Peer Community Journal

Section: Genomics
Conference: JOBIM

Research article

Published
2025-12-16

Cite as
David Robelin, Sébastien Déjean and Sabine Mercier (2025) *localScore: an R package to highlight optimal and suboptimal segments in a sequence with associated p-values computation*, Peer Community Journal, 5: e141.

Correspondence
sebastien.dejean@math.univ-toulouse.fr

Peer-review
Peer reviewed and recommended by PCI Genomics, https://doi.org/10.24072/pci.genomics.100420

localScore: an R package to highlight optimal and suboptimal segments in a sequence with associated p-values computation

David Robelin [1], Sébastien Déjean [2], and Sabine Mercier [2]

Volume **5** (2025), article e141

https://doi.org/10.24072/pcjournal.650

Abstract

Highlighting atypical segments of a sequence is an important goal in very diverse domains. In the case where no prior information on the length of the segment to be highlighted is known, Karlin and Altschul defined, in 1990, the local score for biological sequence analysis, and an asymptotic approximation of its distribution was proposed in 1992. There are now many other theoretical results that can be used to establish the p-value of the local score in different contexts. We have developed an R package bringing together these results for a sequence modelled by independent and identically or Markovian distributed variables. It calculates the local score, the sub-optimal scores and their positions, and proposes to establish the *p*-value of the local score using the various theoretical methods available to date. An automatic analysis is also proposed to apply the most appropriate method depending on the sequence analyzed. Here we present the software package and various application examples. Comparisons with other tools used depending on the context of the application are also given. The localScore package is available on CRAN under the GPL-2 license (core program) and various licenses for the embedded Eigen library.

[1]INRAE, INPT, ENVT, Université de Toulouse, GenPhySE UMR 1388, Castanet-Tolosan, France, [2]Université de Toulouse, UT2J, UT Capitole, INUC, INSA, CNRS, Institut de Mathématiques de Toulouse, UMR 5219, Toulouse, France

Introduction

Highlighting atypical periods or segments in sequences is an issue of interest in many fields, such as Bioinformatics and Genomics, Biosurveillance, Ecology and Environmental Sciences, Epidemiology and Health Sciences, Finance, Reliability and Quality Control, Telecommunication Sciences, and many others... Scan statistics are a popular approach to address such a problem. But, a major challenge with the scan statistics has been finding analytical results regarding its statistical significance. Indeed, all data collection has natural variability, which can be called "noise," and establishing statistical significance allows us to know whether an observation exceeds this random variability. Calculating statistical significance is a fundamental tool to ensure that observations are not simply the result of chance, but reflect real and relevant phenomena, thus ensuring the reliability and relevance of conclusions. Unfortunately, in the case of scan statistics, the computation required to obtain exact p-values is complex and often impractical (Glaz et al., 2001, p.371). Good approximations have been proposed in special cases (Naus, 1982, Wallenstein and Neff, 1987). The main advantages of scan statistics are: the mathematical tools for establishing theoretical results on the statistical significance of the results allow generalization to dimensions 2 and sometimes 3 with certain model assumptions on the observations; convergence rate results are available in some cases; an asymptotic approximation; results also exist for continuous sequences. However, these advantages do not occur for longitudinal and discrete data. The disadvantages of scan statistics, in our opinion, lie in the choice of a pre-fixed window length and in the fact that the results are limited to precise models on the observations (Poisson, uniform,...). The work of Naus and Wallenstein, 2006 and Nagarwalla, 1996 extends the scan statistics to a statistic with a variable window, whose size does not need to be chosen *a priori*. This work is based on the use of a Generalised Likelihood Ratio Test (GLRT) and allows the window size to be varied within a given interval. Nargawalla (Nagarwalla, 1996) provided a simple algorithm for implementing his method, and the Monte Carlo hypothesis testing is used to obtain the p-value, since it is impossible to obtain the null distribution for the statistic under consideration by simulation. These questions about variable-width scan statistics are still relevant (Cucala, 2017; Glaz et al., 2001; Wang and Glaz, 2014). Cucala (Cucala, 2008; Cucala, 2017) proposes a statistic for variable-width scan statistics for longitudinal data not relying on usual likelihood ratio methods, and statistical significance is again established by appropriate simulation.

In comparison, the local score allows one to completely avoid the choice of window size; the statistical significance on its null distribution can be established theoretically by any model on independent and equally distributed observations; and theoretical results exist for the Markov-like dependence models on the data succession. Karlin and Altschul, 1990 defined the local score statistic to analyze biological sequences: it corresponds to the maximum cumulative value of a given property over every possible segment in a sequence, considering segments of any position and any length (see Equation (1)). Karlin and Dembo, 1992 proposed asymptotic approximations of the distribution of the local score when the length of the sequence is growing to infinity, thus the statistical significance can be computed. A generalization of this approximation for the sequence comparison case has been developed in BLAST Software[1], but to our knowledge no development has been done for a single sequence analysis case. At present, results exist that consider independent or dependent models on the sequence. Those results include: improvements of the approximations of Karlin *et al.* (see Cellier et al., 2003 for the independent model and Grusea and Mercier, 2020 for the Markov model); exact methods (see Mercier and Daudin, 2001 for the independent model and Hassenforder and Mercier, 2007 for the Markov model); a result on the distribution for the pair of the local score value and the length of the segment that realizes the local score value (see Chabriac et al., 2014; Lagnoux et al., 2017). We developed the package `localScore` (Simon et al., 2023) for the software `R` (R Core Team, 2024). In the package, we focus on the different ways to establish the statistical distribution of the local score in a sequence modeled by independent and identically distributed (I.I.D.) random variables, or by Markovian random variables. For a more elaborated model of the sequence, a Monte Carlo function is available, provided that the user can simulate sequences from his model. The local

[1] https://blast.ncbi.nlm.nih.gov/

score corresponds to the optimal segment achieving the maximum score. However, sub-optimal segments are also interesting, but their common analysis leads to a multiple test. In the case of the I.I.D. model, it was shown (Fariello et al., 2017, supplementary material) that the distribution of the local score is the one to use on all sub-optimal segments of the sequence.

The remainder of the article contains a brief presentation of the main theoretical backgrounds implemented in the `localScore` package. Then the package is described in Section Software through a standard workflow to follow and a description of the main functions. Section Illustrations presents examples of using `localScore` in four different domains for which we compare the results of usual methods in the corresponding area: biological sequence analysis and a comparison with sliding window statistics; a signal detection context and a comparison with control charts; epidemiology and a comparison with scan statistics; a genomic sequence analysis. The first example on biological sequences, and the last one on genomic sequences, also show the use of local score distribution for a multiple test on suboptimal segments.

Theoretical background

Let us consider a sequence as a succession of components that belong to a finite set $\mathcal{A}$. It can be for example a DNA sequence with $\mathcal{A} = \{A,\ C,\ G,\ T\}$. Let us define a score scheme or a score scale as a function s that assigns a real number to any letter of $\mathcal{A}$. The score can, for example, quantify a physico-chemical property. See the website of Protscale[2] for an illustration of different score scales in biological sequence analysis context. Let $\mathbb{A} = (A_i)_{1 \leqslant i \leqslant n}$ be a sequence, and let us denote $X_i := s(A_i)$, for $i \geqslant 1$, the scoring sequence associated to the sequence $\mathbb{A}$ based on the score function s. Examples of scoring functions are presented in Section Illustrations.

With $X_0 := 0$, the local score M_n of one sequence $\mathbb{X}$ of length n is defined by

$$M_n := \max_{0 \leqslant i \leqslant j \leqslant n} \sum_{k=i}^{j} X_k \ . \tag{1}$$

In Mercier and Daudin, 2001 the authors proved that the local score can also be defined as: $M_n := \max_{0 \leqslant i \leqslant n} U_i$ with $U_0 := 0$ and $U_{i+1} := \max(U_i + X_{i+1}, 0)$ the Lindley, or CUSUM process, associated to the sequence $(X_i)_{1 \leqslant i \leqslant n}$. The Lindley process defines nonnegative excursions and the height of the highest one is equal to the local score. The other excursions are called the sub optimal segments.

The local score approach makes it possible to avoid choosing a segment length in the absence of prior information about it. Let us present below the two main kinds of results, approximation and the exact method, when the random variables $(A_i)_{1 \leqslant i \leqslant n}$ are independent and identically distributed (I.I.D.) and so are the $(X_i)_{1 \leqslant i \leqslant n}$.

Karlin and Dembo approximation

The asymptotic approximation of Karlin and Altschul, 1990 and Karlin and Dembo, 1992 corresponds to an asymptotic result converging to a Gumbel distribution when the length of the sequence n tends to infinity. This result stands on the two following hypotheses: The average score must be nonpositive, $\mathbb{E}[X] < 0$, and a nonnegative score must be possible, $\mathbb{P}(X > 0) > 0$. We have

$$\lim_{n \to +\infty} P\left(M_n \leqslant \frac{\ln n}{\lambda} + x\right) = e^{-K^* e^{-\lambda x}} \tag{2}$$

where λ and K^* depend on the score distribution. For a score distribution denoted as follows with $p_i = P[X = i]$ for $i = 0, 1, \ldots, u$, and $q_j = P[X = -j]$ for $j = 1, \ldots, v$, and $\sum_{i=0,\ldots,u} p_i + \sum_{j=1,\ldots,v} q_j = 1$, let us define the following polynomial $P(x) = x^u \cdot (E[x^{-X} - 1])$. We get

$$P(x) = \sum_{i=1}^{u} p_i \cdot x^{u-i} + (p_0 - 1) \cdot x^u + \sum_{j=1}^{v} q_j \cdot x^{u+j} \ , \tag{3}$$

[2] https://web.expasy.org/protscale

of degree $u + v$ the range of possible scores. Under the hypothesis $E[X] < 0$ (see proposition 1 [Distribution for the maximal partial sums] of Cellier et al., 2003; Karlin and Dembo, 1992), the polynomial P has only two real positive roots that are of simple multiplicity: 1 and R with $0 < R < 1$. The parameter $\lambda = \ln(1/R)$ and K^* is obtained using the other roots of P.

