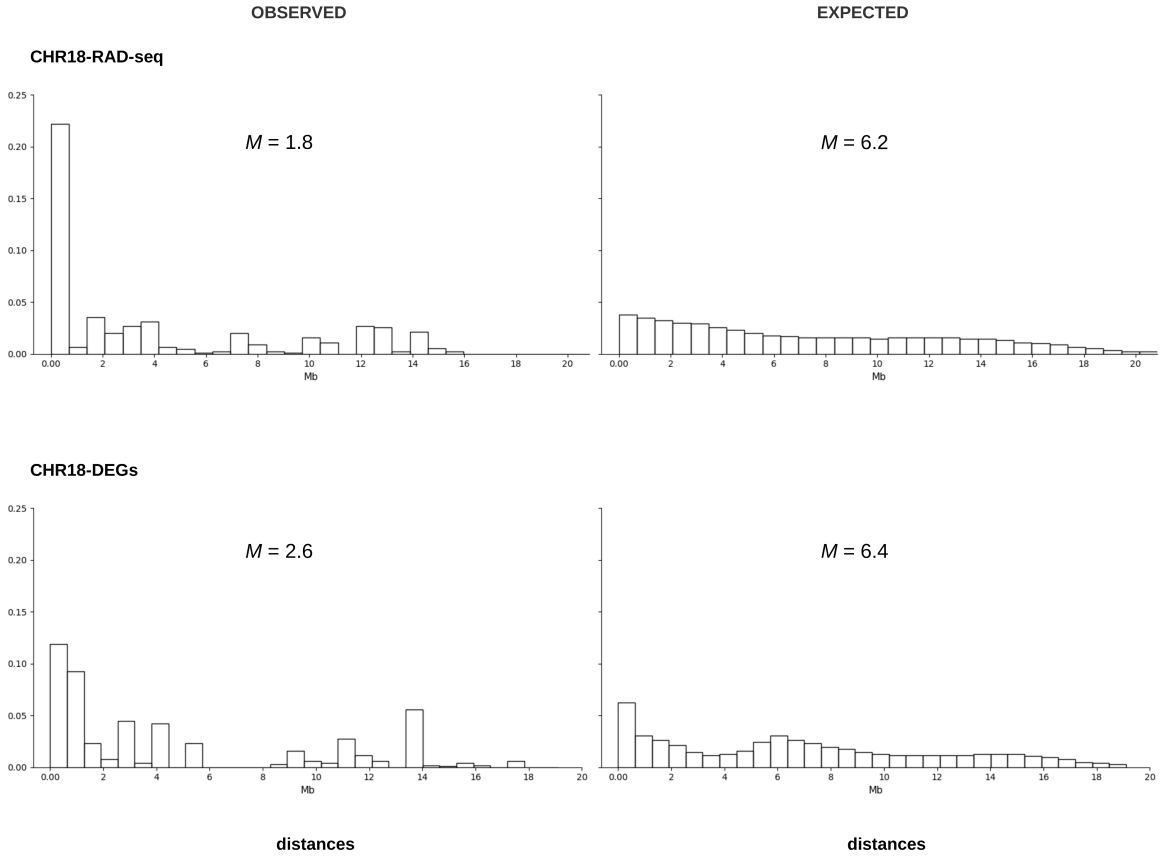

Figure 1. Histogram of observed (left) and expected by chance (right) distances between candidate sites for chromosome 18, obtained using four different selection detection methods. *M*: median measured in Mb units.

The expected median distance exceeds 6 Mb, whereas the observed median falls between 2-3 Mb. For distances of 2 Mb or less, the expected (random) profile would predict a frequency of around 10%. However, a higher frequency of 20-25% is observed (Figure 1). It would be interesting to study this chromosome with a maximum distance of 2 Mb, which, for the distance between pairs, implies $D = 2 \times 10^6 / 3 \approx 667$ kb. Therefore, we now focus only on the intersections of this chromosome (*--notest --dist* 667000 *--chr-id* 18). As expected, the total number of intersections on the chromosome increases, but interestingly they seem to

fall within a relatively short region (see supplementary material for all intersection results). For RAD-seq, the range spans 17,054,167 - 17,734,067 bp, and for DEGs, it is 16,810,598 - 16,985,155 bp. The minimum distance between DEGs candidates and RAD-seq candidates has been reduced to only 69 kb, with a maximum distance of less than 1 Mb. Therefore, this region should be further explored, as well as others that may arise if we study intersections at somewhat longer distances.

False-Positive Rate Under Controlled Null Scenarios

To investigate the false-positive rate of the tests, simulations were parameterized using the SNP distribution observed in the empirical dataset analysed in the previous section. As explained previously we have considered several scenarios regarding the spatial distribution of SNPs: (i) uniformly distributed SNPs, (ii) SNPs clustered toward one or both chromosomal extremes, (iii) SNPs clustered toward the center of the chromosome, and (iv) SNPs following the empirical distribution observed in the original dataset.

