KLinterSel: Intersection among candidates of different selective sweep detection methods

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

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

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

*To whom correspondence should be addressed.

keywords: selective sweep candidates, Kullback-Leibler divergence, permutation test, genomics, *Cerastoderma edule*.

Abstract

Motivation: Studies that detect signals of selection in genomes often employ more than one method to increase the reliability of their findings. In fact, a usual recommendation for these studies is to apply different methodologies and select overlapping genomic regions. However, in many cases, this may be misleading because the regions under study are not independent. The overlap between candidate regions may not be due to the robustness of the methods, but rather to the structure of the data itself.

Results: This paper contributes to solving this issue by developing a test that compares, for a set of SNPs, the observed distribution of distances between candidates from different methods with that expected by chance for the same SNP dataset. The test is incorporated into the KLinterSel program, which also calculates clusters of sites that intersect between different methods within a given distance. As a proof of concept, we applied the program to compare the candidates of four selection detection methods applied to the study of

divergent selection against the parasite *Marteilia cochillia* in common cockle (*Cerastoderma edule*).

Availability and implementation: Source code and documentation is hosted at GitHub (https://github.com/noosdev0/KLinterSel). For accessibility, KLinterSel has pre-built binaries for Windows, Linux and macOS (arm64).

Contact

acraaj@uvigo.es

Supplementary information

Links to additional methods/data available on Zenodo, or reference to online-only Supplementary methods and data available at the journal's web site.

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 (Stephan, 2019). Numerous methods are designed to detect traces of selective sweeps in genomes (Panigrahi *et al.*, 2023; Soni *et al.*, 2023). However, these methods face challenges because similar genomic patterns can arise from different scenarios (Johri *et al.*, 2022; Soni *et al.*, 2023; Soni and Jensen, 2024). Some authors (Johri *et al.*, 2022) 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 (Horscroft *et al.*, 2019). The problem with this approach is that there is often little overlap between candidates (Schlamp *et al.*, 2016), 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 (Lotterhos *et al.*, 2018; Lotterhos, 2019).

In this paper, we propose a complementary alternative to validate candidates obtained with different methods. The central idea involves calculating the distribution of distances between candidates from different methods, then generating the expected distribution of random distances through permutations of the original SNP dataset. If the observed median distance is less than or equal to the expected (random) distribution median, the p-value of the Kullback-Leibler divergence between the observed and expected distributions is calculated. Furthermore, the program also calculates, for a given distance *D*, the intersection between candidates from different methods that are within that distance. Both the test and the intersection calculation can be applied to any number of candidate lists obtained by the corresponding detection methods. The methodology is detailed below along with an application example to data from a previous study on selection to resistance to the common cockle parasite *Marteilia cochillia*.

Implementation

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 distribution of distances

Once the data are loaded, the program calculates all pairwise distances between the candidate positions of the different methods to obtain the observed distribution of distances (D_o). The program then calculates the distribution of distances expected by chance (D_e): first, the positions associated with each method are randomly calculated; for example, if a method has three candidate positions on chromosome 1, three positions are sampled from the original data for that chromosome. This is done for each chromosome and method. Once we have all random positions from each method, all pairwise distances between the positions of the different methods are calculated. This process is repeated *R* times, with a default *R* = 10,000, although this number can be changed using the --*perm* argument. Next, we calculate across the *R* distributions the mean of the distances, with distances within each distribution ordered from smallest to largest. Specifically, we calculate the mean of the values corresponding to the smallest distance in each distribution, followed by the mean of the values for the next smallest distance, and so on, up to the mean of the

values corresponding to the largest distance. This process produces the distribution expected by chance, given the original data and the number of candidates for each method. This is done for each chromosome.

The computational complexity associated with calculating the expected distribution of distances is $O(Rk^2n^2)$, where k is the number of methods and n is the maximum number of significant findings in any of the methods. The program calculates the memory available to process all permutations and, if necessary, attempts to switch to multiprocessor mode, dividing the workload when possible. Anyway, if n is high, depending on the computer, it may be necessary to reduce the number of permutations to 1,000. Tests conducted barely show differences when using 1,000 or 10,000 permutations.

Kullback-Leibler (KL) test

If the median M_o of the observed distance distribution is less than or equal to the median of the expected random distribution M_e, we calculate the Kullback-Leibler divergence between the observed and expected distributions $KL(D_o \,||\, D_e)$. The p-value of that divergence is obtained by comparing the divergence obtained with the one corresponding to each sample of the expected distribution such that

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

where R is the number of permutations (by default 10,000), I is the

indicator function which is 1 when the condition is true and 0 otherwise, and where the expected divergences KL(D_i || D_e) are considered only for those that satisfy that their median is less than or equal to the expected median, $M_i \leq M_e$.