At present, the method used to deduce the roots of the polynomial is based on the Bairstow method (Bairstow, 1920).

The parameter λ corresponds to the single root of a polynomial of degree equal to the amplitude of the scores (maximum score minus minimum score) and checking $\mathbb{E}[\exp(\lambda X)] = 1$. The existence of λ is ensured by the assumption $\mathbb{E}[X] < 0$. The set of other roots is also used for the calculation of K^* notably using a square matrix called Vandermonde comprising at each line a geometric progression associated with one of the roots of the polynomial. This gives the following approximation for an observed local score a

$$P(M_n \leqslant a) \approx e^{-K^* n e^{-\lambda a}} . \tag{4}$$

The approximation given in (4) is very accurate for sequence lengths larger than thousands and very fast to obtain, but must be avoided for sequences of less than a hundred components.

Karlin by Monte Carlo

The Karlin and Dembo approximation in (2) calculates the value of two parameters λ and K^* according to the values and the distribution of the scores. Here we propose to estimate them using a Monte Carlo approach. This method is useful in the case of a sequence of scores that is too long to perform a direct Monte Carlo as it does not require simulating full-length sequences. Formula (4) can be linearized in λ and K^* using logarithms as long as n is large enough. That leads to the following formula:

$$\ln\{-\ln\{P(M_n \leqslant a)\}\} \approx \ln K^* - \lambda a + \ln n . \tag{5}$$

Given the previous formula, the Karlin by Monte Carlo procedure consists in:

(1) Choosing a sequence length n_{sim} for the simulation big enough to have a satisfying Karlin and Dembo approximation and small enough to be computed with reasonable resources.
(2) Simulating sequences of size n_{sim}.
(3) Calculating the local score of each sequence to derive empirical distribution function of the local score for sequences of size n_{sim}.
(4) Deriving estimation of λ and K^* by a linear regression on the empirical distribution function using Formula (5), i.e., $\hat{\lambda} = -\hat{b}$ and $\hat{K^*} = \exp(\hat{a})/n_{sim}$ where $\hat{a}$ and $\hat{b}$ are respectively the slope and the intercept of the regression.
(5) Using Karlin and Dembo approximation to calculate the p-value of the local score observed on the full sequence of size n.

Daudin

For a an observed local score value, the exact method in the I.I.D. case is based on an appropriate stopped process constructed to be a Markov chain and taking its values in $\{0, \ldots, a\}$. Let us denote $P = (P_{ij})_{0 \leqslant i,j \leqslant a}$ its corresponding transition matrix. Mercier and Daudin, 2001 proved that

$$(\forall a \geqslant 0) \quad \mathbb{P}(M_n \geqslant a) = (P^n)_{(0,a)} . \tag{6}$$

There is no restriction on the sign of the average score when using the exact method. This method is accurate and sufficiently fast as long as the value n and the value a for which the p-value is calculated are not too large. In fact, a square matrix of size a is store into memory and exponentiated to the value n. As there is a limit to the exponentiation of P for n tending towards infinity, it is not necessary to use the correct value of n for sequences longer than a hundred thousand, to obtain an accurate value, but a smaller value can be used.

Improved approximation of Karlin et al. in the I.I.D. case

An improved approximation to that proposed in Karlin and Dembo, 1992 is proposed in Cellier et al., 2003. As with the Karlin *et al.* method, it is necessary to compute the roots of the same polynomial which are then used in several steps to compute the additive correction terms in order to improve the Karlin *et al.* approximation. For large a values, we have

$$P(M_n \leqslant a) \approx (1 - \sum_{i=1} K_i R_i^a)^{\frac{n}{\mu}+1} \tag{7}$$

with $(R_i)_i$ the roots of module strictly less than 1 of a polynomial directly defined with the score distribution (see Formula 3). The degree of this polynomial is equal to the range of possible scores. Based on the two assumptions used in the work of Karlin and Dembo, there is a unique positive real root of modulus less than 1, $e^{-\lambda}$, with λ defined in Equation (2). The parameters K_i and μ are also derived from the distribution of scores and from calculations based on the Vandermonde matrix and certain resolutions of systems of equations. The improved approximation in (7) is accurate and fast for n values of several hundred, but should be avoided for sequence lengths of less than a hundred.

All the above theoretical results must be considered as complementary for practical application according to the scoring scheme, with its range, the sign of the average score and the length of the sequence to be analyzed.

Software features and contents

Workflow

A tentative workflow using `localScore` might look like this:

(1) Transform the component of a given set of sequences into score sequences using a given score function.
(2) Learn the distribution of scores on score sequences.
(3) Compute the local score of each sequence.
(4) Calculate the corresponding p-values using the automatic method for the calculated local score value, the corresponding sequence length and the global score distribution.

Main functions

Following the workflow outlined above, here are the main functions that can be used at each stage.

To obtain a score sequence: A sequence of components, such as DNA, can be transformed into a sequence of scores using the `CharSequence2ScoreSequence` function. Integer or real scores can be taken into account.

To learn a distribution: Several functions can be used to learn the distribution of the components of given sequences or scores. For example, the empirical distribution of a numerical sequence or a list of sequences is constructed by `scoreSequences2probabilityVector`.

To calculate the local score: The function `localScoreC` calculates the local score for a sequence of integer or real scores. It provides the local score and all sub-optimal segments with their associated scores. The functions `suboptimalSegment` or `lindley` can be used to obtain the others locations of the different realisations of the local score. `lindley` is also useful to graphically representing the interersting region of the sequence.

To calculate the corresponding p-values: Next, the following functions offer different methods for calculating the p-values associated to the local score of a sequence:

- `karlin`: The Karlin *et al*'s approximation (see (4)). This method requires a nonpositive average score, $\mathbb{E}[X] < 0$, and integer scores, and is best suited to long sequences with length greater than a few thousand components, depending on the expectation of the score distribution.

- `mcc`: An improved approximation of the previous one, presented in Cellier et al., 2003. This method also requires a nonpositive average score, $\mathbb{E}[X] < 0$, and integer scores. It is more suitable for sequences with a few hundreds components, depending on the expectation of the score distribution.
- `daudin`: An exact method for integer scores is also integrated and can be used regardless the sign of the expected score (see (6)). This method is computationally suitable for not too long sequences, but several thousand components can be easily handled. The implementation is based on the exponentiation of a square matrix of size a, with a a given local score value.
- `monteCarlo`: A classical Monte Carlo method taking a random generator function as parameter.
- `karlinMonteCarlo`: A mixture of the Karlin *et al.*'s and the Monte Carlo methods. It allows for an approximate distribution with a shorter computation time than the empirical Monte Carlo method, for very long sequences. This mixed method also requires $\mathbb{E}[X] < 0$.

We have also developed the `automatic_analysis` function for less experienced users. As its name suggests, this function automatically selects the most appropriate p-value method for the data entered by the user according to the configuration described in Table 1. The function calculates the p-value based on the length of each entered sequences. It can use an empirical score distribution based on the input data or a distribution provided by the user. By setting the `method_limit`, the user can also decides up to which sequence length the computationally intensive methods (`daudin`, `exact_mc`) should be used to calculate the p-value.

Table 1 – Adequate methods to compute the local score p-value depending on the average score value $\mathbb{E}[X]$ and the sequence length n order ; with E : `daudin()` ; MCC : `mcc()` ; K : `karlin()` ; MC : `monteCarlo()` ; MC-K : `karlinMonteCarlo()`.

n	< 100	$10^2 \leqslant \cdot < 10^3$	$10^3 \leqslant \cdot < 10^4$	$\geqslant 10^4$
$\mathbb{E}[X] < 0$	E ; MC	E ; MCC ; MC	E ; MCC ; MC	MCC ; K ; MC ; MC-K
$\mathbb{E}[X] \geqslant 0$	E ; MC	E ; MC	E	

Inputs / outputs

Inputs. When the workflow starts, the first input is a sequence. This can be imported into `R` from an ASCII file using standard reading functions such as `read.table` and related functions. For users wishing to analyze biological sequences composed of nucleotides or amino acids, the package can also handle FASTA files as input. In FASTA files, each sequence is preceded by a title (marked with a ">") and a line break. A sequence occupies one line, followed by a line break and a line containing only a tab.

Additionally, if no sequence is passed to the `automatic_analysis` function, it allows the user to choose a FASTA file. In this case, and if the user has not provided a scoring system (which can be done by passing a named list with the appropriate scores for each character), the second file dialog box appears. The latter allows you to choose a file containing the score, and if the user provides an additional column for probabilities, these are also used - see the "File Formats" section in the vignette for details.

Score files can also be imported in a standard way from an ASCII file. Such a file must contain a header and each row contains a letter and its score. Optionally, a probability for each score can also be provided.