False-positive rates for HGkI were consistently below the nominal level across all window sizes and scenarios, indicating conservative behaviour. In contrast, the T_{KL} test showed false-positive rates that closely matched the nominal level in the per-chromosome analysis (Figure 2). Similar results were obtained for the family-wise error rate (FWER) under the experiment-wise decision mode (any-chromosome estimation mode).

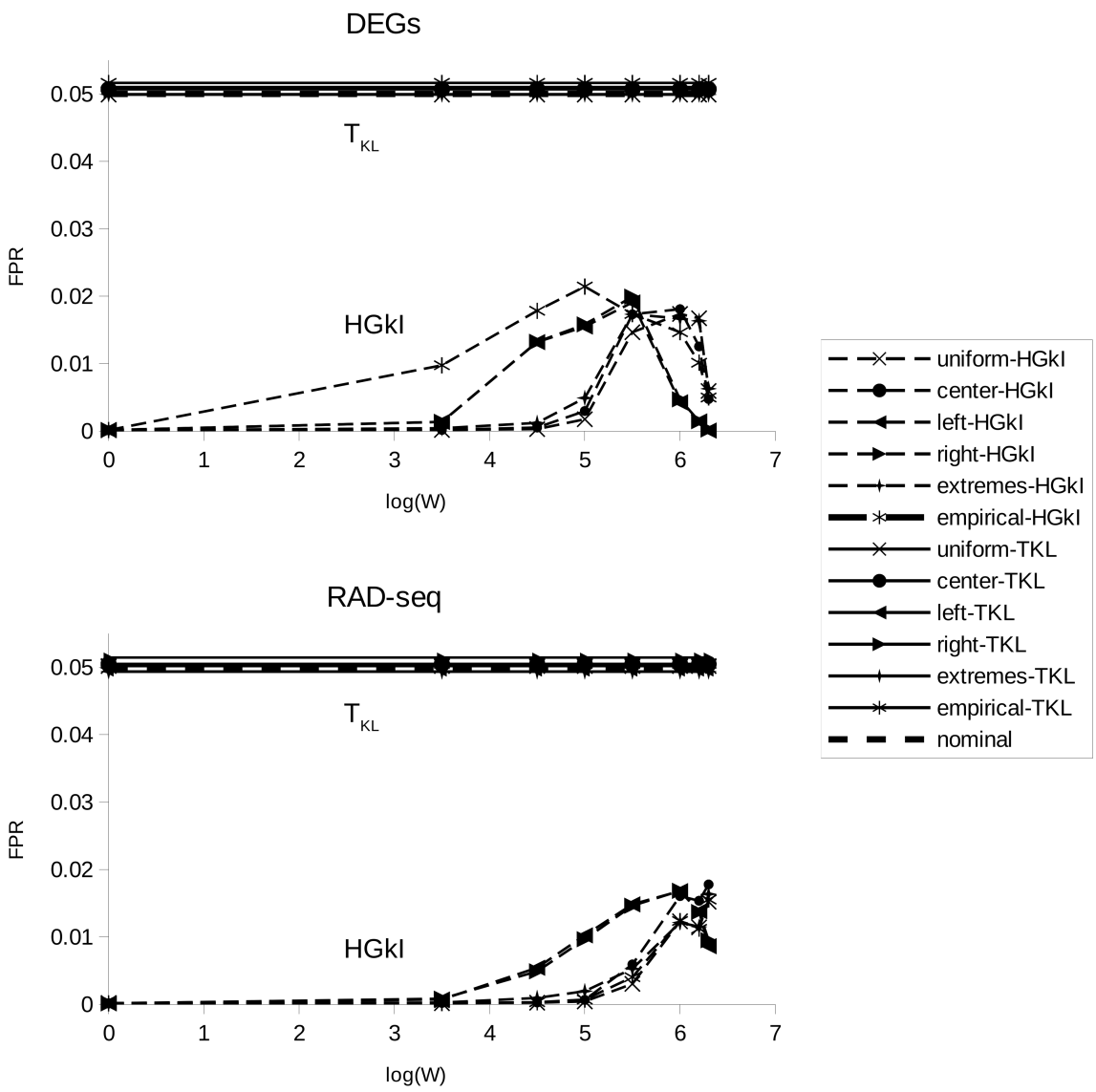

Figure 2. Calibration of the false-positive rates (FPR) of the HGkI and T_{KL} tests under different null scenarios (uniform, empirical, and related variants; see Methods) for RAD-seq and DEG data. The FPR was evaluated across different window sizes (*W*) used in the HGkI test to detect coincident signals between datasets. Each point represents the mean per-chromosome FPR across all chromosomes, calculated as the proportion of simulations in which the test detected a significant overlap between candidate sets. Standard errors were negligible and are not shown. Note that T_{KL} does not depend on window size; therefore, its FPR values are constant across *W* and are shown repeatedly for comparison.

