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, Monte Carlo test, genomics, *Cerastoderma edule*.

Abstract

Studies aiming to detect signals of selection in genomes often apply multiple methods to increase confidence in their results, typically selecting genomic regions that overlap across approaches. However, such overlap can be misleading when the genomic regions under study are not independent. In these cases, coincident candidates may arise from the structure of the data itself rather than from true methodological robustness. To address this issue, we present a statistical test that compares, for a given set of SNPs, the observed distance profile between candidate sites detected by different methods with the distance profile expected by chance for the same dataset. This test is implemented in the KLinterSel program, which additionally identifies clusters of sites jointly detected by several methods within a user-defined distance threshold. As a proof of concept, we applied KLinterSel to evaluate the overlap among candidates from four selection-detection methods investigating divergent selection associated with resistance to the parasite *Marteilia cochillia* in the common cockle (*Cerastoderma edule*). KLinterSel statistically evaluates and visualizes the agreement between observed and expected-by-chance distance profiles. It uses Python's numerical libraries and vectorized operations for computational efficiency and includes

multi-process parallelization options for memory-intensive datasets. Source code and documentation are available on GitHub (https://github.com/noosdev0/KLinterSel), and pre-built binaries for Windows, Linux, and macOS (arm64) facilitate broad accessibility.

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 propose a complementary approach to validate candidate SNPs obtained with different detection methods. The key idea is to compute the set of pairwise distances between the candidates detected by the different methods, and then generate the expected

set of distances under a null model by repeatedly permuting the original SNP positions. For each permutation, the distances are computed and ordered, and the resulting ordered distance vectors are averaged across permutations to obtain the expected profile of ordered distances under the null hypothesis. If the median of the observed ordered distances is smaller than the median of the expected profile, we compute a test statistic T_{KL}, defined as a KL-like (Kullback–Leibler–like) discrepancy measure between the observed ordered distance vector and its expected counterpart. A p-value for T_{KL} is then obtained by Monte Carlo resampling, using the same procedure employed to generate the expected distances.

In addition, the program 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 KL-like test and the intersection analysis can be applied to any number of candidate lists produced by different detection methods. The methodology is detailed below, along with an application example using data from a previous study on selection for resistance to the common cockle parasite *Marteilia cochillia*.

Materials and Methods

Input data

When calling the program, we should indicate the name of the files to be analyzed, the first of which contains the positions of all analyzed SNPs, and the subsequent files the sites detected by different methods. One file is needed for each detection method, with a minimum of two methods required. Thus, the minimum number of files will be three: the first with all the SNPs and at least two more with the results of the two methods to compare.

The maximum number of files is not limited except by the computer's capacity. Specifications for the file formats can be found in the program's manual.

Observed and expected distance profiles

Once the data are loaded, the program computes all pairwise distances between the candidate positions detected by the different methods, producing the vector of observed distances D_o. To estimate the distance pattern expected by chance, the program generates random candidate positions under a null model based on the original SNP dataset. Specifically, for each method and each chromosome, the program samples (without replacement) the same number of SNP positions as the method originally reported for that chromosome. Once the random positions have been generated, all pairwise distances between the candidate sets of the different methods are computed, producing a vector of simulated distances. This resampling procedure is repeated R times (default R = 10,000; adjustable via the *--perm* argument). For each resampling replicate, the distances are sorted from smallest to largest, yielding an ordered distance vector. The expected distance profile D_e, is then obtained by averaging the ordered vectors across the R replicates: that is, by calculating the mean of the smallest simulated distance across replicates, the mean of the second smallest, and so on, up to the mean of the largest. This produces the expected profile of ordered distances under the null model, given the genomic data structure and the number of candidates reported by each method. The procedure is performed independently for each chromosome.

The computational complexity of computing the expected distance profile is $O(Rk^2n^2)$, where

k is the number of methods and *n* is the maximum number of candidate positions reported by any method. The program checks the available memory to determine whether all permutations can be processed sequentially; if needed, it attempts to switch to multiprocessor mode to divide the workload. In practice, if *n* is large, it may be necessary to reduce the number of permutations to 1,000 depending on the available hardware. Tests performed show negligible differences between using 1,000 and 10,000 permutations.