Numerical outputs. The main numeric output is provided by the `localScoreC` function. It contains a list with the following attributes:

- The value of the local score and the start and end indices of the segment achieving this optimal score.
- All local maxima of the Lindley process (only strictly positive excursion) and their start and end indices.

- The recording times of the Lindley process.

Any method calculating *p*-values only provides the value obtained.

Graphical outputs. Graphical outputs can be optionally displayed by the `monteCarlo` and the `karlinMonteCarlo` functions. They represent the distribution of all simulated local scores and the cumulative distribution. The Lindley process plot provides a convenient representation of the data.

Example data

Some data we propose to analyze in the section are already integrated into the package for illustration purposes. `Seq1093` is a real biological sequence with 1093 characters referring to Q60519 queries in UniProt Data base[1]. `SeqListSCOPe` contains 285 protein sequences ranging in length from 31 to 404. They are referenced as `CF_scop2dom_20140205aa` in the Structural Classification Of Proteins database (SCOP)[2]. `SJSyndrome` is a dataset of 824 rows, each describing an appearance of Stevens-Johnson syndrome described by 15 covariates including Case ID, Initial FDA Received Date, days since last FDA. The third column is the number of days between two adverse events. `Aeso` consists of the individual birthdates of over 35 cases of the congenital oesophageal and tracheoesophageal fistula malformations seen at Birmingham hospital.

Illustrations

We illustrate the use of the `localScore` package on four examples in different fields. First, one of the biological sequences embedded in the package is used as a toy example to show the basic usage of the package. Similarly, we illustrate how to deal simultaneously with a set of sequences. Then, we analyze two medical data sets to show how the local score can be used to detect an eventual shift in sequential observations. The last subsection deals with studying a chromosome to associate genomic regions with phenotype differentiation. We also present, for each case, the results of other methods. The R scripts are provided in Robelin et al., 2025 (Appendixes).

Biological sequences

We first describe how to analyze one single sequence and then show how to process a set of multiple sequences at once.

One single sequence. See Robelin et al., 2025 (Appendix A1-3) for corresponding code.

Several sequences are already included in the package. Let us use the object `Seq1093`, corresponding to the protein Q60519 SEM5B_MOUSE[3]. With 1093 characters, we consider this to be a sequence whose statistical significance can be established by almost all proposed methods (see Table 1).

The `CharSequence2ScoreSequence` function converts the character sequence into a score sequence using the `HydroScore` object, which provides the correspondence between an amino acid and its score according to Kyte & Doolittle hydrophobic scoring scale (Kyte and Doolittle, 1982). The local score computation can then be performed with `localScoreC`, to provide the following result.

```
$localScore
value begin   end
   65   956  1001

$suboptimalSegmentScores
      value begin  end
 [1,]    40     1   20
```

[1]https://www.uniprot.org
[2]https://scop.mrc-lmb.cam.ac.uk/
[3]https://www.uniprot.org/uniprot/Q60519

```
 [2,]    10    71   73
 [3,]    23    80   99
[...]
[75,]     3  1079 1079
[76,]     2  1083 1083
[77,]     6  1089 1090

$RecordTime
  [1]    0   70   77   78   79  112  113  115  [...]
 [19]  129  176  177  178  179  180  218  219  [...]
[...]
[181] 1055 1061 1062 1063 1069 1070 1071 1072 [...]
[199] 1093
```

We retrieve only the local score value for further use when calculating the *p*-value.

To do so, the function `scoreSequences2probabilityVector` builds an empirical distribution from the sequence.

```
   -5    -4    -3    -2    -1     0     1     2     3     4     5
0.074 0.203 0.020 0.075 0.212 0.078 0.000 0.071 0.094 0.144 0.028
```

The exact method (see (6), function `daudin`) can then be used to compute the *p*-value. The approximate method of Karlin *et al.* (see (4)) can be performed equivalently with the `karlin` function.

The two *p*-values are quite close (0.072 for the exact method, 0.076 for the approximate one).

In comparison, here are the results obtained with ProtScale Expasy web tool on the same sequence. ProtScale computes and represents the profile on a selected protein produced by any amino acid scale and accumulates the score values over a sliding window of a chosen size. Note that the possible window sizes are restricted to odd values from 3 to 21. We used the hydropathicity scale proposed by Kyte and Doolittle, 1982. We chose a size equal to 21 which is the closest value to the length of the optimal segment given by the local score approach without any prior information on it. The results are presented in Figure 1. We can observe one main peak and the numerical output (not shown) gives us a window value equal to 2.195 and a center index equal to 989 (begin index 979; end index 999). This segment corresponds to the one highlighted by the local score but with a length equal to 45 with a beginning index of 956 and an end index of 1001. The local score *p*-value allows us to say that this region is not statistically significant. For a window size equal to 9 corresponding to the one given by default, we can observe several picks with a similar value before the one we discussed previously. We have also represented the corresponding Lindley process using the `lindley` function.

A set of sequences. See Robelin et al., 2025 (Appendix A1-3) for corresponding code.

The data consists of a list of 285 character strings with their entry codes as names extracted from the Structural Classification Of Proteins database (SCOP)[2]. More precisely this data contain the 285 protein sequences of the data called “CF_scop2dom_20140205aa” with sequence length from 31 to 404.

This sequence is a part of the package and can be loaded and briefly explored. For instance, here is the first sequence of the set (`P50456`). It is composed of 165 characters.

```
P50456
"ARDVIQVVIDHNVGAGVITDGHLLHAGSSSLVEIGHTQVDPYGKRCYCGNHGCLETIAS
VDSILELAQLRLNQSMSSMLHGQPLTVDSLCQAALRGDLLAKDIITGVGAHVGRILAIMV
NLFNPQKILIGSPLSKAADILFPVISDSIRQQALPAYSQHISVEST"
```

Overall, accross all sequences, sequence lengths range from 31 to 404 with a median length of 102 and a mean length of 122.

[2]https://scop.mrc-lmb.cam.ac.uk/

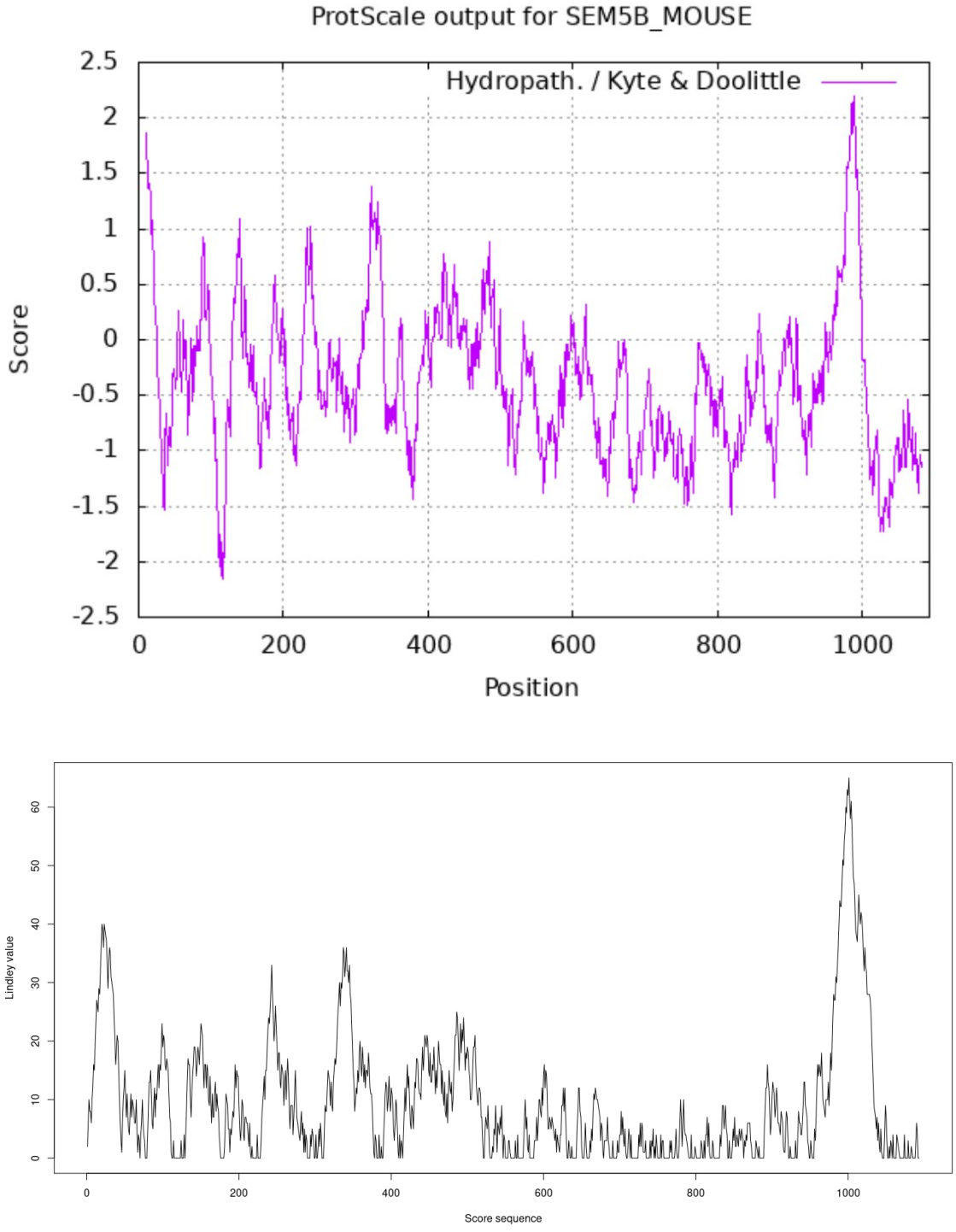

Figure 1 – Top: Graphical output of the results provided by the Expasy ProtScale web tool for the corresponding sequence Q60519, the Kyte and Doolittle scale and a window size equal to 21. Bottom: Lindley process calculated with the `localScore` package.