Intersection lists

The results of the KL test provide information on chromosomes containing matches or distance patterns between candidates from different methods that are significantly smaller than those expected by chance. Additionally, we are interested in identifying groups of candidates from different methods at a physical distance ≤ *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 ≤ *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).

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* (Pampín et al., 2023). 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 (Szpiech and Hernandez, 2014) and, the J_{HAC} method included in the iHDSel program (Carvajal-Rodríguez, 2024).

Data

We will work with two types of data analysed in Pampín et al. (2023): 6,077 SNPs from RAD-seq 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 (Delaneau *et al.*, 2019) with default parameters for a low density array was used to phase both datasets.

Selection detection methods

Pampín et al. (2023) inferred candidate SNPs for the RAD-seq data using FDIST implemented in ARLEQUIN v3.5 (Excoffier and Lischer, 2010), while for DEGs, F_{ST} was estimated using Genepop 4.7.5 (Rousset, 2008), 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 (Pampín *et al.*, 2023)). We will add three more groups of candidates

derived from three methods outlined before for detecting selective sweeps in haplotypes, XP-EHH (Sabeti *et al.*, 2007) and XP-nSL (Szpiech *et al.*, 2021), 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 distribution. The program has a --*paint* option that allows generating histograms of the observed and expected-by-chance distributions (Figure 1). To define intersections between the candidates we will use a distance of 1 kb (--*dist* 1000) as the maximum distance between pairs.

Results

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, 18, and 19 showed significantly different patterns from random expectations. Chromosome 15 was marginally significant (p-val = 0.07). 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, 15, 16, and 18 showed

significantly different patterns. Comparing both datasets (RAD-seq and DEGs), common chromosomes are 12, 18, and marginally 15 (see the distribution 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, and the results for the 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 and 18. As expected, most intersections occur between XP-EHH and XP-nSL (20 and 10 intersections with RAD-seq for chromosomes 12 and 18, respectively, while 17 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 material). 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, two sites -16,983,304 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 with both datasets. The minimum distance between RAD-seq and DEGs candidates is 0.3 Mb. Figure 1 shows the distance distribution between candidates of the four methods for this chromosome; the left panel shows observed distributions, while the right panel those expected by chance.