Power Analysis Under Controlled Alternative Scenarios

The results of the power analysis for the hotspot and compression models are shown in Figure 3.

Hotspot and compression models (two clusters)

Under the hotspot model, the HGkI test showed high power (≈0.9–1) for both RAD-Seq and DEG datasets for single SNP windows (W = 1) and for window sizes ranging from 3 kb to 0.3 Mb. For larger windows (≥ 1 Mb), power decreased, particularly for DEG datasets under the uniform, center, and empirical scenarios (Figure 3). This stronger decrease in HGkI power for DEG datasets at large window sizes likely reflects the higher SNP density of these datasets. When window size increases relative to the spatial scale of hotspot signals, the number of genomic windows decreases and the probability that each method occupies a given window increases. Consequently, the expected number of coincident windows under the hypergeometric null model also increases, reducing the contrast between observed and expected overlaps and potentially decreasing the power of HGkI. This effect is weaker in RAD-Seq datasets because their lower SNP density results in fewer occupied windows.

For the same hotspot model, the T_{KL} test (which does not depend on window size) showed greater variability across spatial scenarios. Power was highest (≈0.6–0.8) under the uniform, center, and empirical scenarios, with slightly higher values for DEG datasets.

Under the compression model with two clusters, the pattern for HGkI was reversed: single-SNP windows (W = 1) or small windows showed lower power, whereas larger windows produced much higher power, approaching 1. This occurs because SNPs under selection are distributed across each cluster rather than concentrated at exactly the same positions, larger

windows increase the probability that candidates detected by different methods fall within the same window. In contrast, the power of T_{KL} varied less across spatial scenarios, typically remaining within the range of ≈0.6–0.8 and again showing slightly higher values for DEG datasets (Figure 3).