The `CharSequence2ScoreSequence` function transforms the protein sequence into a score sequence using the `HydroScore` object. The score corresponding to each amino acid is:

A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
2	3	−4	−4	3	0	−3	5	−4	4	2	−4	−2	−4	−5	−1	−1	4	−1	−1

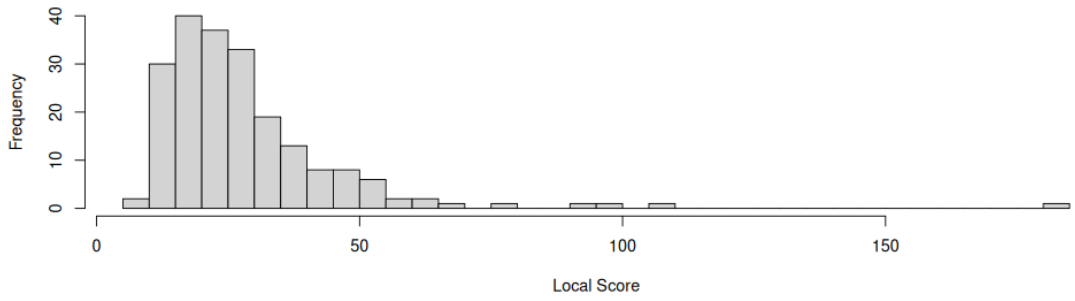

Figure 2 – Distribution of the local score of every sequence in the object `MySeqList`.

Then we use `automatic_analysis` function to perform the most appropriate method to compute the p-value of the local score of each sequence.

The results can then be investigated. For instance, the first sequence of the list has a local score value equal to 62, and the segment that realizes this maximum begins at index 4 and finishes at index 144. Its p-value equals 6.39%.

We can easily extract the first 10 p-values, the 5 smallest p-values, the significant sequences, or their local score values using the `sapply` function as indicated in Robelin et al., 2025 (Appendix A1-3).

Figure 2 displays the distribution of the local scores computed on the different sequences.

The methods used to calculate p-values for each sequence are stored in the component `method` of the output of the `automatic_analysis` function.

Since the maximum sequence length is 404, it is expected for the exact method to be used for all sequences in the database. The score distribution used to calculate the p-value for each local score is the empirical distribution estimated over the entire data set. It can be represented using the `scoreSequences2probabilityVector` function.

```
   -5    -4    -3    -2    -1     0     1     2     3     4     5
0.055 0.264 0.022 0.041 0.148 0.072 0.000 0.105 0.052 0.175 0.067
```

Stevens-Johnson syndrome data

In other fields, such as Telecommunication Sciences or Quality Control to name just two, where the goal is to highlight a change or a breakpoint in the signal sequence, the data are analyzed as soon as they are collected. In these application fields, no score scales is proposed or constructed as is the case in biological sequence analysis. When testing at each time i, the null hypothesis H_0: "The observations $(A_k)_{1\leqslant k\leqslant i}$ follow the distribution f_θ with parameter $\theta = \theta_0$" vs H_1: "The observations $(A_k)_{1\leqslant k\leqslant i}$ follow f_{θ_1} with $\theta_1 \neq \theta_0$", it is common to define the score of a given observation A_i at time i by the following Log Likelihood Ratio:

$$x_i = s(A_i) = \ln\left(\frac{f_{\theta_1}(A_i)}{f_{\theta_0}(A_i)}\right) .$$

Such a score function is used in this subsection and in the Subsection *Congenital oesophageal atresia data*.

The local score can also be used to detect a potential shift in sequential observations. Here, we propose to analyze data relating to the onset of Stevens-Johnson syndrome, a serious dermatological disease due to a drug allergy. The detection of an atypical cluster of Stevens-Johnson syndrome cases can allow us to analyze the possible causes of these exceptional occurrences of side effects. It is reasonable to assume that the cause, or causes, common to all patients, whether directly related to the treatments or external to them, are more easily detected when an atypical cluster appears. Preventive actions can then be proposed and/or implemented to limit such occurrences in the future.

The data present 824 occurrences of adverse events leading to 823 values of Time Between two adverse Events (TBE) in days. The third column of the dataset `days.since.last.fda` corresponds to the number of days elapsed since the last event (sequence of the Time Between Events). See appendix A2 in Robelin et al., 2025.

The TBE sequence can be modeled by a geometric distribution. An estimation of its parameter is given by

```
R> p0Hat <- 1/(mean(DatesTBE[1:n]) - 1)
```

The estimated value (0.0453) corresponds to the probability of observing an adverse event on a given day among the entire population studied. Let denote $(T_i)_{1\leqslant i\leqslant n}$ the TBE observations. At each time i, we wish to test the following hypotheses : H_0 "The observations $(T_k)_{1\leqslant k\leqslant i}$ follow a geometric distribution with parameter p_0" vs H_1 "The observations $(T_k)_{1\leqslant k\leqslant i}$ follow a geometric distribution with parameter $p_1 = 1.5 \cdot p_0$", with p_0 and p_1 in $]0, 1[$.

Let us define:

$$LLR(T) = \ln \frac{f_1(T)}{f_0(T)}$$

with f_j the probability density function of a geometric variable of parameter p_j for $j = 0, 1$. At each time i, the local score of the sequence $(LRR(T_k))_{1\leqslant k\leqslant i}$ and its corresponding p-value are calculated using the package. More precisely, we calculate $LLR = \lfloor E \cdot \ln \frac{f_1(T)}{f_0(T)} \rfloor$ with E a tuning parameter that allows a larger range of possible nonnegative scores. Using this tuning parameter does not change the segment that achieves the local score or its p-value (see Fariello et al., 2017 Supplementary materials, for more details), but it does highlight suboptimal segments that might be of interest. Here, we have at least 3 nonnegative scores for $E = 8$.

We calculate the sequence of scores and the local score for each sequence up to index i for sequential analysis. An alarm can be set when the p-value of an observed local score value is below a given nominal level, typically 5% or 1%. To establish the p-value, it is necessary to know the distribution of scores under the hypothesis H_0. This distribution can be established theoretically, but to simplify presentation, we empirically estimate the distribution of scores on the data. The first and last six empirical estimates are:

```
[1] -109  -75  -61  -56  -54  -46
0.0012151 0.0012151 0.0012151 0.0036452 0.0012151 0.0012151
[1] -2 -1 0 1 2 3
0.036452 0.093560 0.117861 0.134872 0.227217 0.160389
```

Note that not all possible values between the minimum and the maximum scores are present (e.g. between -109 and -75). The vector of the score distribution must be fulfilled with 0.

The average score under H_0 is negative (-1.633) so any method, exact as approximate, can be used to calculate the statistical significance of the local score. We use the exact method with `daudin` function as the sequence lengths allows.

Figure 3 illustrates the example on the first 300 observations where a first alarm, using a nominal level $\alpha = 5\%$, appears at index 229.

In order to assess the robustness of the analysis, different values of the parameter p_1 were tested (p_1=0.05, 0.048 and 0.055 with adapted values for E to obtain at least 3 nonnegative scores, E=29, 47 and 15 respectively). Each case leads to a similar result. Figure 3 representing the p-values at each index, can be considered as a control chart generally used to analyze online sequences in industrial data (see for example the first and the most famous control chart defined in 1930 and called the Shewhart chart, see Shewhart, 1931 mainly used for the Gaussian distribution). We can observe a single clear alarm at index 229. In Mercier, 2020 for a Gaussian model, it is shown that using the local score avoids false alarm better than usual control charts and allows detecting existing parameter variation in a competitive average time.

Let us observe the g chart, a Shewhart chart adapted for geometric distribution, and proposed in the package `qcc` in Figure 4.

Here the lower control limit (LCL) is equal to 0 and has no direct use. The twelve points upper than the upper control limit (UCL) in red are not "bad" alarms as they correspond to a longer-than-expected serie between two adverse events and are therefore considered an improved situation.

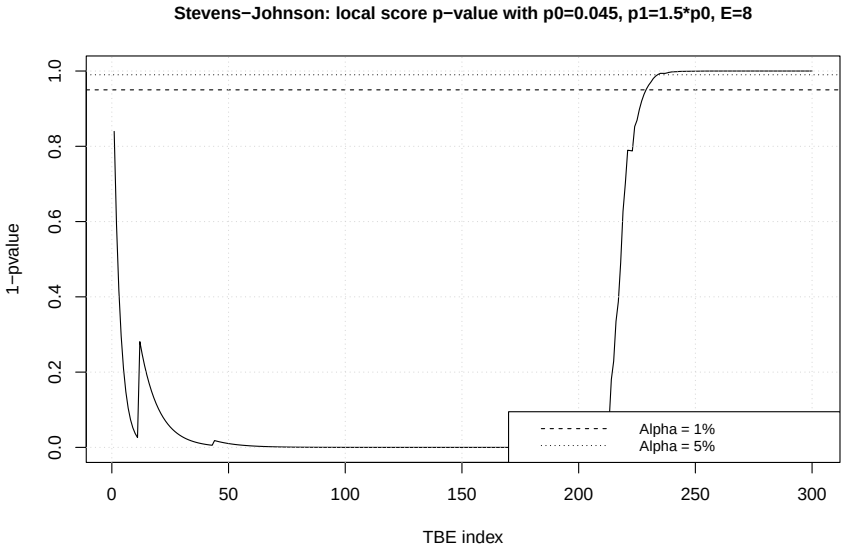

Figure 3 – Stevens Johnson syndrome: a unique alarm at index 229.