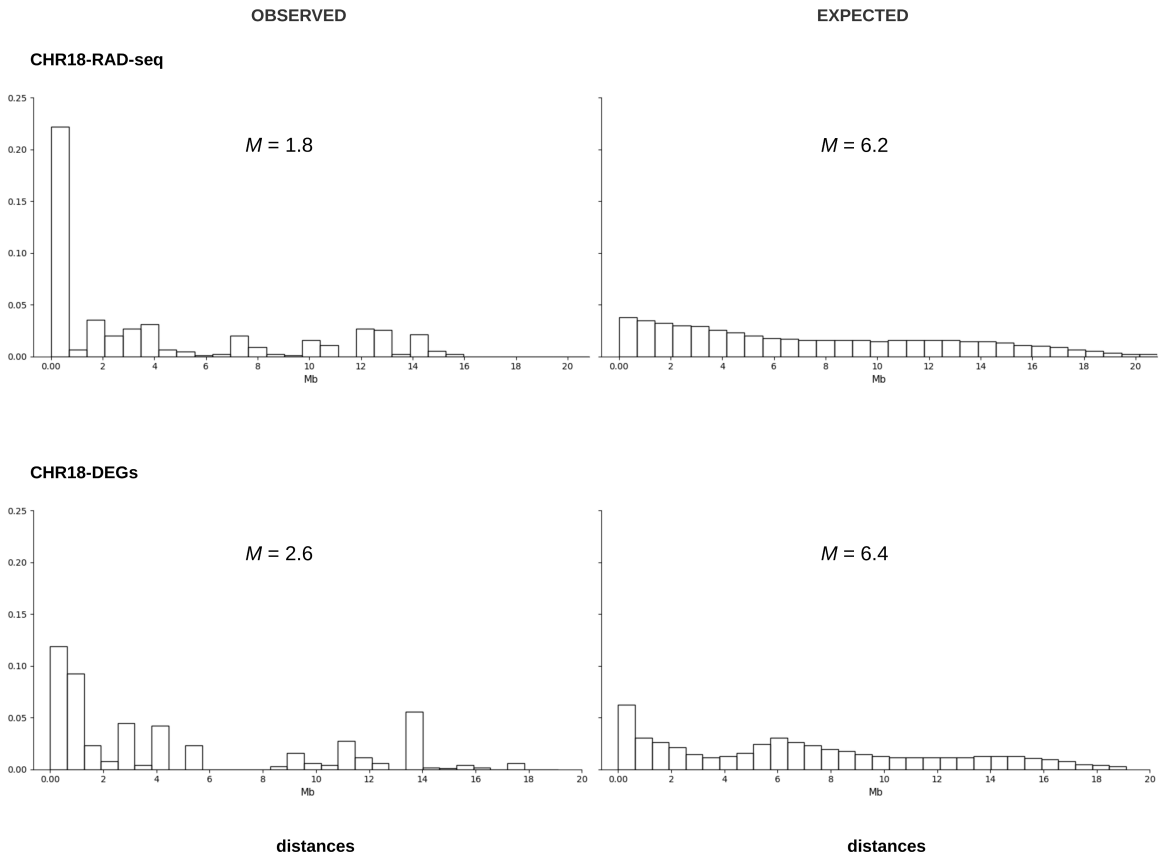

Figure 1. Histogram of observed (left) and expected by chance (right) distances between

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) distribution 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.

Conclusion

This work shows the utility of the KLinterSel program to study the agreement of four different selective analyses on genomic data to detect candidates for divergent selection in *Cerastoderma edule*. By equating the

genetic map to the physical map, two of the methods follow similar approaches (XP-EHH and XP-nSL) and coincident results were expected; the other methods use different strategies to detect the effect of selection. KLinterSel shows how the candidates obtained by the four methods are significantly closer than expected by chance for several chromosomes, but not for others. Additionally, it provides the quadruples, triples, and pairs of coinciding sites for a desired distance. The program can plot the observed and expected-by-chance distributions. KLinterSel utilizes Python's numerical library and its vectorization options to perform calculations efficiently. For cases with high memory consumption, it also includes multi-process parallelism options.

Acknowledgements

Supplementary data

Raw results files are available at Zenodo.

Funding

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. Funding for open access charge: Universidade de Vigo/CISUG.

References

Carvajal-Rodríguez,A. (2024) 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*, **9**, bpae089.

Delaneau,O. *et al.* (2019) Accurate, scalable and integrative haplotype estimation. *Nat Commun*, **10**, 5436.

Excoffier,L. and Lischer,H.E.L. (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, **10**, 564–567.

Horscroft,C. *et al.* (2019) Sequencing era methods for identifying signatures of selection in the genome. *Briefings in Bioinformatics*, **20**, 1997–2008.

Johri,P. *et al.* (2022) Recommendations for improving statistical inference in population genomics. *PLOS Biology*, **20**, e3001669.

Lotterhos,K.E. *et al.* (2018) Analysis validation has been neglected in the Age of Reproducibility. *PLOS Biology*, **16**, e3000070.

Lotterhos,K.E. (2019) The Effect of Neutral Recombination Variation on Genome Scans for Selection. *G3 Genes|Genomes|Genetics*, **9**, 1851–1867.

Pampín,M. *et al.* (2023) 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*, **10**.

Panigrahi,M. *et al.* (2023) Landmarks in the history of selective sweeps. *Anim Genet*.

Rousset,F. (2008) genepop’007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources*, **8**, 103–106.

Sabeti,P.C. *et al.* (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature*, **449**, 913–8.

Schlamp,F. *et al.* (2016) Evaluating the performance of selection scans to detect selective sweeps in domestic dogs. *Molecular Ecology*, **25**, 342–356.

Soni,V. *et al.* (2023) Evaluating power to detect recurrent selective sweeps

under increasingly realistic evolutionary null models. *Evolution*, **77**, 2113–2127.
Soni,V. and Jensen,J.D. (2024) Temporal challenges in detecting balancing selection from population genomic data. *G3 Genes|Genomes|Genetics*, jkae069.
Stephan,W. (2019) Selective Sweeps. *Genetics*, **211**, 5–13.
Szpiech,Z.A. *et al.* (2021) Application of a novel haplotype-based scan for local adaptation to study high-altitude adaptation in rhesus macaques. *Evolution Letters*, **5**, 408–421.
Szpiech,Z.A. and Hernandez,R.D. (2014) selscan: An Efficient Multithreaded Program to Perform EHH-Based Scans for Positive Selection. *Molecular Biology and Evolution*, **31**, 2824–2827.