BIOINFORMATICS

Vol. 00 no. 00 2014
Pages 1–7

LandScape: A Simple Method to Aggregate P-Values and Other Stochastic Variables Without a Priori Grouping

Carsten Wiuf[1,*,†], Jonatan Schaumburg-Müller Pallesen[2,3,4,5,†], Leslie Foldager[4,5,6] and Jakob Grove[2,3,4,5]

[1] Department of Mathematical Science, University of Copenhagen, Denmark
[2] Department of Biomedicine, Aarhus University, Denmark
[3] iPSYCH, The Lundbeck Foundation Initiative for Integrative Psychiatric Research, Denmark
[4] iSEQ, Centre for Integrative Sequencing, Aarhus University, Denmark
[5] Bioinformatics Research Centre, Aarhus University, Denmark
[6] Department of Animal Science, Aarhus University, Denmark
[*] Corresponding author
[†] Joint first authors

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: In many areas of science it is custom to perform many, potentially millions, of tests simultaneously. To gain statistical power it is common to group tests based on a priori criteria such as predefined regions or by sliding windows. However, it is not straightforward to choose grouping criteria and the results might depend on the chosen criteria. Methods that summarise, or aggregate, test statistics or p-values, without relying on a priori criteria, are therefore desirable.

Results: We present a simple method to aggregate a sequence of stochastic variables, such as test statistics or p-values, into fewer variables without assuming a priori defined groups. We provide different ways to evaluate the significance of the aggregated variables based on theoretical considerations, using ideas from random walk theory, and resampling techniques. Validity of the method was demonstrated using simulations and real data analysis.

Discussion: Our method may be a useful supplement to standard procedures relying on evaluation of test statistics individually. Moreover, by being agnostic and not relying on predefined selected regions, it might be a practical alternative to conventionally used methods of aggregation of p-values over regions.

Availability and implementation: The method is implemented in Python and available online at http://www.jsmp.dk/landscape.

Contact: wiuf@math.ku.dk.

Supplementary information: Proofs of statements in the main text, additional remarks about the statistical procedures, further simulation results and information about the implemented method are available at Bioinformatics online.

1 INTRODUCTION

It is commonplace in many areas of science to perform thousands or millions of statistical tests simultaneously. To control the number of false discoveries, it is standard practice to adjust for multiple testing, which however might reduce the statistical power of each individual test severely. For sequentially ordered tests, such as they occur in genomics, association mapping, and time-series analysis, the tests might not only be of interest individually, but also on an aggregated level. Some power may therefore be gained by combining consecutive tests in a region.

To aggregate over a single region, many methods exist (Hendricks *et al.*, 2014). Many of these are inspired by an observation by Fisher (1932), that if k independent tests are performed with p-values $p_1, \ldots, p_k$, then $-2\sum_i \log p_i \sim \chi^2(2k)$. Importantly, Fisher's aggregated statistic can be statistically significant while none of the p-values individually are. Thus, it is possible to detect a combined effect that does not show in the individual tests. For dependent tests, the gold standard is to obtain the distribution of Fisher's test statistic by resampling.

In order to scan through the complete sequence of tests using region based methods, either the tests might be grouped according to some priorly defined regions or some sort of sliding window scheme could be employed (Gladwin, 2012). The latter requires choosing window size and overlap between subsequent windows. These choices may greatly influence the power to detect effects in a particular region. An interesting approach is based on clustering of points or consecutive successes (significant tests) in sliding windows, called scan statistics (Naus, 1982). As an alternative to aggregation, the correlation structure in the data has been used to estimate an "effective number of independent tests", which might then be used as an adjustment factor in multiple testing (Cheverud, 2001; Nyholt, 2004). A recent approach takes a different avenue and perform hierarchically based tests, whereby sequentially ordered hypotheses are identified that cannot all be true at the same time (Meijer *et al.*, 2015). The hierarchically structured procedure increases the power substantially (Meijer *et al.*, 2015).

We propose an agnostic method to summarise sequentially ordered stochastic variables, such as test statistics, without a priori grouping. The procedure is inspired by Random Walk theory (Karlin and Altschul, 1990; Karlin and Dembo, 1992; Iglehart, 1972). It travels along the sequence searching for a stretch of consecutive

values that jointly have a high "score". In this way, we end up with a random number of tests, aggregated over random (and generally unequal) numbers of variables. If the variables are independent then theory predicts the approximate distribution of the aggregated value. When this assumption is not fulfilled, resampling techniques are used to obtain the distribution.

2 AGGREGATION OF VARIABLES

2.1 A motivating example

Consider a sequence of random variables Z_k, where k denotes the position. We think of Z_k as (a transformation of) a test statistic or a p-value. The variable Z_k may be quite general, but suppose for now that $Z_k = 1$ if the k-th test is significant at level γ and otherwise let $Z_k = -1$. A long sequence of mainly positive Z_ks may be considered unlikely and indicative of deviation from the null hypothesis. We provide a definition of such sequences and an algorithm to find them. Specifically, we identify intervals $[n, m]$ (called maximal segments), such that the score $U_{nm} = \sum_{i=n}^{m} Z_i$ is positive, no smaller interval of $[n, m]$ has the same or a larger score, and such that $[n, m]$ is as long as possible. See Fig. 1A for an illustration.

In Section 2.2 we describe mathematically how to construct the maximal segments. They fall in two classes, *dependent* and *independent* segments, which we characterise in Section 2.3. The score of a dependent segment is dominated by the score of a corresponding independent segment.

