iHDSel software: The Price equation and the population stability index to detect genomic patterns compatible with selective sweeps. An example with SARS-CoV-2.

Antonio Carvajal-Rodríguez[1,*]

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

*To whom correspondence should be addressed.

keywords: Price equation, information theory, population stability index, Jeffreys divergence, haplotype allelic class, selective sweep.

Abstract

A large number of methods have been developed and continue to be developed for detecting the signatures of selective sweeps in genomes. Significant advances have been made, including the combination of different statistical strategies and the incorporation of artificial intelligence (machine learning) methods. Despite these advances, several common problems persist, such as the unknown null distribution of the statistics used, necessitating simulations and resampling to assign significance to the statistics. Additionally, it is not always clear how deviations from the specific assumptions of each method might affect the results.

In this work, allelic classes of haplotypes are used along with the informational interpretation of the Price equation to design a statistic with a known distribution that can detect genomic patterns caused by selective sweeps. The statistic consists of Jeffreys divergence, also known as the population stability index, applied to the distribution of allelic classes of haplotypes in two samples. Results with simulated data show optimal performance of the statistic in detecting divergent selection. Analysis of real SARS-CoV-2 genome data also shows that some of the sites playing key roles in the virus's fitness and immune escape capability are detected by the method.

The new statistic, called J_{HAC}, is incorporated into the iHDSel software available at https://acraaj.webs.uvigo.es/iHDSel.html.

Introduction

Evolutionary biology studies the factors that affect genetic variability in populations and species. The main processes that influence the evolution of this variability include mutation and recombination, genetic drift, migration, and natural selection. Natural selection, in addition to affecting the allele carrying a beneficial mutation, impacts the neutral alleles of loci linked to the selective one, producing what is known as genetic hitchhiking (Smith and Haigh 1974; Kaplan et al. 1989), which leads to a selective sweep (Berry et al. 1991; Stephan 2019), meaning a loss of diversity around the selected site. These sweeps can be complete or incomplete, strong or soft, and they can even overlap (Johri, Stephan, et al. 2022). Regarding the detection of the footprint left by selective sweeps in genomes, from the earliest methods that explored haplotype patterns, whether by studying homozygosity (Sabeti et al. 2007), its diversity (Kimura et al. 2007), or interpopulation differentiation (Chen et al. 2010), among others, a great number of methods have been developed and continue to be developed. Significant advancements have been made, including the use of summary statistics, the combination of different statistical strategies, and the incorporation of artificial intelligence-based methods (Horscroft et al. 2019; Stephan 2019; Abondio et al. 2022; Arnab et al. 2023; Panigrahi et al. 2023; Whitehouse and Schrider 2023).

Methods for detecting selective sweeps require the existence of haplotypic data. Despite improvements in the efficiency and accuracy of methods for estimating haplotypes (Delaneau et al. 2019; Meier et al. 2021; Shipilina et al. 2023), in non-model species (understood as those in which, whether or not a genome has been sequenced, it is poorly annotated and has not traditionally been a model species in the pre-genomic era), haplotype-based detection methods are still not widely used. Instead, it is more common to use interpopulation methods based on detecting molecular markers with excessively high differentiation values, known as "outliers". But even in the case of model species, the use of haplotype-based methods to detect selective sweeps presents the problem that the same genomic pattern that could be produced by a selective sweep could also be explained under different scenarios related to factors as diverse as the quality and characteristics of the sampled data, biological characteristics related to mutation and recombination rates, as well as demographic history and the effects of purifying and background selection (Johri, Aquadro, et al. 2022; Soni et al. 2023; Soni and Jensen 2024).

Part of this problem arises from the lack of knowledge of the null distribution of the statistics used, which requires simulating the neutral biological scenario. But overall, it is clear that although a statistical tool can detect a specific genomic pattern in the data, it is unlikely that that pattern could be due solely to the effect of a selective scan. It may do so in some scenarios, but not in others.Therefore, to validate a candidate SNP or region as a result of a selective process, it is first necessary to prove that the statistic does not generate false positives in realistic scenarios in terms of demography and other evolutionary parameters of interest. Subsequently, functional validation of these candidate loci will always be necessary (Johri, Aquadro, et al. 2022). This does not preclude that the development of statistical tools to detect genomic patterns that may be related to selective sweeps remains of great interest. It would also be interesting if that statistic had a known null distribution.

When studying a selective sweep, we can trace its effect over time (directional selection) or across space (divergent selection). Therefore, if we use two samples to compare the effect of the sweep, they can be separated by time or space. Detecting the footprint of natural selection in genomes in general, and specifically divergent selection, is important for studying speciation processes (Galindo et al. 2021) and climate adaptation (Folkertsma et al. 2024), but also for more immediate effects such as resistance to infections in commercially important marine species (Pampín et al. 2023; Vera et al. 2023).

In this work, I propose a statistic that uses the population stability index, also known as Jeffreys divergence, to compare the distribution of allelic classes of haplotypes (Labuda et al. 2007; Hussin et al. 2010) between two populations or samples. To develop the statistic I use the informational interpretation of the Price equation (Price 1972; Frank 2012a) defined for the haplotype allelic class trait. The advantage of this statistic is that it follows a chi-square distribution when the null hypothesis (equal distribution of haplotype classes among samples) is true. This not only increases computational efficiency by several orders of magnitude but also allows for the testing of biological models expected to deviate from this hypothesis, including the presence of local selection and its corresponding selective sweep. Below, I will present the development of the statistic and then demonstrate its behavior with both simulated and real genomic data from various samples of the SARS-CoV-2 virus.

The Price equation and the population stability index to compare population genomes

Price equation

The Price equation in its most general formulation describes the change between two populations at any scale, spatial or temporal (Frank 2012a; Frank 2017). The equation partitions the change into a part due to natural selection and another part due to other effects. We compare two populations or frequency distributions which can be separated by space and/or time. Natural selection causes populations to accumulate information, which is measured in relation to the logarithm of biological fitness *m*= log(ω), where ω is the relative fitness (Frank 2012b; Frank 2012a).

Therefore, let *z* be a character that takes different values z_i with associated frequency p_i in population *P* and with frequency q_i in population Q. If we consider the logarithm of fitness as the character, *z*=*m*, we have that the mean change in *m* due to the effect of natural selection in one or the other population is (Frank 2012a)

$$\Delta_s \bar{m} = J(p,q) = \beta_{mw} D_w \text{ where } D_w = \frac{V_w}{\bar{w}} \qquad (1)$$

where *J* is the Jeffreys divergence or population stability index, *p* and *q* the frequency of the different values of *m* in the populations *P* and Q respectively, and β_{mw} is the regression of *m* on the absolute fitness *w*.

However, it is possible to use scales other than the fitness logarithm to measure information, with the key element being the regression of values in the new scale on fitness

(Frank 2013). Therefore, to detect the effect of natural selection from genomic data, it will be necessary to measure those genomic patterns with high regression values on biological fitness. In this work, we propose the haplotype allelic class (HAC) as a suitable pattern to capture the increase in information generated by natural selection, whether in temporal comparisons (directional selection) or spatial comparisons (divergent selection).

Haplotype allelic class (HAC)

Haplotype allelic classes were initially introduced in (Labuda et al. 2007) and later used to detect genomic patterns caused by selective sweeps (Hussin et al. 2010) and divergent selection (Carvajal-Rodríguez 2017).

Consider a sample of sequences and compute the reference haplotype *R* as the one formed by the major allele of each site. Now, consider for the same or another sample of sequences, the haplotypes of length *L*+1 centered in a given candidate SNP *c* and define the mutational distance between any haplotype and the reference *R* as the Hamming distance between the haplotype and the reference i.e. the number *h* of sites in the haplotype carrying an allele different to the one in *R*. Each group of haplotypes having the same *h* will constitute an haplotype allelic class (HAC, Labuda et al. 2007; Hussin et al. 2010). The HAC distribution is estimated from the distribution of the *h* values in a sample.