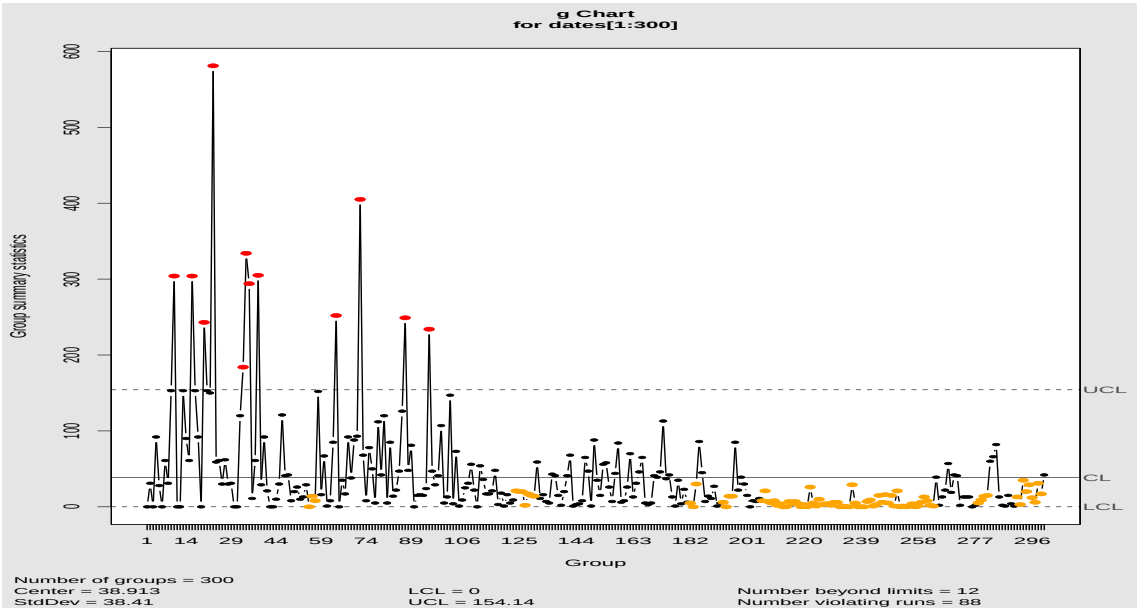

Figure 4 – Stevens Johnson syndrome - Shewhart g chart: A lot of alarms are pointed out. We can observe a violating run, corresponding to a particularly numerous successive points under the central control limit, beginning at index 206 and including the index 229 of the alarm of the local score chart.

We observe several violations run in orange, corresponding to a particularly high number of successive points below the central control limit, which are considered alarms. One of these is particularly long, starting at index 206 and including index 229 of the local score plot alarm.

Regarding the local score and g plots, we suggest that the series of points below the lower control limit before index 206 in the g plot could be considered false alarms.

Congenital oesophageal atresia data

The data include the individual birthdates of $n = 35$ cases of congenital esophageal and tracheoesophageal fistulas seen in a hospital in Birmingham, UK, over a period of 2191 days, from 1950 to 1955, the first day being January 1, 1950 (see Knox, 1959). Glaz et al., 2009

presents in Chapter 17 different works on these data, based on the use of scan statistics. In this section, we first present the results of the scan statistics analyses proposed in Glaz et al., 2009, and then two different approaches based on the local score. The discrete scan statistics $S_{n,k}$, with $k \leqslant n$ two positive integers, of a sequence $(S_i)_{1\leqslant i\leqslant n}$ of n binary trials (1: success, 0: failure) has been defined as the maximum number of successes among k consecutive trials. Consider a discrete sequence $(S_i)_{i=1\ldots n}$. We have $S_{n,k} = \max_{1\leqslant i\leqslant n-k+1} \sum_{j=i}^{i+k-1} S_j$. We also deduce the Time Between two Events (or "successes"). The TBE sequence is modeled by a geometric distribution with parameter $p = 0.016$. The R code is provided in Robelin et al., 2025 (Appendix A1-3).

Scan statistic approach. Considering the date sequence, Glaz et al., 2009 gives the values of the scan statistics for different choices of window length k; the corresponding statistical significance; and the position of the window that realizes the maximal value. They also present the method of Nagarwalla, 1996 using a scan statistics with a variable window size for which the statistical significance is established by the Monte Carlo method. The results are presented in Table 2.

Table 2 – Results of the scan statistic approaches.

k	value	p-value	begin	end
100	7	0.08833	1233	1305
200	10	0.04993	1233	1390
300	15	0.00141	1233	1491
365	16	0.00271	1233	1583
Nagarwalla	15	0.00580	1233	1491

We can observe that the different statistical significances are very different depending on the choice of the window size: Using nominal level equal to 1%, we obtain not significant p-values for a window size $k = 100$ or $k = 200$ and significant values for $k = 300, 365$ and for the Nagarwalla's method.

Log Likelihood Ratio test and local score approach. Here, we propose to consider a possible drift of the parameter p. Consider H_0: $p = p_0$ and H_1: $p = p_0 \cdot (1+\delta)$ for a given δ value. First consider $\delta = 5\%$.

We associate to the Time Between Events sequence (called `tbe` in the code provided in Robelin et al., 2025 (Appendix A3), the following sequence of scores, calculated with

$$X(tbe) = \lfloor E \cdot \ln\left(\frac{f_1(tbe)}{f_0(tbe)}\right)\rfloor$$

with f_i being the probability of a geometrically distributed random variable with parameter p_i, for $i = 0$or1; E is a tuning parameter we have previously presented in Subsection *Stevens-Johnson syndrome data*. Here, it is rated at 62.

This leads to the following sequence of scores:

```
-6  -5  -4   1 -21  -2  -2  -3   2   3   2   2   2   1  -1
 2   2   0   2   2   2   1  -2  -3   2  -2  -4   0  -1   2
 2   2   1   0   1
```

Score distribution. Let us calculate the distribution of scores in two ways: the first, based on the estimation of the occurrence of the score on the observed sequence, and the second, based on theoretical work on a geometric model, which leads to a more precise distribution. The first six and last six values of the probability score and the theoretical score are

```
    ProbScore   ProbScoreTheo
-21 0.02857143   0.0001725077
-20 0.00000000   0.0002380570
-19 0.00000000   0.0003285136
-18 0.00000000   0.0004533418
-17 0.00000000   0.0006256021
```

```
-16 0.00000000    0.0008132325
 [...]
-2 0.11428571      0.07592718
-1 0.05714286      0.09869924
 0  0.08571429      0.14228152
 1  0.14285714      0.19634549
 2  0.37142857      0.27095262
 3  0.02857143      0.01597444
```

Next, for the given offset δ, we calculate the local score value and its p-value using the two different score distributions. Since the sequence length is very short, $n = 35$, we use the exact method to determine the p-value with the function `daudin`. The start and end indices are also given.

```
$localScore
value begin   end
   22     9    22
$suboptimalSegmentScores
     value begin end
[1,]     1     4   4
[2,]    22     9  22
$RecordTime
[1] 0 1 2 3 5 6 7 8
```

The segment that achieves the local score value begins at date index 1233 and ends at date index 1491, which corresponds to the segment highlighted by the scan statistics approach with a window size $k = 300$. Its statistical significance for the observed local score is approximately 0.026.

One could argue that the choice of the window length in the scan statistics method is not necessarily the same as in the local score method. However, the choice of a δ value to construct the scoring function based on p_0 and p_1 is different. Without prior knowledge of the length of the segment to be highlighted, it is easier to choose the smallest drift to detect. Let us consider a set of different δ values, ranging from 1% to 5%.

Table 3 – Value of the local score, its statistical significance, the position of the start and end indices in the sequence *tbe* and in the sequence *date*, for different values of δ. We also give in the second column the tuning parameter E used to obtain at least three nonnegative scores.

δ	E	Local score	$p-$value	b.tbe	e.tbe	b.date	e.date
0.01	302	22	0.02962169	9	22	1233	1491
0.02	152	22	0.02835914	9	22	1233	1491
0.03	102	22	0.02757781	9	22	1233	1491
0.04	77	22	0.02735163	9	22	1233	1491
0.05	62	22	0.02647577	9	22	1233	1491

We can observe in Table 3 that the value of the local score does not change and neither the segment that achieves the local score: See $b.tbe$ (*respectively* $b.date$) the start index in the tbe (*resp.* date sequence and see $e.tbe$ (*respectively* $e.date$) the end index in the tbe (*resp.* date) sequence. Moreover, the statistical significance is quite constant and around 3%.

Direct analysis on the 0-1 sequence. Below, we propose to analyze the initial sequence of occurrences (0-1) without constructing the TBE sequence. The model is then based on a Bernoulli distribution, still with parameter $p_0 = 0.0159$. Consider a drift $p_1 = 1.05 \cdot p_0$. We have two different score values: $(\ln(p_1(1-p_0)/(p_0(1-p_1))) + \log((1-p_1)/(1-p_0)))$ corresponding to 1 and $\ln((1-p_1)/(1-p_0))$ to 0.