Results

Kullback-Leibler-like (T_{kl}) test

If the median M_o of the observed ordered distances is less than or equal to the median M_e of the expected ordered distances under the null model, we compute a KL-like discrepancy measure between the observed and expected ordered distance vectors, denoted $T_{KL}(D_o \,||\, D_e)$:

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

Before computing this KL-like statistic, both the observed and expected ordered distance vectors (*D*) are normalized to sum to one, resulting in normalized profiles (*d*) indexed by their ordering position. This normalization allows the KL-like measure to be computed in a mathematically well-defined manner while still capturing differences in the relative weighting of small versus large distances. Note that this is not a Kullback–Leibler divergence in the strict information-theoretic sense, as these vectors are not probability distributions. Although the normalized vectors do not represent probability distributions of distances in the usual sense, they provide a consistent basis for comparing the observed and expected

distance profiles. The p-value of this statistic is obtained by Monte Carlo resampling, comparing the observed discrepancy with the discrepancies obtained from each permutation of the SNP dataset:

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

where R is the number of permutations (default 10,000), I(-) is the indicator function, equal to 1 when the condition is true and 0 otherwise.

In addition, for consistency with the test logic, only those resampled discrepancies $T_{KL}(D_i \,||\, D_e)$ for which the median of the resampled distances satisfies $M_i \leq M_e$ are considered.

Intersection lists

The results of the T_{KL} test 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).

Bias due to reduntant methodologies

A potential issue arises when, either by oversight or lack of awareness, two selection-

detection methods that are very similar, i.e., based on essentially the same logic, are used together. In such cases, it is expected that the reported candidates will be similar in both number and genomic location. This can produce an apparent excess similarity that does not reflect true selective signal, but rather the redundancy or high correlation between the methods. In practice, however, the opposite situation is more common: even methods that are conceptually similar tend to yield results that differ substantially both in number and in genomic position.

In any case, KLinterSel adopts a conservative strategy by default. If two input files contain a similar number of candidates and at least 95% of the candidates in one file are included in the other, the file representing the subset is removed from the analysis. This behaviour is applied only when both files contain at least five candidates each. The user may disable this option using *--permissive*, in which case no files are excluded even under extreme overlap.

Example of application

Below, 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* [10]. 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 [11] and, the J_{HAC} method included in the iHDSel program [12].

Data

We will work with two types of data analysed in Pampín et al. [10]: 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 [13] with default parameters for a low density array was used to phase both datasets.

Selection detection methods

Pampín et al. [10] inferred candidate SNPs for the RAD-seq data using FDIST implemented in ARLEQUIN v3.5 [14], while for DEGs, F_{ST} was estimated using Genepop 4.7.5 [15], and SNPs with the most extreme F_{ST} values were selected. 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, Table 3 [10]). We will add three more groups of candidates derived from three methods outlined before for detecting selective sweeps in haplotypes, XP-EHH [16] and XP-nSL [17], implemented in the selscan program, and J_{HAC} in the iHDSel program. 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 options

We use 10,000 permutations (default) to calculate the expected distance profile. The program has a --*paint* option that allows generating histograms of the observed and

expected-by-chance distance profiles (Figure 1). To define intersections between the candidates we will use a distance of 1 kb (--*dist* 1000) as the maximum distance between pairs.

Pattern of distances

For the RAD-seq data, 6077 SNPs were studied across 19 chromosomes. Pampín23 identified 153 candidates, XP-EHH 129, XP-nSL 135, and J_{HAC} 141. Chromosomes 1, 2, 6, 7, 12, 15, 18, and 19 showed significantly different patterns from random expectations. For the DEGs data, 13004 SNPs were studied. Pampín23 identified 123 candidates, while XP-EHH 400, XP-nSL 356, and J_{HAC} 339. Chromosomes 3, 5, 11, 12, 13, 14, 15, 16, and 18 showed significantly different patterns.