Thus, in a given haplotype with the candidate SNP position *c* in the middle, for each position other than *c* we count the outcome $X_k = I(s_k \neq r_k)$ were s_k is the allele in the position *k* of the haplotype, r_k is the allele in the reference and *I*(A) is the indicator variable taking 1 if A is true and 0 otherwise. Therefore, the *h* value of an haplotype of length *L*+1 is

$$h=\sum_{k=1}^{L+1} X_k \quad \text{where}\, k\neq c\, , X_k=I\left(s_k\neq r_k\right)\text{and}\, h\in[0,L] \qquad (2)$$

The idea behind using *h*-values to detect selective sweeps is that if one allele increases in frequency due to the effect of selection, the higher frequency alleles from adjacent sites will be swept along with the selected allele so that these haplotypes will have many common alleles with the reference configuration, i.e., an *h*-value close to zero.

Information for haplotype allelic classes: the population stability index

Let h_i be the HAC value that satisfies *h*=*i* with $i\in[0,L]$ then for a sample of n_1 sequences in *P*, the frequency of h_i is

$$P_i=\#h_i/n_1 \quad \text{with} \quad \sum_i P_i=1$$

similarly, for a sample of n_2 sequences in *Q*, the frequency of h_i is

$$Q_i=\#h_i/n_2 \quad \text{with} \quad \sum_i Q_i=1$$

In previous works, studying the distribution of alleles around a candidate site in both samples *P* and *Q* , has been performed comparing in several ways the HAC variances of the partitions that have the reference allele or not in the different samples (Carvajal-Rodríguez 2017; Gabián et al. 2022). There are some problems with this type of approach as the unknown distribution of the defined statistics or a loss of power when using homogeneity variance tests. Here, I rely on the abstract model of the Price equation as proposed by Frank (Frank 2012a; Frank 2013; Frank 2017; Frank 2020) to calculate, using Jeffreys divergence, the change caused by selection in the distribution of HAC values between two populations.

Number of classes and smoothing

For a total of *L*+1 different classes the Jeffreys divergence is (Kullback 1997)

$$J_{HAC}=\frac{n_1 n_2}{n_1+n_2}\sum_{i=0}^{L}\left(P_i-Q_i\right)\ln\frac{P_i}{Q_i}$$

However, computing J_{Hac} in this way could suffer from the curse of dimensionality if eventually $L > n_1+n_2$ which will cause the presence of the different classes to be scarce. To alleviate this problem we will group the values in *K* (*K*≤*L*) HAC classes. The number of classes *K* is an important parameter because too many classes have the dimensionality issue but too few classes will have low power for the distribution comparison. A heuristic conservative guess is *K*=*L*/2 when *L*>=15 or *K*=*L* otherwise.

Given *K*, we will group uniformly the *h* values into *K* groups so that the first group indicates classes with less than (100/*K*)% of minor alleles, the next corresponds to classes with more than (100/*K*)% but less than 2×(100/*K*)%, until the last group with more than (*K*-1)×(100/*K*)% but equal or less than 100%. The class with 100% of minor alleles is included in this last group.

Thus, for population *P*, the frequency P'_i of each group of classes is

$$\begin{cases} S_i=\sum_{j=u}^{U}\#h_{j-1}/n_1 \ \text{where}\, i\in[0,K],\, u=1+\frac{L}{K}i \ \text{and}\ U=\frac{L}{K}(i+1) \\ i\in[0,K-1]:\ P'_i=S_i \\ i=K: \qquad P'_K=S_K+\#h_L/n_1 \end{cases} \tag{3}$$

However, note that the Jeffreys divergence is defined only if *P* and *Q* have no zeros. To avoid zeros we use additive smoothing (Manning et al. 2008) with a pseudocount α=0.5 for each possible outcome so that S_i and P'_k in (3) become

$$S_i = \sum_{j=u}^{U} (\# h_{j-1} + \alpha) / (n_1 + \alpha K)$$

$$P'_k = S_K + (\# h_L + \alpha) / (n_1 + \alpha K)$$

So, for *K* groups of HAC classes, the Jeffreys divergence for comparing the HAC distribution between populations *P* and *Q* finally is (c.f. eq. 5.10 in Kullback 1997 p. 130)

$$J_{HAC} = \frac{n_1 n_2}{n_1 + n_2} \sum_{i=0}^{K} (P'_i - Q'_i) \ln \frac{P'_i}{Q'_i} \quad \text{with values in } [0, +\infty) \quad (4)$$

Note that J_{HAC} is also known as the population stability index and is asymptotically distributed as Chi-square with *K*-1 degrees of freedom.

The advantage of using (4) in the context of studying the genomic footprint of selection is that, contrary to other statistics, it can be approached by a chi-square distribution providing a faster approach as we can avoid performing computationally expensive simulations or resampling.

Phenotypic scale, linkage disequilibrium and window size

Phenotypic scale

The gain in information caused by the effect of natural selection as expressed in (1) depends on the log-fitness *m* and if we measure the frequency of the h_i classes instead of fitness classes, the relationship between the average change in the *h* distribution and the gain in information will depend on the regression of *h*-values on fitness as follows (Frank 2013)

$$\Delta_s \bar{h} = \beta_{hw} D_w \text{ where } D_w = \frac{V_w}{\bar{w}}$$

thus, if we use the HAC values to compute *J* we obtain J_{HAC}

$$J_{HAC} = \beta_{hw} D_w = \frac{\beta_{hw}}{\beta_{mw}} J$$

The quantity $\beta_{hw}/\ \beta_{mw}$ is the change in phenotype (HAC values) relative to the change in information (Frank 2013). Therefore, if there is perfect fit between ln(*P*/*Q*) and *m* then J_{HAC}=*J*.

The regression of *h* on *w* will be high when it is fitness that is distributing the classes of *h*, which requires that there are indeed one or more sites under selection within the haplotype window. However, this is a necessary but not sufficient condition. Price's equation for total change indicates that the average variation in phenotype *h* has two components: one due to selection and the other due to other causes, including changes in the components of the phenotype that are transmitted (Δ*h*)

$$\Delta \bar{h} = \Delta_s \bar{h} + q' \Delta h$$

In our context, the change in *h* not caused by selection may be due to, besides mutation, the effect of recombination on haplotypes, which in turn will depend on the window size. Therefore, we are interested in using window sizes that correspond to haplotype blocks in order to minimize Δ*h*.

Window size

The program computes haplotype blocks and set the candidate position *c* in the middle of each block. An haplotype block is computed as a sequence of reference SNPs with lenght *W* that satisfies $r^2(c\text{-}W/2,c\text{-}W/2+1),\ldots, r^2(c\text{-}1,c), r^2(c,c+1), r^2(c+x\text{-}1,c+x),\ldots, r^2(c+W/2\text{-}1, c+W/2),\ldots$ where *r* is the correlation coefficient calculated from the sample of size *n* so that $Pr(nr^2)\leq\alpha$ and nr^2 has a Chi square distribution. Furthermore, for a given SNP *c*+1 to be included in the block, it is also required that $D'(c, c+1)\geq 0.4$, where *D*' is the normalized linkage disequilibirum (Lewontin 1964). The block is extended until any of both conditions is rejected i.e. $Pr(nr^2_{c+x-1,c+x})>\alpha$ or $D'(c+x\text{-}1, c+x)<0.4$.

Optionally, the program can use an outlier as the putative center of a block and build the block around it. In this case, the condition for defining a block is more liberal, allowing blocks that have a mean normalized linkage disequilibrium value greater than zero. The reason is that the outliers may have been part of older blocks, so we use the minimum condition that the average linkage of reference alleles is greater than zero assuming that, if they are not the product of selective sweep, the distribution of haplotypic classes will not be affected, the latter will be checked in the next section by simulation.