These first score values lead us to choose a tuning parameter E equal to 1000 in order to maintain the proportion between these two values and obtain integer values allowing the use the exact method. Remember that this modification of the score function only affects the local score value, but not the segment that achieved it, nor the statistical significance. The two possible scores are then: -1 and 48. We observe that the segment that achieved the local score is still the same as with the geometric model with a starting index of 1233 and an ending index of 1491.

Let us examine the statistical significance. But first, let us give the distribution of the scores. The probabilities under H_0 associated with the scores are equal to $1 - p_0$ for -1 and p_0 for 48.

The obtained p-value (0.0250) is comparable to that of the previous study using the local score approach and the geometric model on the TBE sequence.

Both studies based on the local score highlight the same segment as the sweep statistic with a window size of 300 and that highlighted by the Nagarwalla's method with a variable window. This segment is statistically significant in each method. The local score avoids the need to choose a window length and allows for theoretically establishing statistical significance.

Genomic regions associated with phenotypic differentiation of European local pig breeds

The original dataset is based on European local pig breeds characterized genetically using polled DNA sequencing data, and phenotypically using breed-level phenotypes related to stature, fattening, growth, and reproductive performance. It is composed of 19 populations of European local pig breed populations and 7 industrial breeds populations. Genetic diversity is assessed through a medium-density SNP (Single Nucleotide Polymorphism) leading to 16,403,270 SNPs spanning 18 chromosomes of the pig genomes after filtering out SNPs with missing data. The second part of the original dataset consists of phenotypic characterizations of each breed combined into four distinct groups summarizing stature, fattening, growth, and reproductive performance. The objective of the study published in Poklukar et al., 2023 is to detect genomic regions exhibiting signatures of selection linked to phenotypic traits in order to discover potential candidate genes that may be undergoing adaptation to specific environments. The methodology in Poklukar et al., 2023 uses the same approach as Coop et al., 2010, which leads to the development of a Bayes Factor measuring the link between phenotypic and genotypic variations for each SNP. Statistical significance is then assessed SNP by SNP, correcting the multiple test problem with a False Discovery Rate (FDR) approach from Benjamini and Hochberg, 1995. They finally revealed 234 regions associated with traits of stature, fatness, growth, or reproduction.

Here, we propose using a local score approach to analyze the final dataset containing Bayes factors associated with stature traits, kindly provided by the authors of Poklukar et al., 2023. For each of the approximately 16 million SNPs covering 18 pig chromosomes, we have the SNP positions and associated Bayes Factor statistics. Table 4 shows the number of points for each chromosome. Note that a Bayes Factor is a real number, and p-values associated with the local score can not be directly assessed by the function `karlin`, `mcc` nor `daudin` as their associated methodologies require integer scores. One proposed solution is to discretize the scores. In the second part of this illustration, we also assess the effect of this discretization on the results, comparing three schemes: 1. real scores 2. scores multiplied by 10 and rounded to the closest unit 3. scores rounded to the closest unit. Due to the length of the sequence, we evaluate the p-values using the `karlinMonteCarlo` function, see Formula (5). See the R code in Robelin et al., 2025 (Appendix A4).

Data analysis. To analyze this big data file, we proceed chromosome by chromosome and use the `R` library `sqldf` to load the data. Refer to the `R` code in Robelin et al., 2025 (Appendix A4).

Without compromising generality, we present here the detailed analysis and results for chromosome 1. Note that the empirical expectation of the score is strictly negative and the scores are strictly positive, as expected for a meaningful local scoring analysis. Figure 5 shows the observed SNP score along chromosome 1 and the associated Lindley process.

The local score, its position, and all the suboptimal scores are calculated by the `localScoreC` function. The local score on chromosome 1 is 702.7715 and is realized by the segment situated in position (86184149, 86566846) with a p-value $< 10^{-16}$.

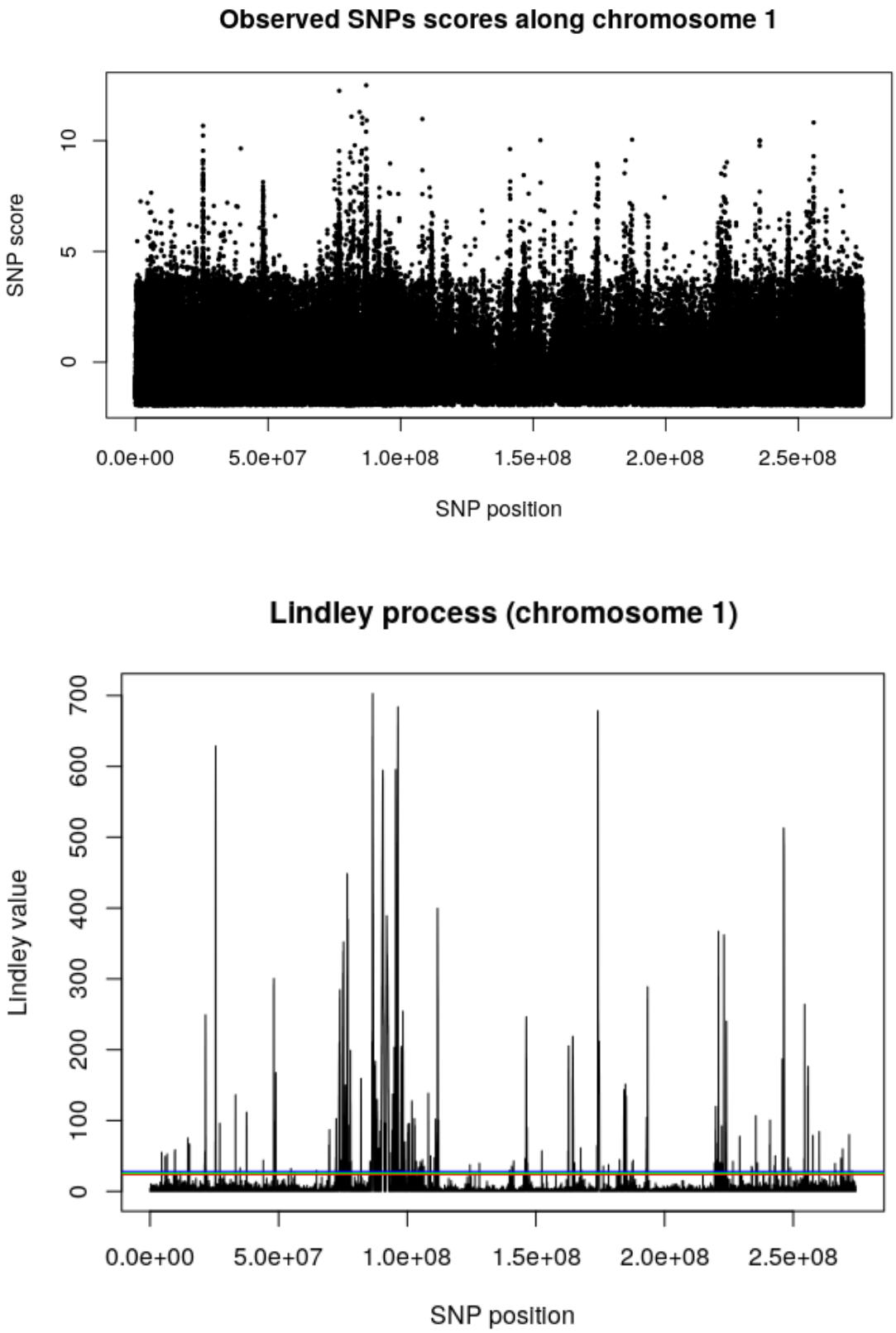

Figure 5 – Observed SNP scores along chromosome 1: Top) SNP Score values; Bottom) Associated Lindley process, with horizontal lines representing thresholds associated associated with local score statistical significance at the 5% (red), 1% (green) and 1‰(blue) levels.

Table 4 – Number of SNPs by pig chromosome in the dataset.

Chromosome	SNPs count
1	1427539
2	1072176
3	946332
4	923731
5	792366
6	1206701
7	899034
8	1110422
9	1020531
10	695091
11	664610
12	563400
13	1225446
14	1029162
15	899320
16	694637
17	570599
18	417965

In the same way, we assess the statistical significance of the scores of sub-optimal segments. As mentioned in Fariello et al., 2017, the local score threshold given for a first-order risk α also ensures a first-order risk α for at least one false positive among all excursions above this threshold. In other word, all excursions above this threshold can be considered as significant sub-optimal segments scores. On chromosome 1, we found a total of 67535 segments with positive cumulative scores, from which 225 segments appear to be significant at 5%-level, 210 segments at 1%-level, and 183 segments at 1‰-level. See the code in Robelin et al., 2025 (Appendix A4) to obtain these results.

Chromosome 1 contains a total of 1421525 SNPs with individual scores. The proportion of SNPs present in significant segments relative to this total is approximately 4% (see Table 5). In Poklukar et al., 2023, the Bayes factor-based whole-genome analysis retains 2 significant segments at threshold 5%, corrected for multiple tests, compared to 225 segments significantly detected by the local score approach for the stature trait.

Table 5 – Numbers and proportions of SNPs present in a significative segment according to the test threshold.