It is worth noting that we performed 19 T_{KL} tests for each dataset, and therefore it makes sense to consider a correction for multiple comparisons, which can be applied using any of the many available methods and software tools [18]. In our case, we used the program Myriads1.3 [19], and after correction, either controlling the FDR at 5% or accepting only q-values (estimates of the minimum pFDR at which each test would be considered significant) below 5%, all significant results were retained except for chromosome 15 for the RADs and chromosome 14 for the DEGs.

Comparing both datasets (RAD-seq and DEGs), common chromosomes are 12, 15 (before multiple testing correction) and 18 (see the histograms of the distance profiles for chromosome 18 in Figure 1). Now, we focus on intersections between candidates from different methods. All information, including candidate numbers per chromosome and

method, histograms, and the results for the T_{KL} test for each chromosome, is available in the Supplementary Files.

Candidate intersections between methods

For chromosomes with a significant distance pattern, the program calculates all intersection pairs between candidate methods, then groups of three, and so forth, up to the combination of all methods. An intersection distance of 1 kb (--*dist* 1000), with 4 methods, implies that candidates may be within 3 kb from each other in some pairs of methods, but no more than 1 kb apart in others. As proof of concept, we focus only on those chromosomes common to both data types: 12, 15 and 18. As expected, most intersections occur between XP-EHH and XP-nSL (20, 4 and 10 intersections with RAD-seq for chromosomes 12, 15 and 18, respectively, while 17, 28 and 23 for the same chromosomes with DEGs). The overlap of XP-EHH and XP-nSL with Pampín23 and J_{HAC} is low (maximum 3 sites, see DEGs and RAD-seq intersection files in Supplementary Files). We focus exclusively on candidates present in the intersection of all four methods. For RAD-seq, only one SNP at position 17,284,544 on chromosome 18 is found. For DEGs, sites 16,983,304, 16,984,143, 1,985,090 and 16,985,155 also on chromosome 18 are identified. Chromosome 18 shows a significant distance pattern with both datasets and is the only one with an intersection among all four methods in both datasets. The distance between RAD-seq and DEGs candidates is 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.