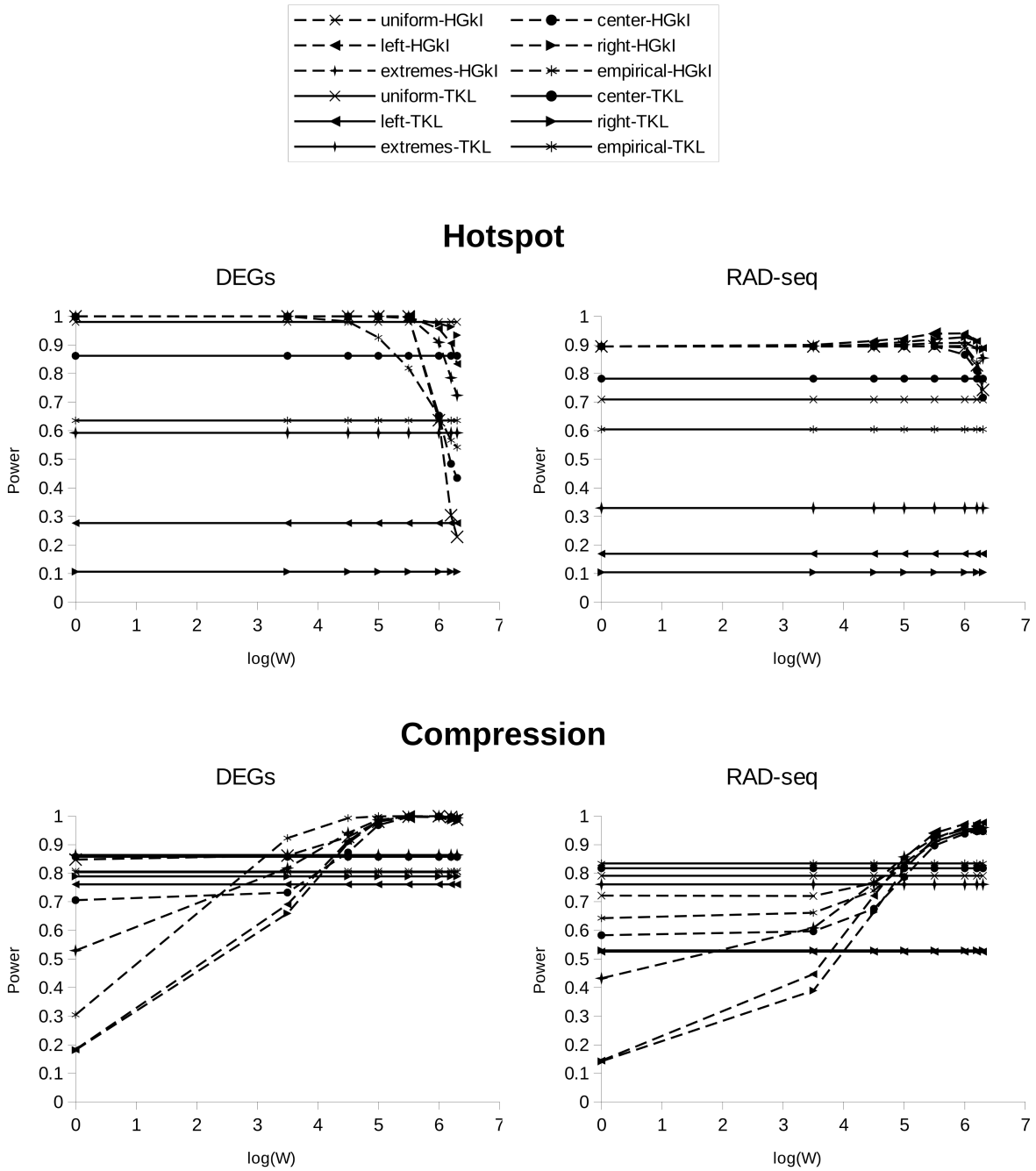

Figure 3. Power analysis of the HGkI and T_{KL} tests for RAD-seq and DEG data under different null scenarios (uniform, empirical, and related variants; see Methods) and two alternative spatial models of SNP clustering (hotspot and compression). In the hotspot model, candidate SNPs are concentrated within a single localized genomic region, whereas in the compression model they are distributed across two localized clusters shared by the different methods. Statistical power was evaluated across window sizes (*W*) used in the HGkI test. Each point represents the mean power across chromosomes, defined as the proportion of simulations in which the test detected significant overlap between candidate sets. Standard errors were negligible and are not shown. Note that T_{KL} does not depend on window size; therefore, its power values are constant across *W* and are shown repeatedly for comparison.

Dispersed signal model: compression with eight clusters

Figure 3 shows that under compression models with two clusters, the power of HGkI increases with window size. This raises the question of how far this trend extends when candidate signals are distributed across a larger number of clusters.

As shown in Figure 4, a model with eight clusters (shared fraction $\phi = 0.9$ and cluster spread $\delta = 1.8$ Mb) applied to a uniform genomic SNP distribution produces dispersed coincidences along the chromosome. Under this scenario, HGkI only detects significant signals for relatively large windows (≈1–4 Mb). In contrast, the T_{KL} test shows intermediate power under the same conditions, but still outperforms HGkI in this dispersed-signal scenario. For both methods, power is generally higher when using the DEG datasets. A similar behaviour was observed under the empirical scenario.

These results suggest that HGkI power depends strongly on the spatial scale defined by the window size, being highest when the window size matches the spatial extent of candidate clustering. In contrast, T_{KL} shows more stable behavior across spatial scenarios because its statistic relies on the spatial distribution of SNPs rather than on window-based coincidence counts.

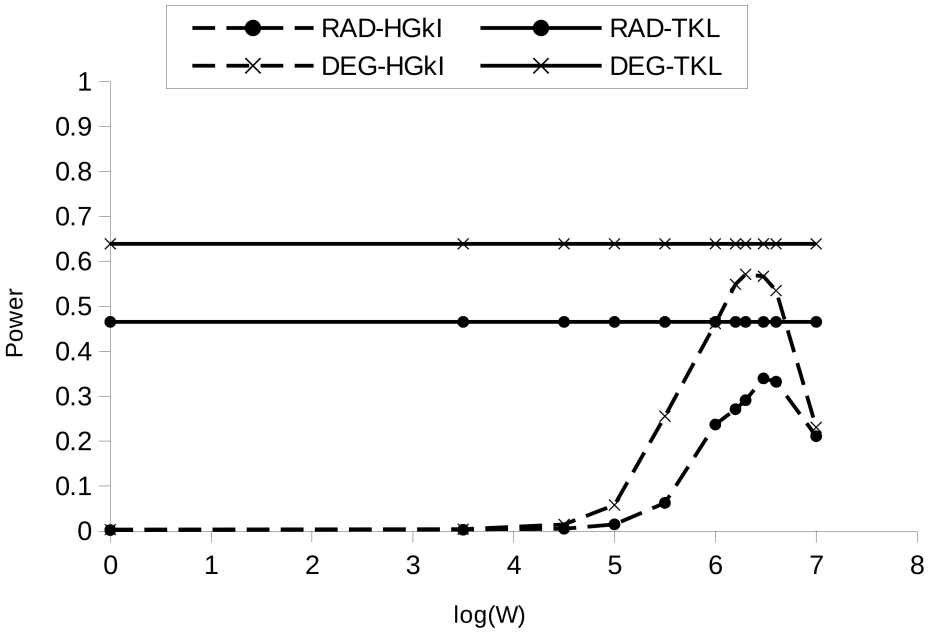

Figure 4. Power analysis of the HGkI and T_{KL} tests for RAD-seq and DEG data. Simulations were performed under a uniform null scenario and under a compression model of spatial SNP clustering with 8 clusters, δ = 1.8 Mb and a shared fraction φ = 0.9, which generates sparse localized concentrations of candidate SNPs along the chromosome. Statistical power was evaluated across increasing window sizes (*W*) used in the HGkI test to explore the saturation of detection power, with *W* ranging up to 10^7. Each point represents the mean power across all chromosomes, calculated as the proportion of simulations in which the test detected a significant overlap between candidate sets. Standard errors were negligible and are not shown. Note that T_{KL} does not depend on the window size; therefore its power values are identical across *W* and are shown repeatedly for comparison.