Test threshold	0.05	0.01	0.001
Number of SNPs present in significant segment	54129	52954	51357
Proportion of SNP present in significant segment	0.04	0.04	0.04

Score discretization assessment. Three scoring schemes are compared: a) real scores as given by the input Bayes factor b) decimal scores multiplied by 10 and then rounded c) scores rounded to the nearest whole number. b) and c) give integer scores. Figure 6 shows the empirical distributions of each scoring system obtained from chromosome 18. Other chromosomes show very similar distributions (not shown).

For each chromosome in the entire genome, Table 6 summarizes the number of significant segments detected applying a threshold of 5%, 1% and 1‰. These numbers are also indicated according to the scoring system.

Let us also examine the influence of the three scoring systems on the length of segments that achieve the local score: Figure 7 shows comparable boxplots of the (log)-length of the detected segments below the 5% threshold for the three scoring schemes.

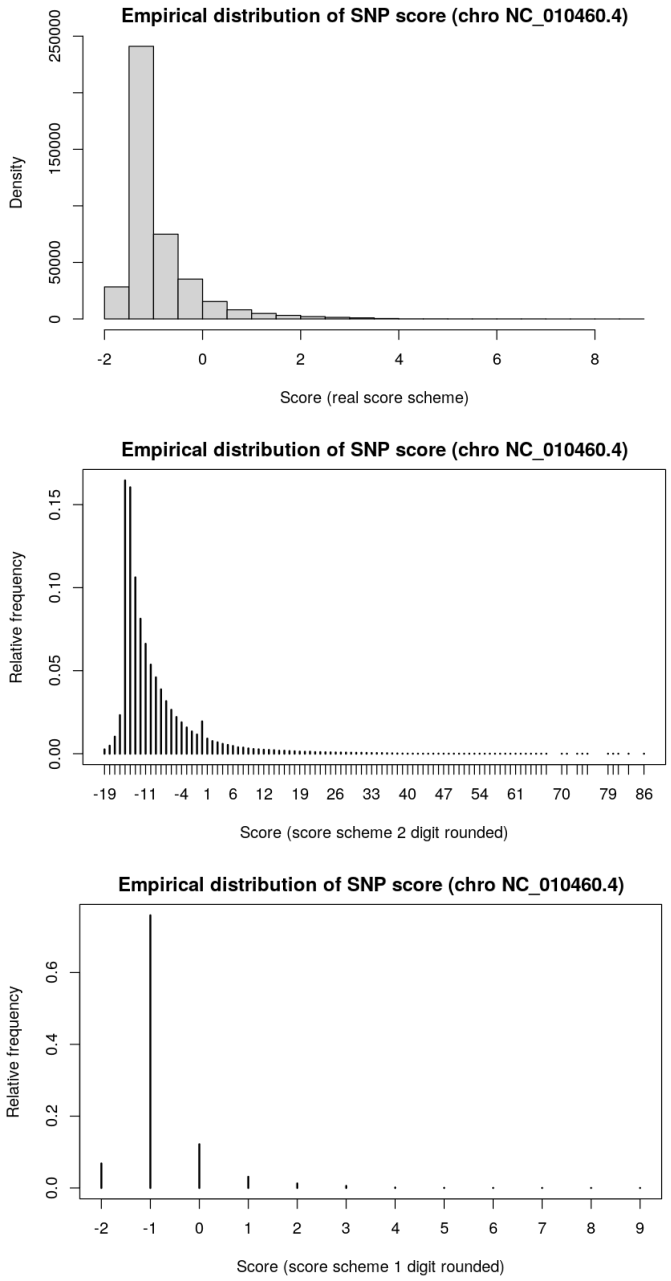

Figure 6 – Empirical distributions of SNP scores obtained on chromosome 18 for three scoring scheme: 1. Real score 2. Two-digits rounded score 3. One-digit rounded score.

Considering the segments obtained with the real scores as a reference, Table 7 displays the numbers of false positive and false negative segments that occur with 1-digit scores and 2-digit scores. Caution should be taken as a threshold is applied to p-value below 5%, which may change the list of significant close to the threshold. Given the wide range and distribution skewness of the real scores, the rounded 2-digit scores only slightly change the results, missing only 9 segments (0.5%) over 1882 real segments on the whole scale and falsely detecting 45 segments (2.3%) over 1919 detected segments. Note that the performance of the brutal unit rounded score essentially reflects the real segment detected with only 76 missing segments (4%) (false negative) over 1882 and falsely detecting (false positives) 157 segments (8%) over 1938 detected segments.

Table 6 – Numbers of significant segments detected applying a threshold of 5%, 1% and ‰. These numbers regarding the scoring scheme are also indicated.

chromosomes		Real scores			2-digits scores			Unit scores		
		5%	1%	5‰	5%	1%	5‰	5%	1%	5‰
1	NC_010443.5	226	212	185	225	213	185	218	201	179
2	NC_010444.4	70	60	49	70	59	48	67	55	44
3	NC_010445.4	56	54	46	56	53	45	56	53	45
4	NC_010446.5	90	78	70	87	78	67	89	80	71
5	NC_010447.5	94	85	67	94	85	68	105	92	72
6	NC_010448.4	93	83	74	102	91	78	109	96	88
7	NC_010449.5	94	83	71	96	85	72	96	84	72
8	NC_010450.4	131	118	108	131	118	107	127	121	112
9	NC_010451.4	132	120	105	134	123	107	145	136	120
10	NC_010452.4	108	97	86	111	98	86	123	104	91
11	NC_010453.5	67	58	50	66	57	50	60	51	44
12	NC_010454.4	52	46	40	55	47	41	56	46	41
13	NC_010455.5	161	153	130	162	156	135	160	150	133
14	NC_010456.5	135	122	105	143	126	112	140	127	107
15	NC_010457.5	142	124	110	157	138	120	152	133	113
16	NC_010458.4	74	69	64	75	71	65	81	76	69
17	NC_010459.5	70	62	57	68	61	57	68	61	57
18	NC_010460.4	87	71	61	87	73	61	86	73	63
Total		1882	1695	1478	1919	1732	1504	1938	1739	1521

Table 7 – Considering the real scoring scheme as the detection reference, the table shows the number of segments that differ from the reference for the 2-digit rounded scoring scheme, and the unit rounded scheme on the whole genome analysis. "False negative": number of segment that are present in the reference, but not in the considered scoring scheme; "False positive": number of segments significantly detected but not present in the reference.

score scheme	Real	2-digits	1-digit
Total detected segments (<5%)	1882	1919	1938
False negative		9 (0.5%)	76 (4%)
False positive		45 (2.3%)	157 (8%)

Computational details

The results in this paper were obtained using `R` 4.3.1. `R` itself and all packages used are available from the Comprehensive `R` Archive Network (CRAN) at `https://CRAN.R-project.org/`. Computation times are discussed in Robelin et al., 2025 (Appendix B).

Conclusion

When no a priori information is known about the length of the segments to be highlighted, the local score is a dedicated tool to exploit and complement sliding windows or scan statistics methods.

The `localScore` package allows to calculate statistical significance in different contexts and thus to distinguish segments with atypical optimal scores from those appearing randomly. The package groups together various functions allowing in particular to visualize and point out the highlighted regions. Different methods for evaluating statistical significance are proposed. A function allows to perform this calculation by automatically selecting the method most suited to the context, according to the length of the sequence, and the average score under a given or learned model. If the local score was initially defined for the identification of atypical regions