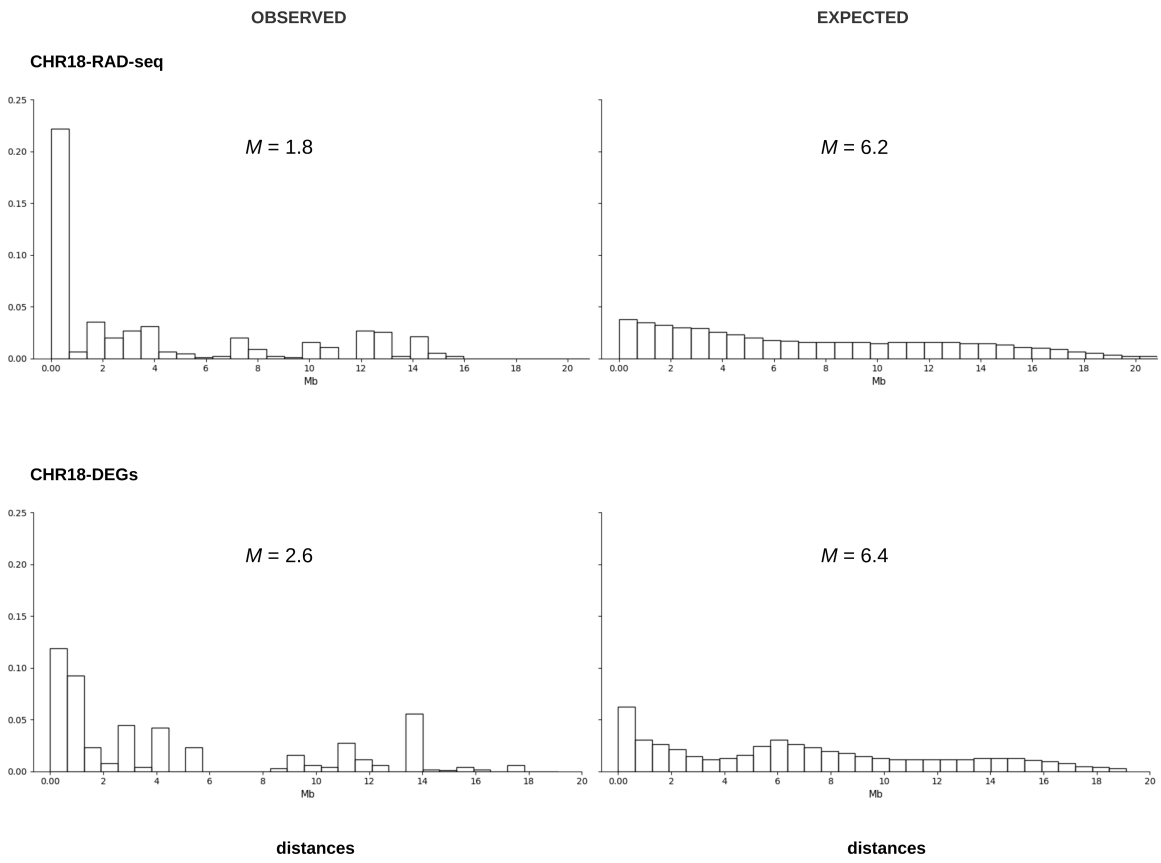

Figure 1. Histogram of observed (left) and expected by chance (right) distances between candidate sites for the chromosome 18, obtained using four different selection detection methods. *M*: median measured in Mb units.

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 $D = 2\times10^6 / 3 \approx 667$ 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.

Discussion

It is common practice to use multiple selection-detection methods in parallel to identify patterns associated with the action of natural selection on genomes, under the assumption that agreement among methods increases confidence and robustness of the inferred candidates [6]. The difficulty with this approach is that, in practice, different methods often show very limited overlap [7], which has led to questioning the reliability of results obtained through individual or combined methods. Ideally, the methods applied should rely on different methodological principles; however, as noted above, even conceptually similar methods frequently produce results that differ substantially both in number and in genomic location. In this context, agreement between SNPs identified by different selection-scan methods, particularly when they fall in nearby genomic regions, is commonly used as

complementary evidence to support their reliability as candidate selective markers. Nevertheless, such agreement is not always straightforward to assess when hundreds of candidates are involved and exact matches are rare, which is not surprising, as different methods often detect signals within the same genomic region or gene but not necessarily the exact SNP, due to differences in methodology and linkage-drag effects.

T_{KL} test

The T_{KL} test, implemented in the program KLinterSel, thus adds a conservative step to the strategy of selecting candidates based on multiple selection-detection methods: it tests whether the degree of overlap among the candidates identified by different methods is expected by chance, given the empirical distribution of SNPs in the genome, or whether the observed co-occurrence is unlikely under neutrality and therefore potentially indicative of a genuine selective signal. Furthermore, SNP positions used for selective sweep detection are rarely uniformly distributed across the genome. Clustering tends to arise due to linkage disequilibrium, recombination landscapes, genotyping design, or filtering procedures. Using a uniform null model may therefore lead to expected distances that are larger than those typically present in the empirical SNP set. By contrast, KLinterSel's default strategy, i.e., generating expected distance profiles directly from the empirical distribution of SNP positions, is inherently more conservative, as it mitigates this discrepancy and provides a null expectation that more faithfully reflects the underlying genomic data structure. Nonetheless, the program also allows the user to generate the expected distribution under a uniform assumption when this is preferred or necessary for comparison with other methods or theoretical expectations.

Importantly, we are not proposing a new selection-detection method. Rather, we provide an *additional step*, applied after running several existing selection-scan methods (a very common practice), to statistically test whether the observed overlap between their candidate SNPs could be explained by chance. The program also produces the empirical distance distributions of the original SNP data and provides explicit intersection sets among methods for any user-defined distance. If one prefers exact matches only, one can simply set the distance to zero.