Simulations

The same simulated data as in (Carvajal-Rodríguez 2017; Gabián et al. 2022) were used. Two populations of 1000 facultative hermaphrodites were simulated under divergent selection and different conditions about mutation, recombination, migration and selection. Each individual consisted of a diploid chromosome of length 1Mb.

Input setting for the simulations

A minor allele frequency (MAF) value of 0.01 was used. As we have already seen, the program allows defining the window or haplotypic block size automatically, using the correlation between pairs of sites to define the block size and placing the central SNP as a candidate or, alternatively, it uses the detected outliers as candidate SNPs and then calculates the window size. Both methods were used. All other parameters were as defined by default (see the program manual). An example of the command line to launch case C1 (Table 1) and analyze the 1,000 files located in subfolder C1 and using the automatic calculation of blocks (-useblocks 1) is:

```
-path /home/data/C1/ -runs 1000 -input Om_SNPFile_Run -format ms -sample 50
-minwin 11 -output JHAC_C1_ -maf 0.01 -useblocks 1 -doEOS 1 &
```

The *-doEOS* tag indicates whether we want (1, default) or not (0) to run in addition the EOS outlier test (Carvajal-Rodríguez 2017). If the calculation without blocks is used (-useblocks 0) the doEOS tag must necessarily be set to 1.

Simulation results

In the following tables the results of power (Tables 1-3) and false positive rate (Table 4) after analyzing 1000 replicates of each scenario are presented. In summary, for haplotypes with linkage and the selective site in the center of the chromosome, when using the automatic blocks system, the power is equal to or greater than 95%, regardless of mutation and recombination rates. As expected, if the sites are not linked, the method does not work because there is no selective sweep (Table 1). When the position of the selective site moves away from the center of the chromosome (Table 2), the power remains high. Localization improves as recombination increases and as the marker is located closer to the center. In the case of multiple selective sites (Table 3), the power to detect at least three is above 75% when using automatic blocks but only detects one (97% power) in the case of blocks centered on outliers. In general, for blocks centered on outliers, the power is slightly lower, but in some cases, the localization was considerably more accurate.

Table 1. Percent power for detecting divergent selection by J_{Hac} in simulated data with the selective site in the middle. The power was computed as 100×the number of replicates where selection was detected/1000. In parentheses, the corresponding value when the blocks were built around outliers instead of finding the blocks automatically, if the value is equal the = symbol appears. Genome size is 1Mb. Population size *N*= 1000. *T*: number of generations. Population mutation rate θ = 4*Nμ*. Population recombination rate ρ = 4*Nr*. *s*: selection coefficient. *Dist*: average distance in Kb from the detected position to the actual effect, given only when ρ>0. *W*: average size, in number of SNPs, of the haplotypes analyzed. Significance level α = 0.05. Each case was replicated 1,000 times.

Case	*T*	*θ*	*ρ*	*s*	% power	Dist Kb	*W*
C1	10^4	12	0	± 0.15	100 (98)	-	14 (13)
C2	10^4	12	4	± 0.15	100 (98)	42 (38)	14 (13)
C3	10^4	12	12	± 0.15	100 (96)	4 (10)	13 (12)
C7	5×10^3	60	0	± 0.15	100 (94)	-	13 (12)
C8	5×10^3	60	4	± 0.15	100 (85)	37 (14)	13 (12)
C9	5×10^3	60	60	± 0.15	98 (80)	14 (2)	12 (11)
C13	10^4	60	0	± 0.15	100 (100)	-	14 (=)
C14	10^4	60	4	± 0.15	99 (100)	126 (15)	14 (13)
C15	10^4	60	60	± 0.15	95 (91)	19 (2)	13 (12)
C15Indep	10^4	60	∞	± 0.15	0 (2*)	- (-)	- (11)

* Note that this 2% results from using the outlier-centered haplotype method. When directly inspecting outliers with the EOS method, the power was 78%.

Table 2. Percent power for detecting divergent selection by J_{Hac} in simulated data with the selective site in different locations. The power was computed as 100×the number of replicates where selection was detected/1000. In parentheses, the corresponding value when the blocks were built around outliers instead of finding the blocks automatically, if the value is equal the = symbol appears. Genome size is 1Mb. Population size *N*= 1000. *T*: number of generations. Population mutation rate θ = 4*Nμ*. Population recombination rate ρ = 4*Nr*. *s*: selection coefficient. *Loc*: true relative position of the selective site. *Dist*: average distance in Kb from the detected position to the actual effect, given only when ρ>0. *W*: average size, in number of SNPs, of the haplotypes analyzed. Significance level α = 0.05. Each case was replicated 1,000 times.

Case	*T*	*θ*	*ρ*	*s*	Loc	% power	Dist Kb	*W*
C13loc0	10^4	60	0	± 0.15	0.0	100 (99)	-	14 (13)
C13loc10	10^4	60	0	± 0.15	0.01	100 (99)	-	14 (13)
C13loc100	10^4	60	0	± 0.15	0.1	100 (98)	-	14 (13)

C13loc250	10^4	60	0	± 0.15	0.25	100 (99)	-	14 (13)
C14loc0	10^4	60	4	± 0.15	0.0	98 (93)	300 (262)	14 (13)
C14loc10	10^4	60	4	± 0.15	0.01	98 (96)	285 (292)	14 (13)
C14loc100	10^4	60	4	± 0.15	0.1	99 (96)	180 (229)	14 (13)
C14loc250	10^4	60	4	± 0.15	0.25	99 (98)	62 (114*)	14 (14)
C15loc0	10^4	60	60	± 0.15	0.0	86 (79)	211 (189)	13 (11)
C15loc10	10^4	60	60	± 0.15	0.01	87 (80)	198 (170)	13 (11)
C15loc100	10^4	60	60	± 0.15	0.1	91 (89)	106 (70)	13 (12)
C15loc250	10^4	60	60	± 0.15	0.25	92 (89)	37 (14)	13 (12)

* Several runs with average F_{ST} > 0.5 and no outliers, so the 90th percentile was considered.

Table 3. Percent power for detecting divergent selection by J_{Hac} in simulated data for a polygenic model with 5 selective sites uniformly distributed in the chromosome. The power was computed as the number of replicates where selection was detected. In parentheses the corresponding % power when the blocks were built around outliers instead of finding the blocks automatically, if the value is equal the = symbol appears. Genome size is 1Mb. Population size N= 1000. Number of generations T=10^4. Population mutation rate $\theta = 4N\mu$=60. Population recombination rate $\rho = 4Nr$=60. Selection coefficient per site s=± 0.032 . W: average size, in number of SNPs, of the haplotypes analyzed. Each case was replicated 100 times.

Case	Candidate	% power	W
C15poly	1	99 (97)	15 (18)
C15poly	2	89 (0)	16
C15poly	3	75 (0)	16
C15poly	4	59 (0)	16
C15poly	5	44 (0)	16

Finally, in the neutral simulations where there was no selective site (Table 4), the false positive rate conservatively remains below the expected 5%, both using automatic blocks and those centered on outliers, with one exception corresponding to the effect of bottlenecks. When a bottleneck occurs, it can generate linkage disequilibrium that could resemble the effect of a selective sweep, thus increasing the possibility of false positives. In our case, we observed that J_{Hac} becomes liberal with 13% when the blocks are centered around the outliers, which means an 8% excess over the expectation. The explanation for this happening with blocks centered on outliers but not with automatic ones is that, as previously indicated, the construction of blocks centered on outliers is somewhat more liberal, validating as blocks those regions that have an average disequilibrium greater than 0. A conservative option available for the above exception is to set the window size to a higher value, say 25 or 50, which solves the problem and sets the false positive rate to just 2%. While for the corresponding selective case when we run the program with these window sizes the power is 90%.