Effects of chromosome length and SNP density on statistical power

To better understand the factors affecting detection power, we calculated, for both RAD-seq and DEG datasets, the Spearman correlations between statistical power and chromosome length (ρ_1), and between statistical power and the number of SNPs (ρ_2) across chromosomes. Chromosomes differ substantially in length and in SNP number and density, which may influence the performance of spatial tests. Notably, SNP density (mean SNPs per Mb) is approximately two to three times higher in DEG than in RAD-seq datasets.

Because SNP spatial distribution can affect these correlations, we focus on the empirical scenario (Table 3), which is the most informative for interpreting the real data analyzed in

the section with real data. Correlations for all scenarios are provided in the CSV files in the *Results Power Study* folder of the Supplementary Files.

In simulated scenarios where power was close to one across chromosomes (power ≈ 0.99–1), correlations with chromosome length or SNP number were not informative due to a ceiling effect, as the very low variance in power leads to unstable estimates.

Table 3. Spearman correlations between statistical power and chromosome length (ρ_1), and between statistical power and the number of SNPs (ρ_2), across chromosomes for DEGs and RAD-seq under the empirical scenario and hotspot and compression models (2 and 8 clusters). For the HGkI test, results correspond to a window size of 1 Mb.

	Hotspot			Compression-2			Compression-8		
	power	**ρ_1**	**ρ_2**	**power**	**ρ_1**	**ρ_2**	**power**	**ρ_1**	**ρ_2**
RAD-seq HGkI	0.90	0.21	0.20	0.95	0.29	0.30	0.31	0.59	0.64
RAD-seq T_{KL}	0.61	0.48	0.51	0.83	0.44	0.47	0.46	0.69	0.75
DEGs HGkI	0.65	0.60	0.46	1	-0.23	-0.18	0.76	0.72	0.79
DEGs T_{KL}	0.64	0.21	0.46	0.80	0.44	0.65	0.58	0.56	0.64

Hotspot model

For RAD-seq data, both HGkI and T_{KL} showed broadly similar relationships between statistical power and chromosome characteristics across window sizes. Under the hotspot model, correlations between power and chromosome length (ρ_1) and SNP number (ρ_2) were generally moderate and of comparable magnitude, typically ranging between 0.2 and 0.5 for both the empirical (Table 3) and uniform scenarios (Supplementary Files). In DEGs, a contrasting pattern emerged: $\rho_1 > \rho_2$ for HGkI, whereas $\rho_1 < \rho_2$ for T_{KL}. This reflects the different statistical nature of the tests. HGkI is sensitive to absolute spatial scale, showing stronger dependence on chromosome length, whereas T_{KL} relies on the distribution of inter-SNP distances and is therefore largely insensitive to chromosome size but responsive to the number of SNPs. In RAD-seq, the lower SNP density likely makes both factors contribute more similarly.

Compression models

Under compression models, a similar pattern was observed, with the highest correlations in the model with eight clusters. The negative correlations for HGkI with DEGs under the two-cluster compression model likely reflect a ceiling effect, as average power across chromosomes was close to one.

Overall, these results indicate that HGkI correlations largely reflect the geometric partitioning imposed by chromosome length and window size, whereas T_{KL} correlations are more strongly influenced by the amount of spatial information (i.e., SNP density) contained in the SNP distribution.

Discussion

It is common practice to apply several selection-detection methods in parallel when searching for genomic signatures of natural selection, based on the assumption that agreement among methods increases confidence in the inferred candidates [6]. In practice, however, different approaches often show only limited overlap [7], which has raised questions about the reliability of results obtained either from individual methods or from their combined use. Ideally, the methods applied should rely on distinct methodological principles. Yet even conceptually similar approaches frequently yield results that differ substantially, both in the number of candidates detected and in their genomic location.

In this context, concordance among SNPs identified by different selection scans, particularly when they occur in nearby genomic regions, is often interpreted as supporting evidence when prioritizing candidate loci for further investigation. Assessing such concordance, however, is not always straightforward when hundreds of candidates are involved and exact matches are rare. This is not surprising, as different methods may detect signals within the same genomic region or gene without necessarily identifying the same SNP, owing to methodological differences and linkage-drag effects.