Analysis of real data

We illustrated the utility of KLinterSel by analysing the concordance among four different selection scans applied to genomic data to detect candidates for divergent selection in *Cerastoderma edule*. Using the physical map as a proxy for the genetic map, two of the methods (XP-EHH and XP-nSL) are expected to yield somewhat similar results; however, their overlap remained below 95%, and thus both files were retained. The other two methods were based on entirely different logic. We then focused on chromosomes showing significant distance patterns and containing SNPs identified by all four methods.

Furthermore, we repeated the analysis using two distinct types of data: SNPs located in differentially expressed genes, and anonymous SNPs obtained via restriction site-associated DNA sequencing. We restricted our attention to chromosomes displaying significant patterns under both datasets, as well as positions where all four selection-detection methods coincided in both data types. The T_{KL} test showed that, for more than half of the chromosomes, the observed agreement among the different methods could be explained by

chance given the empirical SNP distributions. In the remaining chromosomes, however, the observed distance patterns were unlikely to arise by chance.

After all filtering steps, we ultimately identified a specific region on chromosome 18 of approximately 0.3 Mb. However, the present study should be considered a proof of concept. Among the various restrictions applied, perhaps the most stringent is the definition of coincident SNPs as those separated by 1 kb or less, a threshold that is usually smaller than the average gene length and, independently of linkage considerations, is objectively conservative. Using more realistic values, such as 10 kb or greater, would substantially increase the number of overlapping patterns and regions worth investigating. Importantly, the choice of this distance threshold does not affect the T_{KL} test itself, which compares complete distance profiles and is therefore independent of the distance D used to define overlapping regions among methods.

As a final remark, it is important to note that a chromosome whose distance pattern is not significantly different from that expected by chance does not imply that its candidate sites are not under selection. Rather, it means that the coincidence or spatial proximity of candidates detected by different methods cannot, in that case, be taken as evidence of reliability or robustness when prioritizing those sites as stronger candidates.

Limitations of the method

A potential issue may arise when applying this statistic, particularly when two methods that are extremely similar, or effectively based on the same underlying logic, are used together, whether intentionally or inadvertently. In such cases, the number and genomic positions of

the reported candidates are expected to be highly similar. This can create an apparent excess of agreement that does not reflect a true selective signal but rather methodological redundancy or strong correlation between methods. To prevent this, KLinterSel adopts a conservative default rule: if two input files contain a similar number of candidates and at least 95% of the candidates in one file are contained in the other, the subset file is removed from the analysis. Two considerations are relevant here. It makes little sense to use highly similar methodologies to confirm or reinforce selective candidates, and, in fact, this is not standard practice. Moreover, even when similar methods are used, applying the test remains more conservative than not applying it at all.

Concluding remarks

This work illustrates the utility of the KLinterSel program for assessing the agreement among four different selective analyses applied to genomic data to detect candidates for divergent selection in *Cerastoderma edule*. KLinterSel shows that candidates identified by the four methods are significantly closer than expected by chance for several chromosomes, while no such pattern is observed for others. In addition, it reports the quadruples, triples, and pairs of coinciding sites within any user-defined distance, and it can plot both the observed and expected-by-chance distance profiles. To our knowledge, no existing tool integrates these capabilities within a single framework. KLinterSel leverages Python's numerical library and its vectorization options to perform computations efficiently, and it also includes multi-process parallelism options for analyses with high memory 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 Zotero repository, [https://shorturl.at/HwfoC]

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. 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*.
11. 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.
12. 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.
13. 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.
14. 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.
15. Rousset, F. Genepop'007: A Complete Re-Implementation of the Genepop Software for Windows and Linux. *Molecular Ecology Resources* **2008**, *8*, 103–106.
16. 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.
17. 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.
18. 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.
19. Carvajal-Rodríguez, A. Myriads: P-Value-Based Multiple Testing Correction. *Bioinformatics* **2018**, *34*, 1043–1045, doi:10.1093/bioinformatics/btx746.