Table 4. Percent false positive rate for detecting divergent selection in simulated neutral data. In parentheses the corresponding value when the blocks were built around outliers instead of finding the blocks automatically, if the value is equal the = symbol appears. Genome size is 1Mb. Population size *N*= 1000. *T*: number of generations. Population mutation rate θ = 4*Nμ*. Population recombination rate ρ = 4*Nr*. %FPR = 100×number of replicates with significant J_{Hac} test/1000. *W*: average size, in number of SNPs, of the haplotypes analyzed. Each case was replicated 1,000 times.

Case	*T*	*θ*	*ρ*	% FPR	*W*
C4	10^4	12	0	0.1 (1)	11 (=)
C5	10^4	12	4	0.3 (2)	12 (11)
C6	10^4	12	12	0.1 (4)	12 (11)
C10	5×10^3	60	0	0 (0.4)	- (11)
C11	5×10^3	60	4	0 (2)	- (11)
C12	5×10^3	60	60	0.2 (3)	12 (11)
C16	10^4	60	0	0.3 (0.4)	12 (11)
C17	10^4	60	4	1 (2)	12 (11)
C18	10^4	60	60	1 (4)	13 (=)
C18Indep	10^4	60	∞	0 (0.2)	- (11)
C18Bottle	10^4	60	60	3 (13)	12 (11)
C18Bottle	10^4	60	60	2	26*
C18Bottle	10^4	60	60	2	51*

* window size set to a specific value

Real data analysis: SARS-CoV-2

SARS-CoV-2 virus genomes stored in the GISAID database (Khare et al. 2021) are indexed by both locality and the time period where they were sampled thus presenting a unique opportunity to apply iHDSel to both time or spatially separated samples. Therefore, as an example of application, we are going to compare SARS-CoV-2 genomes sampled in Spain (SP), England (EN) and South Africa (SA) in periods corresponding to different waves. The findings of this section are based on data associated with 30,274 SARS-CoV-2 genomes available on GISAID up to February 12, 2024, gisaid.org/EN1, gisaid.org/EN2, gisaid.org/EN3, gisaid.org/EN4, gisaid.org/SP1, gisaid.org/SP2, gisaid.org/SA.

The downloaded genomes were complete (>29,000 bp) and of high quality (<1% undefined bases and <0.05% unique amino acid mutations). These datasets were then processed using the Nextclade CLI for quality control (Aksamentov et al. 2021). Briefly, the Nextclade CLI examines the completeness, divergence, and ambiguity of bases in each genome. Only genomes considered ‘good’ by Nextclade CLI were selected.

The samples from England (EN1, EN2,EN3 and EN4) correspond to the period of March 2020, at the beginning of the first wave of the pandemic (EN1, 4820 genomes collapsed to 4227 after quality control), a second sample taken between March 28 and March 31, 2021, inclusive (EN2, 5966 genomes collapsed to 4152 after quality control), a third from June 24 to June 26, 2021, inclusive (EN3, 6886 genomes collapsed to 5844 after quality control), and from October 1, 2023, until January 31, 2024, inclusive (EN4, 3928 genomes collapsed to 3712 after quality control).

The samples from Spain (SP1 and SP2) correspond to the periods June 24, 2021, to July 12, 2021, inclusive (SP1, 6195 genomes collapsed to 4627 after quality control) and October 1, 2023, to January 31, 2024, inclusive (SP2, 1012 genomes collapsed to 221 after quality control).

Finally, the sample from South Africa corresponds to the same period as SP1, June 24, 2021 to July 12, 2021, inclusive (SA, 1467 genomes collapsed to 1327 after quality control).

These samples will allow us to compare population changes in space or time. We will compare genomes from different samples to study if there are genomic patterns that the J_{HAC} test identifies as potentially caused by selection (see below).

Rationale of the comparisons

Spatial comparisons: SP1-SA, EN3-SA, EN3-SP1

These comparisons involve samples from different countries obtained in the same time period of the pandemic. The interest in the comparison with South Africa is that on June 24, 2021 to July 12, 2021, vaccination rates were high in Spain and England but very low in South Africa. Virtually 100% of the Spanish and English population was vaccinated with at least one dose and less than 10% of the South African population (https://ourworldindata.org/covid-vaccinations?country=ZAF).

Temporal comparisons: EN1-EN2, EN2-EN3, EN3-EN4

These comparisons affect the same country but in different periods of the pandemic from the beginning of the first wave to the beginning of 2024 with virtually the entire population already vaccinated several times and the majority variant being Omicron and its subvariants (Brüssow 2022; Wang et al. 2023; Wang et al. 2024).

Spatial comparisons: EN4-SP2

At the end of 2023, the JN.1 subvariant of Omicron, originating from the BA.2.86 lineage, began to spread. This subvariant already carried more than 30 mutations in the spike protein compared to previous subvariants. JN.1 includes the L455S mutation and, by the end of 2023, exhibited a higher reproductive rate than previous sublineages in countries such as Spain, France, and England, with the number of detected JN.1 sequences being higher in England than in Spain (Kaku et al. 2024). During this period, DV.7.1, a sub-lineage of BA.2.75, was highly prevalent in Spain (50% compared to 5% in the UK, https://cov-lineages.org/lineage_list.html) and was considered a variant to monitor, although it was later downgraded. Therefore, the comparison between EN4 and SP2, corresponding to October 2023 - January 2024, is of interest to study the potential patterns of divergent selection in the evolutionary dynamics of Omicron subvariants between these two countries.

Genome alignment and lineage classification

The pooled genomes for each comparison were aligned with the MAFFT FFT-NS-2 program (Katoh and Standley, 2013) with the specific version for SARS-CoV-2 accessible online (https://mafft.cbrc.jp/alignment/server/add_sarscov2.html). Sequences that had more than

5% ambiguous sites were removed and also, to keep the alignment length the same as the input, insertions were deleted. The remaining options were the default. After the alignment, and following the protocol recommended by NextStrain given the possibility of artifactual SNPs located at the beginning and end of the alignment (van Dorp et al. 2020), sites in the first 130 base pairs and the last 50 were removed using the program Mega X (Kumar et al. 2018). Lineages were identified with Nextclade CLI (Table 5).

Table 5. Percentage of SARS-CoV-2 lineages in the analyzed data.

Data	%Alpha	%Beta	%Delta (%AY.4/AY.45)	%Gamma	%Omicron (%JN.1/FLIP/DV.7.1)	%Other (pre-Alpha, Lambda, Mu, recombinants, undefined)
SP1	24	2	70 (2/0)	2	0	2
SA	1	3	94 (0/57)	0	0	2
EN1	0	0	0	0	0	100 (pre-Alpha)
EN2	98	1	0.1	0.1	0	0.8
EN3	1	0.02	98.9 (72/0)	0	0	0.08
EN4	0	0	0	0	96 (39/6/1)	4 (recombinants)
SP2	0	0	0	0	97 (26/12/28)	3 (recombinants)

Input settings for iHDSel

A minor allele frequency (MAF) value of 0.01 was used. The two methods already mentioned were used to define the window size (automatic or outlier-centered blocks) and the results detected by either of the two methods are reported. All other parameters were the ones by default (see program manual). An example of the command line for the comparison between EN3 and SP1 where both samples are in the file EN3_SP1.fas located in the data folder and using the outlier-centered block calculation (-useblocks 0) is:

```
-path /home/data/ -input EN3_SP1.fas -format fasta -output EN3_SP1 -
useblocks 0 -tag ENGLAND &
```

where -*tag* is the argument that defines the word included in the name from the England sequences and that allows to separate both samples.

Similarly, for the temporal comparison between EN2 and EN3

```
-path /home/data/ -input EN2_EN3.fas -runs 1 -format fasta -output EN2_EN3
-useblocks 0 -tag 2021-03 -reference 2
```

where we have added the -*reference* tag to indicate that the EN3 sample should be used as a sample to calculate the blocks and the reference haplotype.