To address this issue, we introduced two complementary statistical methods to quantify agreement among SNPs identified by different selection-scan methods: a parametric test

based on the hypergeometric distribution (HGkI) and a non-parametric, permutation-based test (T_{KL}).

The HGkI test identifies coincident SNPs across methods and evaluates whether the observed number of overlaps exceeds that expected by chance. Besides exact SNP-level coincidences, the test can also be applied using genomic windows, within which all SNPs are collapsed into a single spatial unit. Under this construction, the test evaluates whether the number of coincident windows among methods is greater than expected under the null model. Choosing an appropriate window size is not straightforward. Instead of fixing a single spatial scale, we applied the HGkI test across a range of increasing window sizes and considered chromosomes significant when they showed consistent signals across multiple scales. Increasing window size can improve power when true signals are spatially close but not identical, because nearby candidates are collapsed into the same window. However, larger windows also reduce the number of independent spatial units and increase the probability of random overlaps under the null hypothesis. If windows become too large given the SNP density, spatial resolution is lost and the test becomes uninformative.

Like its parametric counterpart, the T_{KL} test evaluates whether the degree of overlap among candidates identified by different methods can be explained by chance given the genomic distribution of SNPs, or whether the observed co-occurrence is unlikely under neutrality and therefore potentially indicative of a genuine selective signal. Unlike HGkI, however, T_{KL} does not focus on discrete overlaps but instead considers the full distribution of inter-method distances between candidate SNPs. SNP positions used in selective sweep analyses are rarely uniformly distributed across the genome. Clustering commonly arises due to linkage disequilibrium, genotyping design, or filtering procedures. Assuming a uniform null model

may therefore produce expected distances larger than those typically observed in empirical SNP sets. T_{KL} addresses this issue by generating expected distance profiles directly from the empirical distribution of SNP positions. This strategy is inherently more conservative, as it provides null expectations that better reflect the underlying genomic structure. Nevertheless, the program also allows expected distance profiles to be generated under a uniform assumption when this is required for comparison with other methods or theoretical expectations.

When analyzing data, a potential pitfall emerges if two methods that are essentially the same, or operate on the same core principles, are inadvertently or deliberately included. In these cases, the identified candidate regions and their locations will likely be very similar. Consequently, the observed overlap might stem from this methodological overlap rather than a true biological signal of selection. To mitigate this, KLinterSel employs a cautious default setting: if two input files contain a comparable number of candidates and at least 95% of the candidates in one file are also present in the other, the subset file is excluded from the analysis.

Two points are worth noting. First, using nearly identical methodologies to confirm selective candidates provides little additional information and is generally not standard practice. Second, even when similar methods are included, applying the test still represents a more conservative strategy than ignoring the issue altogether.

We illustrated the use of KLinterSel by analysing concordance among four selection scans applied to detect candidates for divergent selection in *Cerastoderma edule*. Although XP-EHH and XP-nSL are expected to produce partially similar results, their overlap remained

below 95%, and all four methods were retained. We focused on chromosomes showing significant overlap and distance patterns, and on positions where all methods coincided.

Analyses were performed on both DEG-associated SNPs and RAD-seq SNPs, restricting attention to chromosomes showing significant patterns in both datasets. For more than half of the chromosomes, both HGkI and T_{KL} indicated that the observed agreement could be explained by chance. For the remaining chromosomes, however, concordance patterns were unlikely under the null model, suggesting non-random agreement among methods.

After filtering, a region of ~0.3 Mb on chromosome 18 was identified, although this result should be considered a proof of concept. The use of a stringent coincidence threshold (≤1 kb) is conservative and likely underestimates the number of relevant regions; larger thresholds would increase overlap detection without affecting T_{KL}, which is based on full distance distributions.

HGkI and T_{KL} provide complementary statistical evaluations of concordance. HGkI is a fast parametric test focused on regional overlap counts, whereas T_{KL} is a Monte Carlo approach based on full distance profiles. As a result, T_{KL} can fail to detect chromosomes where local clustering is offset by long-range dispersion, while HGkI remains sensitive to local overlap signals.

False-Positive rate and power

False-positive analyses show that HGkI is consistently conservative, whereas T_{KL} is well calibrated to the nominal significance level. As a result, HGkI maintains strong control of family-wise error rates across chromosomes, while T_{KL} behaves as expected under the nominal level and may require multiple-testing correction in genome-wide analyses.

Simulation scenarios further illustrate that HGkI and T_{KL} respond differently to spatial configurations: HGkI is more sensitive to localized clustering, whereas T_{KL} captures deviations in full distance distributions, making the two approaches complementary for detecting concordance among candidate sets.