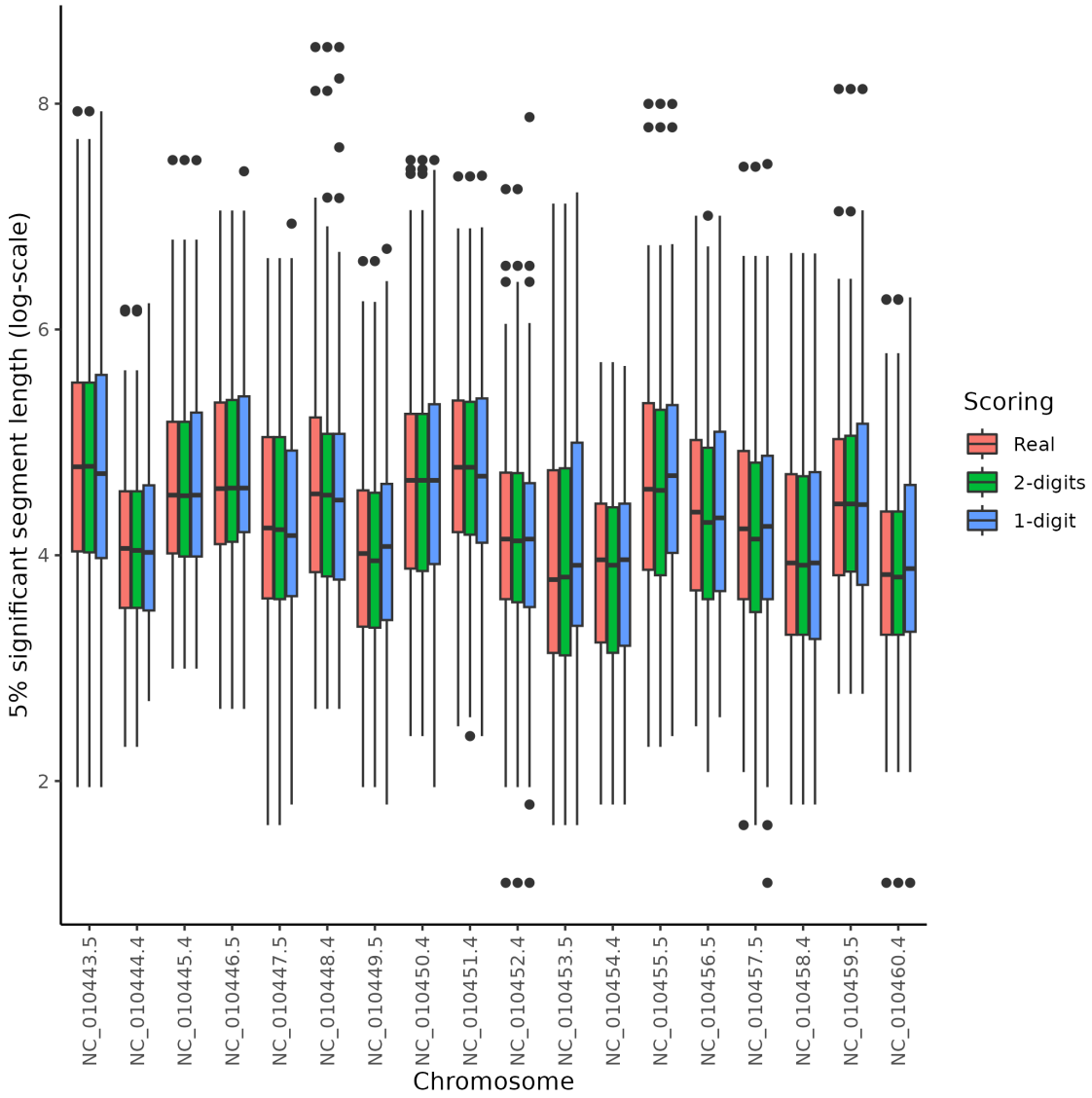

Figure 7 – Log-length of the detected segments below the threshold of 5% for the three scoring schemes by chromosomes.

within biological sequences, it can also be useful in many application areas, as we wanted to illustrate in our examples. It can also be applied to online analyses, in particular to the detection of breakpoints.

The disadvantage of the local score is that it is currently not generalizable to the continuous case or spatial (2-dimension) data, and may require a transformation of the data via a function called the score function, which must allow positive and negative values.

Acknowledgements

Preprint version 5 of this article has been peer-reviewed and recommended by Peer Community In Genomics (https://doi.org/10.24072/pci.genomics.100420; Wang, 2025).

The authors thank Sebastian Simon who started building the package during his internship.

Fundings

This work was supported by the project “Highlight” of the *Excellence Laboratory International Center of Mathematics and Computer Science in Toulouse* (Labex CIMI).

Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article.

Data, script, code, and supplementary information availability

The source code of the `localScore` package is available at: https://cran.r-project.org/web/packages/localScore/index.html. The package also contains the datasets analyzed in the article, except for the Genomic dataset available in Robelin et al., 2025. All the code used to analyze the data and produce the figures are available as supplementary material in Robelin et al., 2025.

References

Bairstow L (1920). *Applied Aerodynamic*. In: Appendix: The Solution of Algebraic Equations with Numerical Coefficients in the Case where Several Pairs of Complex Roots exist. London: Longmans, Green and Company. Chap. Appendix, pp. 551–560.

Benjamini Y, Hochberg Y (1995). *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing*. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Cellier D, Charlot F, Mercier S (2003). *An improved approximation for assessing the statistical significance of molecular sequence features*. *Jour. Appl. Prob.* **40**, 427–441. https://doi.org/10.1239/jap/1053003554.

Chabriac C, Lagnoux A, Mercier S, Vallois P (2014). *Elements related to the largest complete excursion of a reflected BM stopped at a fixed time. Application to local score*. *Stochastic Processes and their Applications* **124**, 4202–4223. https://doi.org/10.1016/j.spa.2014.07.003.

Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010). *Using environmental correlations to identify loci underlying local adaptation*. *Genetics* **185**, 1411–1423. https://doi.org/10.1534/genetics.110.114819.

Cucala L (2008). *A hypothesis-free multiple scan statistic with variable window*. *Biometrical Journal* **50**, 299–310. https://doi.org/10.1002/bimj.200710412.

Cucala L (2017). *Variable Window Scan Statistics: Alternatives to Generalized Likelihood Ratio Tests*. In: *Handbook of Scan Statistics*. Ed. by Joseph Glaz and Markos V. Koutras. New York, NY: Springer, pp. 1–16. https://doi.org/10.1007/978-1-4614-8414-1_36-1.

Fariello M, Boitard S, Mercier S, Robelin D, Faraut T, Arnould C, Le Bihan-Duval E, Recoquillay J, Salin G, Dahais G, Pitel F, Leterrier G, Sancristobal M (2017). *Accounting for Linkage Disequilibrium in genome scans for selection without individual genotypes : the local score approach*. *Molecular Ecology* **26(14)**, 3700–3714. https://doi.org/10.1111/mec.14141.

Glaz J, Pozdnyakov V, Wallenstein S (2009). *Scan statistics - Methods and applications*. Birkhäuser Boston. https://doi.org/10.1007/978-0-8176-4749-0.

Glaz J, Naus J, Wallenstein S (2001). *Scan Statistics*. Springer Series in Statistics. New York, NY: Springer. https://doi.org/10.1007/978-1-4757-3460-7.

Grusea S, Mercier S (2020). *Improvement on the distribution of maximal segmental score in a Markovian sequence*. *Journal of Applied Probability* **57.1**, 29–52. https://doi.org/10.1017/jpr.2019.75.

Hassenforder C, Mercier S (2007). *Exact Distribution of the Local Score for Markovian Sequences*. *AISM* **59**, 741–755. https://doi.org/10.1007/s10463-006-0064-6.

Karlin S, Altschul SF (1990). *Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes*. *PNAS* **87**, 2264–2268. https://doi.org/10.1073/pnas.87.6.2264.

Karlin S, Dembo A (1992). *Limit distributions of maximal segmental score among Markov-dependent partial sums*. *Advances in Applied Probability* **24**, 113–140. https://doi.org/10.2307/1427732.

Knox G (1959). *Secular pattern of congenital oesophageal atresia*. *British Journal of Preventive Social Medicine* **13**, 222–226. https://doi.org/10.1136/jech.13.4.222.

Kyte J, Doolittle R (1982). *A simple method for displaying the hydropathic character of a protein*. *Journal of molecular biology* **157**, 105–132. https://doi.org/10.1016/0022-2836(82)90515-0.

Lagnoux A, Mercier S, Vallois P (2017). *Statistical significance based on length and position of the local score in a model of i.i.d. sequences*. *Bioinformatics* **33**, 654–660. https://doi.org/10.1093/bioinformatics/btw699.

Mercier S (2020). *Transferring biological sequence analysis tools to break-point detection for on-line monitoring: A control chart based on the Local Score*. *Qual. Reliab. Engng. Int.* **36**, 2379–2397. https://doi.org/doi.org/10.1002/qre.2703.

Mercier S, Daudin J (2001). *Exact Distribution for the Local Score of One i.i.d. Random Sequence*. *Jour. Comp. Biol* **8**, 373–380. https://doi.org/10.1089/106652701752236197.

Nagarwalla N (1996). *A Scan Statistic with a Variable Window*. *Statistics in Medicine* **15**, 845–850. https://doi.org/10.1002/(SICI)1097-0258(19960415)15:7/9<845::AID-SIM254>3.0.CO;2-X.

Naus J, Wallenstein S (2006). *Temporal surveillance using scan statistics*. *Statistics in Medicine* **25**, 311–324. https://doi.org/10.1002/sim.2209.

Naus JI (1982). *Approximations for Distributions of Scan Statistics*. *Journal of the American Statistical Association* **77**, 177–183. https://doi.org/10.1080/01621459.1982.10477783.

Poklukar K, Mestre C, Škrlep M, Čandek-Potokar M, Ovilo C, Fontanesi L, Riquet J, Bovo S, Schiavo G, Ribani A, Muñoz M, Gallo M, Bozzi R, Charneca R, Quintanilla R, Kušec G, Mercat MJ, Zimmer C, Razmaite V, Araujo JP, et al. (2023). *A meta-analysis of genetic and phenotypic diversity of European local pig breeds reveals genomic regions associated with breed differentiation for production traits*. *Genetics Selection Evolution* **55**. https://doi.org/10.1186/s12711-023-00858-3.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. URL: https://www.R-project.org/.

Robelin D, Dejean S, Mercier S (2025). *Supplementary material: localScore: an R package to highlight optimal and suboptimal segments in a sequence with associated p-values computation*. https://doi.org/10.57745/TAQPAU.

Shewhart WA (1931). *Economic Control of Quality of Manufactured Product*. New York: D. Van Nostrand Company.

Simon S, Robelin D, Mercier S, Dejean S (2023). *localScore: Package for Sequence Analysis by Local Score*. R package version 1.0.11.

Wallenstein S, Neff N (1987). *An approximation for the distribution of the scan statistic*. *Statistics in Medicine* **6**, 197–207. https://doi.org/10.1002/sim.4780060212.

Wang S (2025). *localScore: finding optimal segments in genetic sequences*. *Peer Community in Genomics*. https://doi.org/10.24072/pci.genomics.100420.

Wang X, Glaz J (2014). *Variable Window Scan Statistics for Normal Data*. *Communications in Statistics - Theory and Methods* **43**, 2489–2504. https://doi.org/10.1080/03610926.2013.782201.