The imprint of selection in the SARS-CoV-2 genomes

Spatial comparisons: SP1-SA (summer 2021)

The SP1 sample has a majority Delta (70%) and Alpha (24%) composition while SA is mostly (94%) Delta (Table 5). The pooled SP1-SA sample consists of 247 SNPs with a frequency greater than 1%. After genome-wide analysis, iHDSel did not find any significant haplotypic blocks in the automatic search nor when focusing on outliers.

Spatial comparisons: EN3-SA (summer 2021)

Both samples are mostly Delta (99% EN3 and 94% SA, Table 5). The pooled EN3-SA sample consists of 107 SNPs with a frequency higher than 1%. After whole genome analysis, iHDSel found one site with the automatic block method (28,282) and five sites centered on outliers (Table 6).

The first site is 7,851, which corresponds to ORF1a 2,529. In the SA sample, 100% of the sequences have the amino acid A, while in EN3, there is 27%A and 73%V, indicating the change A2529V. It is noteworthy that A2529V is one of the main SARS-CoV-2 mutations associated with virus fitness (Jankowiak et al. 2022). Moreover, in a recent study (Garcia et al. 2024) analyzing the evolution of different lineages in relation to the progress of vaccination, the A2529V mutation in ORF1a showed a significant positive correlation between the prevalence of the mutation and vaccination in Norway during the first 9 months of 2021 (including the sampling period of EN3 and SA).

The second site is 13,812, which, after identifying the slippery region (Kelly et al. 2021) and the start of ORF1b at 13,468, corresponds to amino acid 115 in ORF1b (NSP12). This site has 100%M in EN3 but 42%M and 58%I in SA. The change M115I is a characteristic mutation of the AY.45 lineage (Gangavarapu et al. 2023), which is present in SA with a frequency of 57% but is absent in EN3.

The third and fourth sites are mutations corresponding to amino acid changes in the Spike protein. Specifically, T95I represents the change observed between SA and EN3, with I at a frequency of only 8% in SA but 72% in EN3. The other mutation in Spike is G142D, with D present at 62% in SA and 97% in EN3 (Table 6). Both mutations are characteristic of the Delta variants and increase in frequency in Delta Plus (Cai and Cai 2021; Dhawan et al. 2022; Kannan et al. 2022; Mahmood et al. 2022).

The fifth site is position 25,413 of the genome, corresponding to amino acid 7 in ORF3a, with amino acid I in both samples being EN3 (ATC) and SA (ATT|50%C). Therefore, the existence of a significant signal due to different HAC distribution must be caused by accumulated variation in the surrounding sites. Similarly, the sixth and final site corresponds to amino acid 3 of the N protein, with the amino acid being D (GAT) in 99% of the cases in both samples, with practically 1% being L (CTA). Again, the existence of a significant signal due to different HAC distribution is caused by accumulated variation in the surrounding sites.

Table 6. Significant J_{Hac} tests (*p-val*<0.05) for EN3-SA comparison (with 107 SNPs and sample sizes n_{EN3} = 5844, n_{SA}=1327).

EN3-SA		Gene (protein)	AA	%
Block size	Site (+1+130)		(AA in EN3) (AA in SA)	(p1 | p2 |... EN3) : (p1 |p2 |... SA)
41	7851	ORF1a (NSP3)	(V|A) 2529 (A)	(73 | 27):(- | 100)
11	13812	ORF1b (NSP12)	(M) 115 (M| I)	(100):(42 | 58)
30	21846	ORF2 (S)	(I|T) 95 (I|T)	(76 | 24):(8 | 92)
14	21987	ORF2 (S)	(D|G) 142 (D|G)	(97 | 3):(62 | 38)

11	25413	ORF3a	(I) 7 (I)	(100):(100)
14	28282	ORF9 (N)	(D|L) 3 (D|L)	(99 | 1):(99 | 1)

(+1+130): added to the program output position, the +1 to correct the program indexing to 0 and the +130 to correct the eliminated initial positions.

Spatial comparisons: EN3-SP1 (summer 2021)

We already saw that the EN3 genomes are predominantly Delta (99%), while SP1 has 70% Delta genomes and 24% Alpha (Table 5). The combined EN3-SP1 sample consists of 154 SNPs with a frequency greater than 1%. After the whole genome analysis, iHDSel found one significant site. The nucleotide site 7851 corresponds to amino acid 2,529 in ORF1a, which was also significant in the EN3-SA comparison, and we saw that A2529V is one of the main SARS-CoV-2 mutations associated with virus fitness. In this comparison, the change is from 98%A in SP1 to 73%V (27%A) in EN3.

Therefore, regarding the spatial comparisons in the summer of 2021, we see that in the SA and SP1 samples, amino acid 2529 of ORF1a was still A in virtually 100% of the sequences analyzed, while in EN3, only 27% had A and the remaining 73% were already V. This mutation is associated with an advantage for the virus and in relation to vaccination, and indeed, the J_{HAC} statistic detects it as a site with a selective pattern.

Temporal comparisons: EN1-EN2 (March 2020 vs March 2021)

The comparison between the English genomes is between samples separated in time (different waves). These comparisons should be considered with caution as the differentiation between samples is very large. Indeed, the mean F_{ST} in all three comparisons (EN1-EN2, EN2-EN3 and EN3-EN4) is above 0.5. However, the sites detected in the three comparisons correspond to sites with recognized impact on virus fitness.

The genomes in EN1 belong to pre-alpha variants, while the genomes in EN2 are Alpha. The combined EN1-EN2 sample consists of 77 SNPs with a frequency greater than 1%. After the whole genome analysis, iHDSel found six significant sites for the J_{HAC} test. These sites correspond to six Spike mutations, namely amino acids 501, 570, 681, 716, 982, and 1118 (Table 7). All of them correspond to the characteristic Spike mutations of Alpha (Gangavarapu et al. 2023). The only one missing is D614G, although it is included in the detected haplotypic regions. The fact that it does not come out as directly significant may be because the program did not use that position as the center of a haplotypic block, as it detected the other sites as more extreme outliers since 614G has a presence of 61%G in EN1 and 99.9% in EN2. However, when the program is run proposing the nucleotide positions corresponding to tha amino acid 614 as candidates, the result is significant. Therefore, it seems that the haplotypic region including all these mutations has been detected.

Table 7. Significant J_{Hac} tests (*p-val*<0.05) for EN1-EN2 comparison (with 77 SNPs and sample sizes n_{EN1} = 4224, n_{EN2}=4152).

EN1-EN2		**Gene (protein)**	**AA**	**%**
Block size	Site (+1+130)		(AA in EN1) position (AA in EN2)	(p1 | p2 |... EN1) : (p1 |p2 |... EN2)
11	23063	ORF2 (S)	N501Y	(100):(1 | 99)
11	23271	ORF2 (S)	A570D	(100):(2 | 98)
11	23604	ORF2 (S)	P681H	(100):(1 | 99)
11	23709	ORF2 (S)	T716I	(100):(1| 99)
11	24506	ORF2 (S)	S982A	(100):(2| 98)
11	24914	ORF2 (S)	D1118H	(100):(1| 99)

(+1+130): added to the program output position, the +1 to correct the program indexing to 0 and the +130 to correct the eliminated initial positions.

Temporal comparisons: EN2-EN3 (March 2021 vs June 2021)

This is a comparison of Alpha (EN2) with Delta (EN3) genomes. The pooled EN2-EN3 sample consists of 105 SNPs with a frequency greater than 1%. After whole genome analysis, iHDSel found seven significant sites using blocks centered on outliers (Table 8).

These included substitution of relevant Spike amino acids at sites such as 452, 478, 681, and 950 (Kannan et al. 2021). For example, the L452R substitution appears to be associated with evasion of the immune response (He et al. 2022). As well as three sites in the N protein, 63, 203 and 377, which correspond to significant mutations of the delta variant, namely, D63G, R203M, and D377Y (Bhattacharya et al. 2023).