Statistical power in both tests increases with the number of SNPs per chromosome, reflecting the greater spatial information available, and remains robust across typical RAD-seq and transcriptomic datasets. Reduced power is mainly observed in cases of very low SNP density or few candidate loci, where spatial information is limited, although this does not affect type-I error due to chromosome-specific null models. Importantly, lack of significant overlap or distance deviation should not be interpreted as absence of selection, but rather as lack of evidence for non-random concordance. In this context, KLinterSel is not intended to detect selection directly, but to provide a complementary framework to assess whether agreement among candidate SNPs exceeds random expectations.

Conclusions

This work illustrates the utility of KLinterSel for assessing statistical agreement among candidate loci identified by different selection-detection methods, here applied to genomic data from *Cerastoderma edule*. The analysis shows that candidates detected by the four methods are significantly closer than expected by chance in several chromosomes, whereas no such pattern is observed in others. Taken together, our results show that HGkI and T_{KL} capture complementary aspects of spatial agreement and respond to different configurations of candidate distribution. This allows concordance to be evaluated across multiple spatial scales without relying on a single, arbitrarily chosen window size.

Importantly, KLinterSel provides a statistical framework to assess whether agreement among candidate sets exceeds random expectations, rather than establishing their biological validity. In addition, the program reports sets of coincident sites (pairs, triples, and quadruples) within any user-defined distance and provides visualization of both observed and expected distance profiles. To our knowledge, no existing tool integrates these capabilities within a single framework. KLinterSel is implemented in Python and combines efficient vectorized computations with optional multi-process parallelism for analyses with high computational demands.

Acknowledgements

This work was supported by the Marine Science Programme (ThinkInAzul) supported by the Ministerio de Ciencia e Innovación and Xunta de Galicia with funding from the European Union NextGenerationEU (PRTR-C17.I1) and European Maritime and Fisheries Fund, Xunta de Galicia (Grupo de Referencia Competitiva, ED431C 2024/22), Ministerio de Ciencia e Innovación (PID2022-137935NB-I00), Centro singular de investigación de Galicia accreditation 2024-2027 (ED431G 2023/07) and ERDF A way of making Europe.

Data Availability

The datasets generated during the current study are available in the Zenodo repository, [https://zenodo.org/records/16912703?preview=1&token=eyJhbGciOiJIUzUxMiJ9.eyJpZCI6IjM2ZTk2NWExLWE1MWMtNDg0My1iMmUwLTI2MmQzMDhkMDAxOCIsImRhdGEiOnt9LCJyYW5kb20iOiI5N2JmY2Q2ZGMwNzAyMjk5ODM4NTg2ZWQ1MjYyNmY1ZiJ9.RTsOg88slO8M82y2Ig_zmA5rH0r7LNmkWukiTTzktxa0Uk2BLCKN0IFBxwB9xirQZWoBXzMmI25bwwkNrNAjqA]

References

1. Stephan, W. Selective Sweeps. *Genetics* **2019**, *211*, 5–13, doi:10.1534/genetics.118.301319.
2. Panigrahi, M.; Rajawat, D.; Nayak, S.S.; Ghildiyal, K.; Sharma, A.; Jain, K.; Lei, C.; Bhushan, B.; Mishra, B.P.; Dutt, T. Landmarks in the History of Selective Sweeps. *Anim Genet* **2023**, doi:10.1111/age.13355.
3. Soni, V.; Johri, P.; Jensen, J.D. Evaluating Power to Detect Recurrent Selective Sweeps under Increasingly Realistic Evolutionary Null Models. *Evolution* **2023**, *77*, 2113–2127, doi:10.1093/evolut/qpad120.
4. Johri, P.; Aquadro, C.F.; Beaumont, M.; Charlesworth, B.; Excoffier, L.; Eyre-Walker, A.; Keightley, P.D.; Lynch, M.; McVean, G.; Payseur, B.A.; et al. Recommendations for Improving Statistical Inference in Population Genomics. *PLOS Biology* **2022**, *20*, e3001669, doi:10.1371/journal.pbio.3001669.
5. Soni, V.; Jensen, J.D. Temporal Challenges in Detecting Balancing Selection from Population Genomic Data. *G3 Genes|Genomes|Genetics* **2024**, jkae069, doi:10.1093/g3journal/jkae069.
6. Horscroft, C.; Ennis, S.; Pengelly, R.J.; Sluckin, T.J.; Collins, A. Sequencing Era Methods for Identifying Signatures of Selection in the Genome. *Briefings in Bioinformatics* **2019**, *20*, 1997–2008, doi:10.1093/bib/bby064.
7. Schlamp, F.; van der Made, J.; Stambler, R.; Chesebrough, L.; Boyko, A.R.; Messer, P.W. Evaluating the Performance of Selection Scans to Detect Selective Sweeps in Domestic Dogs. *Molecular Ecology* **2016**, *25*, 342–356.
8. Lotterhos, K.E.; Moore, J.H.; Stapleton, A.E. Analysis Validation Has Been Neglected in the Age of Reproducibility. *PLOS Biology* **2018**, *16*, e3000070, doi:10.1371/journal.pbio.3000070.
9. Lotterhos, K.E. The Effect of Neutral Recombination Variation on Genome Scans for Selection. *G3 Genes|Genomes|Genetics* **2019**, *9*, 1851–1867, doi:10.1534/g3.119.400088.
10. Feller, W. *An Introduction to Probability Theory and Its Applications*, *Volume 1*; Wiley, 1968; ISBN 978-0-471-25708-0.
11. Larson, H.J. *Introduction to Probability Theory and Statistical Inference*; Wiley series in probability and mathematical statistics: Probability and mathematical statistics; Wiley, 1982; ISBN 978-0-471-05909-7.
12. Tavazoie, S.; Hughes, J.D.; Campbell, M.J.; Cho, R.J.; Church, G.M. Systematic Determination of Genetic Network Architecture. *Nat Genet* **1999**, *22*, 281–285, doi:10.1038/10343.
13. Huang, D.W.; Sherman, B.T.; Lempicki, R.A. Bioinformatics Enrichment Tools: Paths toward the Comprehensive Functional Analysis of Large Gene Lists. *Nucleic Acids Res* **2009**, *37*, 1–13, doi:10.1093/nar/gkn923.
14. Yeaman, S.; Gerstein, A.C.; Hodgins, K.A.; Whitlock, M.C. Quantifying How Constraints Limit the Diversity of Viable Routes to Adaptation. *PLOS Genetics* **2018**, *14*, e1007717, doi:10.1371/journal.pgen.1007717.
15. Davison, A.C.; Hinkley, D.V. *Bootstrap Methods and Their Application*; Cambridge Series in Statistical and Probabilistic Mathematics; Cambridge University Press: Cambridge,