Table 8. Significant J_{Hac} tests (*p-val*<0.05) for EN2-EN3 comparison (with 105 SNPs and sample sizes n_{EN2} = 4152, n_{EN2}=5844).

EN2-EN3		Gene (protein)	AA
Block size	Site (+1+130)		(AA in EN2) position (AA in EN3)
18	22917	ORF2 (S)	L452R
18	22995	ORF2 (S)	T478K
13	23604	ORF2 (S)	H681R
12	24410	ORF2 (S)	D950N
12	28461	ORF9 (N)	D63G
11	28881	ORF9 (N)	K203M
13	29402	ORF9 (N)	D377Y

(+1+130): added to the program output position, the +1 to correct the program indexing to 0 and the +130 to correct the eliminated initial positions.

Temporal comparisons: EN3-EN4 (June 2021 vs January 2024)

This is a comparison of Delta genomes (EN3) with Omicron genomes (EN4). The pooled EN3-EN4 sample consists of 239 SNPs with a frequency greater than 1%. After whole genome analysis, iHDSel identified several sites with F_{ST} greater than 0.99 and 14 of them in the center of significant blocks (Table 9).

The first site occurs in ORF1a (NSP5) and corresponds to the amino acid change P132H, which is a mutation in a functionally important domain and characteristic of Omicron (Hossain et al. 2022). The remaining sites presented in Table 9 correspond to core Omicron mutations in Spike (Basheer et al. 2023; Chen et al. 2023) including some like S371F, S373P, and S375F, which are related to alterations in binding and entry preference (Hu et al. 2022; Zheng et al. 2023) and also the 'Kraken' subvariant immune escape F486P (Parums 2023). Finally, the synonymous change L18L in ORF7b is within the same haplotypic block as the reversions A82V and I120T in ORF7a, which, when directly contrasted as candidates, were significant.

Table 9. Significant J_{Hac} tests (*p-val*<0.05) for the EN3-EN4 comparison (with 239 SNPs and sample sizes n_{EN3} = 5844, n_{EN4} = 3712).

EN3-EN4		Gene (protein)	AA
Block size	Site (+1+130)		(AA in EN3) position (AA in SP2)
11	10447	ORF1a (NSP5)	P132H
11	22674	ORF2 (S)	S371F
11	22679	ORF2 (S)	S373P
11	22686	ORF2 (S)	S375F
11	22775	ORF2 (S)	D405N
11	22786	ORF2 (S)	R408S
11	22813	ORF2 (S)	K417N
11	22898	ORF2 (S)	G446S
11	22992	ORF2 (S)	S477N
11	23019	ORF2 (S)	F486P
11	23055	ORF2 (S)	Q498R
11	23075	ORF2 (S)	Y505H
11	25000	ORF2 (S)	D1146D
11	27807	ORF7b	L18L (ORF7a A82V, ORF7a I120T)

(+1+130): added to the program output position, the +1 to correct the program indexing to 0 and the +130 to correct the eliminated initial positions.

Spatial comparisons: EN4-SP2

The genomes of both samples are Omicrom but the subvariant composition is different

(Table 5). The pooled EN4-SP2 sample consists of 218 SNPs with a frequency greater than 1%. After whole genome analysis, iHDSel identified four significant sites (Table 10).

The change A427V in ORF1a is characteristic mutation of the DV.7.1 Omicron sublineage (Gangavarapu et al. 2023) wich is virtually absent in EN4 (0.6%) but has a 28% in SP2 (Table 5) which explains the absence of 427V in EN4 and the 29%V in SP2. The same scenario applies to A520V in ORF1b. The other two significant sites belong to Spike. The mutation at 445 would be related to the V445H and V445P changes that seem to favor immune evasion of the virus (Ao et al. 2023; Chen et al. 2023) with the presence of 445V being 30% in SP2 but only 1% in EN4 (Table 10). Finally, L858I is also a characteristic mutation of DV.7.1.

Table 10. Significant J_{Hac} tests (*p-val*<0.05) for the EN4-SP2 comparison (with 218 SNPs and sample sizes n_{EN4} = 3712, n_{SP2}=221). Only amino acids with a frequency equal or greater than 1% are indicated.

EN4-SP2		**Gene (protein)**	**AA**
Block size	Site (+1+130)		(EN4 AA) position (SP2 AA)
11	1545	ORF1a (NSP2)	A427(71%A|29%V)
13	15026	ORF1b (NSP12)	A520(71%A|29%V)
11	22895	ORF2 (S)	(51%H|47%P|1%V)445(31%H|39%P|30%V)
11	24134	ORF2 (S)	L858(71%L|29%I)

(+1+130): added to the program output position, the +1 to correct the program indexing to 0 and the +130 to correct the eliminated initial positions.

Discussion

In this work, a new statistic called J_{Hac} is proposed to detect genomic patterns compatible with selective sweeps. The statistic is constructed from the interpretation in terms of information of the Price equation (Price 1972; Frank 2012a) and consists of the population stability index applied to the distribution of haplotype classes in two samples. The iHDSel program incorporates the statistic along with the calculation of haplotype blocks in such a way that each candidate site is located in the center of a block. J_{Hac} appears to work optimally with simulated data where two populations are subjected to divergent selection under different mutation and recombination conditions. However, if using the program mode that places the outlier sites in the center of the blocks, care must be taken because the false positive rate increases in bottleneck scenarios. A possible correction in these scenarios is to repeat the calculation with a slightly larger window size.

Real SARS-CoV-2 data have also been used to test J_{Hac} in both spatial and temporal comparisons. Some sites known to impact virus fitness and its ability to promote immune escape have been detected.

The Price equation for comparing genomic patterns

The general formulation of the Price equation describes a change between two populations at any scale, spatial or temporal (Frank 2017). The Price equation has been proposed as a unifying principle in evolutionary biology, allowing the formulation and systematization of different evolutionary models and motivating the development of equations and models that reveal invariances and general principles (Luque 2017; Luque and Baravalle 2021). Here, we have used the selective component of the Price equation, specifically its interpretation in terms of information theory (Frank 2012a), which allows the expression of the covariance between fitness and the trait under study in terms of Jeffreys divergence or population stability index. We have defined the trait as the allelic class of haplotypes and used Jeffreys divergence to compare the distribution of the trait between two populations. The change in trait distribution would be compatible with the effect of selective sweeps, whether due to divergent or directional selection, depending on whether we are comparing populations in space or time.

Limitations of the J_{Hac} method

The detection of selective sweeps is affected by different evolutionary and demographic scenarios. Throughout the space of the various parameters (mutation, recombination, background and deletereous selection, etc.) it is not difficult to find scenarios that generate an excess of false positives (Johri, Aquadro, et al. 2022; Soni et al. 2023). In our case, we have seen that some evolutionary scenarios, such as bottlenecks, can generate interpopulation genomic patterns that increase the false positive rate when using automatic window sizes centered on outliers. Although increasing the window size restores control over the false positive rate, it is possible that other scenarios without positive selection could also alter haplotype class patterns.

Moreover, as we have already indicated, the method proposed here arises from the informational interpretation of the selective component of the Price equation. However, it is a statistical decomposition based on covariance, and we know that correlation does not imply causation. There is also no a priori guarantee that the partition between selection and transmission is additive (Okasha and Otsuka 2020). Therefore, J_{Hac} is an indirect method that detects a genomic pattern possibly related to selection but which can also be generated under other circumstances. Hence, the detected sites should be verified through direct methods such as the study of gene function, fitness, etc.

Finally, some genomic patterns of selection correlate with environmental variables, making it difficult to separate both effects (Folkertsma et al. 2024). The method proposed here could be combined with other methods that take this correlation into account.

Concluding remarks

There are many statistics for identifying regions of selective sweeps in genomes, see for example (Horscroft et al. 2019; Stephan 2019; Horscroft et al. 2020; Abondio et al. 2022; Panigrahi et al. 2023). The use of machine learning-based methods to detect selection patterns has been increasing due to their accuracy and ability to handle large amounts of complex data. The underlying idea of all these methods is to use classification algorithms trained with known response data (simulations). That is, if we aim to detect a selection pattern, we train the algorithm with data that we know contains that pattern and with other data without the pattern. Different types of algorithms have been applied: neural networks, extremely randomized trees, and boosting algorithms (Horscroft et al. 2019; Panigrahi et al. 2023). A major advantage of these methods is their power and flexibility, partly due to the ease of incorporating new statistics with minimal changes to the structure of the method. Two recent machine learning methods have been designed to detect genomic signatures caused by natural selection, using a supervised multi-statistic machine learning approach (Arnab et al. 2023; Lauterbur et al. 2023). In this work, we have developed a new statistic, J_{Hac}, which, due to its known null distribution, allows us to efficiently and quite accurately test for the existence of genomic patterns compatible with selective sweeps. Therefore, J_{Hac} could be an additional measure to consider for future AI-based selection detection methods. In addition, J_{Hac} has been incorporated into the iHDSel program (https://acraaj.webs.uvigo.es/iHDSel.html) along with an automatic haplotype block detection system, so it can be run independently or in conjunction with the heuristic EOS outlier detection method (Carvajal-Rodríguez 2017).

Acknowledgements

This work was supported by Xunta de Galicia (Grupo de Referencia Competitiva, ED431C 2020/05), Ministerio de Ciencia e Innovación (PID2022-137935NB-I00) and 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. Funding for open access charge: Universidade de Vigo/CISUG. I gratefully acknowledge all data contributors, i.e., the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which the real data example in this article is based.

References

Abondio P, Cilli E, Luiselli D. 2022. Inferring Signatures of Positive Selection in Whole-Genome Sequencing Data: An Overview of Haplotype-Based Methods. *Genes* 13:926.

Aksamentov I, Roemer C, Hodcroft EB, Neher RA. 2021. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *Journal of Open Source Software* 6:3773.

Ao D, He X, Hong W, Wei X. 2023. The rapid rise of SARS‐CoV‐2 Omicron subvariants with immune evasion properties: XBB.1.5 and BQ.1.1 subvariants. *MedComm (2020)* 4:e239.

Arnab SP, Amin MR, DeGiorgio M. 2023. Uncovering Footprints of Natural Selection Through Spectral Analysis of Genomic Summary Statistics. *Molecular Biology and Evolution* 40:msad157.

Basheer A, Zahoor I, Yaqub T. 2023. Genomic architecture and evolutionary relationship of BA.2.75: A Centaurus subvariant of Omicron SARS-CoV-2. *PLoS One* 18:e0281159.

Berry AJ, Ajioka J, Kreitman M. 1991. Lack of polymorphism on the Drosophila fourth chromosome resulting from selection. *Genetics* 129:1111–1117.

Bhattacharya M, Chatterjee S, Sharma AR, Lee S-S, Chakraborty C. 2023. Delta variant (B.1.617.2) of SARS-CoV-2: current understanding of infection, transmission, immune escape, and mutational landscape. *Folia Microbiol (Praha)* 68:17–28.

Brüssow H. 2022. COVID-19: Omicron – the latest, the least virulent, but probably not the last variant of concern of SARS-CoV-2. *Microbial Biotechnology* 15:1927–1939.

Cai HY, Cai A. 2021. SARS-CoV2 spike protein gene variants with N501T and G142D mutation–dominated infections in mink in the United States. *J Vet Diagn Invest* 33:939–942.

Carvajal-Rodríguez A. 2017. HacDivSel: Two new methods (haplotype-based and outlier-based) for the detection of divergent selection in pairs of populations. *PLoS One* 12:e0175944.

Chen H, Patterson N, Reich D. 2010. Population differentiation as a test for selective sweeps. *Genome Research* 20:393–402.

Chen S, Huang Z, Guo Y, Guo H, Jian L, Xiao J, Yao X, Yu H, Cheng T, Zhang Y, et al. 2023. Evolving spike mutations in SARS-CoV-2 Omicron variants facilitate evasion from breakthrough infection-acquired antibodies. *Cell Discov* 9:1–5.

Delaneau O, Zagury J-F, Robinson MR, Marchini JL, Dermitzakis ET. 2019. Accurate, scalable and integrative haplotype estimation. *Nat Commun* 10:5436.

Dhawan M, Sharma A, Priyanka, Thakur N, Rajkhowa TK, Choudhary OP. 2022. Delta variant (B.1.617.2) of SARS-CoV-2: Mutations, impact, challenges and possible solutions. *Hum Vaccin Immunother* 18:2068883.

van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, Owen CJ, Pang J, Tan CCS, Boshier FAT, et al. 2020. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infection, Genetics and Evolution* 83:104351.

Folkertsma R, Charbonnel N, Henttonen H, Heroldová M, Huitu O, Kotlík P, Manzo E, Paijmans JLA, Plantard O, Sándor AD, et al. 2024. Genomic signatures of climate adaptation in bank voles. *Ecology and Evolution* 14:e10886.

Frank SA. 2012a. Natural selection. V. How to read the fundamental equations of evolutionary change in terms of information theory. *J Evol Biol* 25:2377–2396.

Frank SA. 2012b. Natural selection. IV. The Price equation. *J Evol Biol* 25:1002–1019.

Frank SA. 2013. Natural selection. VI. Partitioning the information in fitness and characters by path analysis. *Journal of Evolutionary Biology* 26:457–471.

Frank SA. 2017. Universal expressions of population change by the Price equation: Natural selection, information, and maximum entropy production. *Ecol Evol* 7:3381–3396.

Frank SA. 2020. Simple unity among the fundamental equations of science. *Philosophical Transactions of the Royal Society B: Biological Sciences* 375:20190351.

Gabián M, Morán P, Saura M, Carvajal-Rodríguez A. 2022. Detecting Local Adaptation between North and South European Atlantic Salmon Populations. *Biology* 11:933.

Galindo J, Carvalho J, Sotelo G, Duvetorp M, Costa D, Kemppainen P, Panova M, Kaliontzopoulou A, Johannesson K, Faria R. 2021. Genetic and morphological divergence between Littorina fabalis ecotypes in Northern Europe. *Journal of Evolutionary Biology* 34:97–113.

Gangavarapu K, Latif AA, Mullen JL, Alkuzweny M, Hufbauer E, Tsueng G, Haag E, Zeller M, Aceves CM, Zaiets K, et al. 2023. Outbreak.info genomic reports: scalable and dynamic surveillance of SARS-CoV-2 variants and mutations. *Nat Methods* 20:512–522.

Garcia I, Lee Y, Brynildsrud O, Eldholm V, Magnus P, Blomfeldt A, Leegaard TM, Müller F, Dudman S, Caugant DA. 2024. Tracing the adaptive evolution of SARS-CoV-2 during vaccine roll-out in Norway. *Virus Evol* 10:vead081.

He P, Liu B, Gao X, Yan Q, Pei R, Sun J, Chen Q, Hou R, Li Z, Zhang Yanjun, et al. 2022. SARS-CoV-2 Delta and Omicron variants evade population antibody response by mutations in a single spike epitope. *Nat Microbiol* 7:1635–1649.

Horscroft C, Ennis S, Pengelly RJ, Sluckin TJ, Collins A. 2019. Sequencing era methods for identifying signatures of selection in the genome. *Briefings in Bioinformatics* 20:1997–2008.

Horscroft C, Pengelly R, Sluckin TJ, Collins A. 2020. zalpha: an R package for the identification of regions of the genome under selection. *The Journal of Open Source Software* 5.

Hossain A, Akter S, Rashid AA, Khair S, Alam ASMRU. 2022. Unique mutations in SARS-CoV-2 Omicron subvariants’ non-spike proteins: Potential impacts on viral pathogenesis and host immune evasion. *Microb Pathog* 170:105699.