1997; ISBN 978-0-521-57471-6.
16. Pampín, M.; Casanova, A.; Fernández, C.; Blanco, A.; Hermida, M.; Vera, M.; Pardo, B.G.; Coimbra, R.M.; Cao, A.; Iglesias, D.; et al. Genetic Markers Associated with Divergent Selection against the Parasite Marteilia Cochillia in Common Cockle (Cerastoderma Edule) Using Transcriptomics and Population Genomics Data. *Frontiers in Marine Science* **2023**, *10*.
17. Szpiech, Z.A.; Hernandez, R.D. Selscan: An Efficient Multithreaded Program to Perform EHH-Based Scans for Positive Selection. *Molecular Biology and Evolution* **2014**, *31*, 2824–2827, doi:10.1093/molbev/msu211.
18. Carvajal-Rodríguez, A. IHDSel Software: The Price Equation and the Population Stability Index to Detect Genomic Patterns Compatible with Selective Sweeps. An Example with SARS-CoV-2. *Biology Methods and Protocols* **2024**, *9*, bpae089, doi:10.1093/biomethods/bpae089.
19. Delaneau, O.; Zagury, J.-F.; Robinson, M.R.; Marchini, J.L.; Dermitzakis, E.T. Accurate, Scalable and Integrative Haplotype Estimation. *Nat Commun* **2019**, *10*, 5436, doi:10.1038/s41467-019-13225-y.
20. Excoffier, L.; Lischer, H.E.L. Arlequin Suite Ver 3.5: A New Series of Programs to Perform Population Genetics Analyses under Linux and Windows. *Molecular Ecology Resources* **2010**, *10*, 564–567, doi:10.1111/j.1755-0998.2010.02847.x.
21. Rousset, F. Genepop'007: A Complete Re-Implementation of the Genepop Software for Windows and Linux. *Molecular Ecology Resources* **2008**, *8*, 103–106.
22. Sabeti, P.C.; Varilly, P.; Fry, B.; Lohmueller, J.; Hostetter, E.; Cotsapas, C.; Xie, X.; Byrne, E.H.; McCarroll, S.A.; Gaudet, R.; et al. Genome-Wide Detection and Characterization of Positive Selection in Human Populations. *Nature* **2007**, *449*, 913–918.
23. Szpiech, Z.A.; Novak, T.E.; Bailey, N.P.; Stevison, L.S. Application of a Novel Haplotype-Based Scan for Local Adaptation to Study High-Altitude Adaptation in Rhesus Macaques. *Evolution Letters* **2021**, *5*, 408–421, doi:10.1002/evl3.232.
24. Carvajal-Rodríguez, A. Bioinformatics of Correction of Multiple Testing: An Introduction for Life Scientists. In *Bioinformatics and Human Genomics Research*; CRC Press, 2021; pp. 342–363 ISBN 1-00-300592-6.
25. Carvajal-Rodríguez, A. Myriads: P-Value-Based Multiple Testing Correction. *Bioinformatics* **2018**, *34*, 1043–1045, doi:10.1093/bioinformatics/btx746.