Hu B, Chan JF-W, Liu H, Liu Y, Chai Y, Shi J, Shuai H, Hou Y, Huang X, Yuen TT-T, et al. 2022. Spike mutations contributing to the altered entry preference of SARS-CoV-2 omicron BA.1 and BA.2. *Emerging Microbes & Infections* [Internet]. Available from: https://www.tandfonline.com/doi/abs/10.1080/22221751.2022.2117098

Hussin J, Nadeau P, Lefebvre J-F, Labuda D. 2010. Haplotype allelic classes for detecting ongoing positive selection. *BMC Bioinformatics* 11:65.

Jankowiak M, Obermeyer FH, Lemieux JE. 2022. Inferring selection effects in SARS-CoV-2 with Bayesian Viral Allele Selection. *PLoS Genet* 18:e1010540.

Johri P, Aquadro CF, Beaumont M, Charlesworth B, Excoffier L, Eyre-Walker A, Keightley PD, Lynch M, McVean G, Payseur BA, et al. 2022. Recommendations for improving statistical inference in population genomics. *PLOS Biology* 20:e3001669.

Johri P, Stephan W, Jensen JD. 2022. Soft selective sweeps: Addressing new definitions, evaluating competing models, and interpreting empirical outliers. *PLOS Genetics* 18:e1010022.

Kaku Y, Okumura K, Padilla-Blanco M, Kosugi Y, Uriu K, Hinay AA, Chen L, Plianchaisuk A, Kobiyama K, Ishii KJ, et al. 2024. Virological characteristics of the SARS-CoV-2 JN.1 variant. *The Lancet Infectious Diseases* 24:e82.

Kannan SR, Spratt AN, Cohen AR, Naqvi SH, Chand HS, Quinn TP, Lorson CL, Byrareddy SN, Singh K. 2021. Evolutionary analysis of the Delta and Delta Plus variants of the SARS-CoV-2 viruses. *J Autoimmun* 124:102715.

Kannan SR, Spratt AN, Sharma K, Chand HS, Byrareddy SN, Singh K. 2022. Omicron SARS-CoV-2 variant: Unique features and their impact on pre-existing antibodies. *Journal of Autoimmunity* 126:102779.

Kaplan NL, Hudson RR, Langley CH. 1989. The “Hitchhiking effect” revised. *Genetics* 123:887–899.

Kelly JA, Woodside MT, Dinman JD. 2021. Programmed −1 Ribosomal Frameshifting in coronaviruses: A therapeutic target. *Virology* 554:75–82.

Khare S, Gurry C, Freitas L, Schultz MB, Bach G, Diallo A, Akite N, Ho J, Lee RT, Yeo W, et al. 2021. GISAID’s Role in Pandemic Response. *China CDC Wkly* 3:1049–1051.

Kimura R, Fujimoto A, Tokunaga K, Ohashi J. 2007. A Practical Genome Scan for Population-Specific Strong Selective Sweeps That Have Reached Fixation. *PLOS ONE* 2:e286.

Kullback S. 1997. Information Theory and Statistics. New edition. Mineola, N.Y: Dover Publications

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular Biology and Evolution* 35:1547–1549.

Labuda D, Labbé C, Langlois S, Lefebvre J-F, Freytag V, Moreau C, Sawicki J, Beaulieu P, Pastinen T, Hudson TJ, et al. 2007. Patterns of variation in DNA segments upstream of transcription start sites. *Hum Mutat* 28:441–450.

Lauterbur ME, Munch K, Enard D. 2023. Versatile Detection of Diverse Selective Sweeps with Flex-Sweep. *Molecular Biology and Evolution* 40:msad139.

Lewontin RC. 1964. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics* 49:49–67.

Luque VJ. 2017. One equation to rule them all: a philosophical analysis of the Price equation. *Biol Philos* 32:97–125.

Luque VJ, Baravalle L. 2021. The mirror of physics: on how the Price equation can unify evolutionary biology. *Synthese* 199:12439–12462.

Mahmood TB, Hossan MI, Mahmud S, Shimu MSS, Alam MJ, Bhuyan MMR, Emran TB. 2022. Missense mutations in spike protein of SARS-CoV-2 delta variant contribute to the alteration in viral structure and interaction with hACE2 receptor. *Immun Inflamm Dis* 10:e683.

Manning CD, Raghavan P, Schütze H. 2008. Introduction to Information Retrieval. Cambridge: Cambridge University Press Available from: https://www.cambridge.org/core/product/669D108D20F556C5C30957D63B5AB65C

Meier JI, Salazar PA, Kučka M, Davies RW, Dréau A, Aldás I, Box Power O, Nadeau NJ, Bridle JR, Rolian C, et al. 2021. Haplotype tagging reveals parallel formation of hybrid races in two butterfly species. *Proceedings of the National Academy of Sciences* 118:e2015005118.

Okasha S, Otsuka J. 2020. The Price equation and the causal analysis of evolutionary change. *Philos Trans R Soc Lond B Biol Sci* 375:20190365.

Pampín M, Casanova A, Fernández C, Blanco A, Hermida M, Vera M, Pardo BG, Coimbra RM, Cao A, Iglesias D, 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* [Internet] 10. Available from: https://www.frontiersin.org/articles/10.3389/fmars.2023.1057206

Panigrahi M, Rajawat D, Nayak SS, Ghildiyal K, Sharma A, Jain K, Lei C, Bhushan B, Mishra BP, Dutt T. 2023. Landmarks in the history of selective sweeps. *Anim Genet*.

Parums DV. 2023. Editorial: The XBB.1.5 ('Kraken') Subvariant of Omicron SARS-CoV-2 and its Rapid Global Spread. *Med Sci Monit* 29:e939580.

Price GR. 1972. Extension of covariance selection mathematics. *Annals of human genetics* 35:485–490.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll

SA, Gaudet R, et al. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* 449:913–918.

Shipilina D, Pal A, Stankowski S, Chan YF, Barton NH. 2023. On the origin and structure of haplotype blocks. *Molecular Ecology* 32:1441–1457.

Smith JM, Haigh J. 1974. The hitch-hiking effect of a favourable gene. *Genetics Research* 23:23–35.

Soni V, Jensen JD. 2024. Temporal challenges in detecting balancing selection from population genomic data. *G3 Genes|Genomes|Genetics*:jkae069.

Soni V, Johri P, Jensen JD. 2023. Evaluating power to detect recurrent selective sweeps under increasingly realistic evolutionary null models. *Evolution* 77:2113–2127.

Stephan W. 2019. Selective Sweeps. *Genetics* 211:5–13.

Vera M, Wilmes SB, Maroso F, Hermida M, Blanco A, Casanova A, Iglesias D, Cao A, Culloty SC, Mahony K, et al. 2023. Heterogeneous microgeographic genetic structure of the common cockle (Cerastoderma edule) in the Northeast Atlantic Ocean: biogeographic barriers and environmental factors. *Heredity* 131:292–305.

Wang X, Lu L, Jiang S. 2023. SARS-CoV-2 Omicron subvariant BA.2.86: limited potential for global spread. *Sig Transduct Target Ther* 8:1–3.

Wang X, Lu L, Jiang S. 2024. SARS-CoV-2 evolution from the BA.2.86 to JN.1 variants: unexpected consequences. *Trends in Immunology* 45:81–84.

Whitehouse LS, Schrider DR. 2023. Timesweeper: accurately identifying selective sweeps using population genomic time series. *Genetics* 224:iyad084.

Zheng B, Xiao Y, Tong B, Mao Y, Ge R, Tian F, Dong X, Zheng P. 2023. S373P Mutation Stabilizes the Receptor-Binding Domain of the Spike Protein in Omicron and Promotes Binding. *JACS Au* 3:1902–1910.