2.2 Segments and scores

Let $\mathbb{K}$ be a finite or infinite set of consecutive positive integers starting at 1. We call such a set an *index set*. The length of $\mathbb{K}$, that is, the number of elements in $\mathbb{K}$, is denoted $|\mathbb{K}|$ and it may be finite or infinite. Let $Z = \{Z_k\}_{k\in\mathbb{K}}$, be a sequence of random variables and let U_{nm} be the partial sums:

$$U_{nm} = \sum_{k=n}^{m} Z_k, \quad n \le m \in \mathbb{K}. \tag{2.1}$$

If $m < n$, we take the partial sum to be zero.

DEFINITION 2.2. *A* segment *for Z is a closed interval $[n, m]$ such that $U_{nk} > 0$ and $U_{km} > 0$ for all $k \in [n, m]$, where we allow m to be infinite. A* maximal segment *for Z is a segment $[n, m]$ such that there are no other segments for Z containing it. The* score *of a maximal segment $[n, m]$ is the partial sum U_{nm}.*

A maximal segment cannot contain an interval $[k, l]$ such that $U_{kl} \ge U_{nm}$. Indeed if this was the case then $U_{nm} = U_{n(k-1)} + U_{kl} + U_{(l+1)m} \le U_{kl}$ and at least one of $U_{n(k-1)}$ and $U_{(l+1)m}$ must be non-positive, contradicting the definition of a maximal segment. Similarly, two different maximal segments, $[n_1, m_1]$ and $[n_2, m_2]$ are always disjoint. Assume conversely that $n_1 \le n_2 \le m_1 \le m_2$. Then for $n_1 \le k < n_2$, we have $U_{n_1 k} > 0$, as $[n_1, m_1]$ is a segment, and $U_{km_2} = U_{km_1} + U_{m_1+1,m_2} > 0$, since both $[n_1, m_1]$ and $[n_2, m_2]$ are segments, and similarly for $n_2 \le k \le m_1$ and $m_1 < k \le m_2$. Thus $U_{n_1 k} > 0$ and $U_{km_2} > 0$ for all $k \in [n_1, m_2]$, which contradict the maximality of $[n_1, m_1]$ and $[n_2, m_2]$.

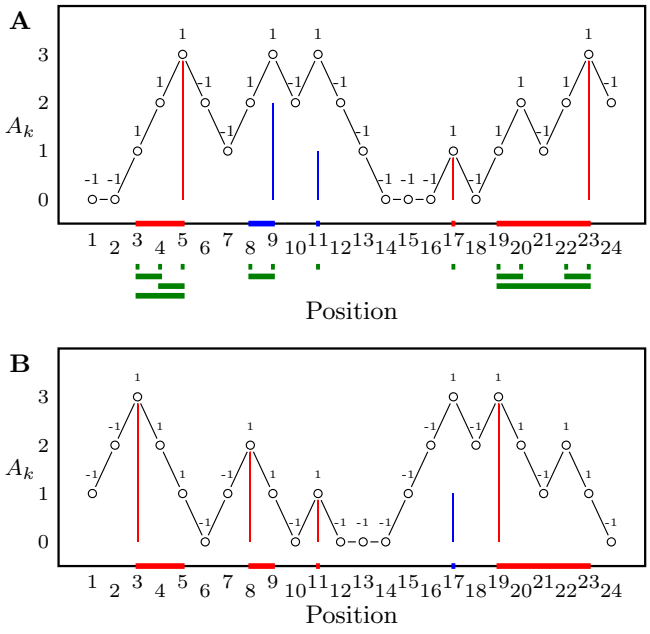

Figure 1. **(A)** Illustrated is a sequence Z_k with values in $\{-1, 1\}$ as in the motivating example (Section 2.1). The value of Z_k is shown above the circles. The landscape is the accumulated sequence $A_k = \max\{0, Z_k + A_{k-1}\}$ (2.4). Green bars show all segments, that is, intervals $[n, m]$ such that $U_{nk} > 0, U_{km} > 0$ (Definition 2.2), for example, $[3, 4]$ is a segment. Maximal segments are coloured red or blue; red are independent segments and blue are dependent segments (Definition 2.8). A dependent segment begins when A_k starts increasing after having decreased to a non-zero value. The vertical red and blue bars indicate the score of the segment (Definition 2.2). **(B)** LandScape picture of the reverse sequence relative to the sequence in (A) (see section 2.4). For the reverse sequence, all maximal segments are independent, except for that containing only position 17 (blue). Non-maximal segments are not shown. Position 1 corresponds to position 24 in (A) and so forth.

It follows that each position, $k \in \mathbb{K}$, is in at most one maximal segment. Hence, we have the following:

THEOREM 2.3. *Let $Z = \{Z_k\}_{k\in\mathbb{K}}$, be a sequence of random variables. Then there is a unique sequence of disjoint maximal segments $J_i = [n_i, m_i]$, $i \in \mathbb{M}$, where $\mathbb{M}$ is some index set, and such that the sequence contains all maximal segments for Z. That is, if I is a maximal segment for Z, then $I = J_i$ for some $i \in \mathbb{M}$.*

The sequence of segments $J_i, i \in \mathbb{M}$, in Theorem 2.3 is said to be *maximal*. If $[n, m]$ is a maximal segment then it follows from Definition 2.2 that $Z_n, Z_m > 0$ and $Z_{n-1}, Z_{m+1} \le 0$. If $Z_k \ge 0$ for all $k \in \mathbb{K}$, then there is at most one segment which also is maximal. If $Z_k > 0$ for at least one k then there is at least one segment, otherwise there are none. Thus, a natural requirement is that Z_k can take positive as well as negative values.

In the following, if $J_i = [n_i, m_i]$, $i \in \mathbb{M}$, is a sequence of maximal segments, we assume that n_i is increasing in i. This can always be achieved, potentially by reordering the segments. It is also to be understood that whenever we talk of segments, it is defined with respect to a given sequence of random variables. The statistics of interest are the scores of the maximal segments.

2.3 Independent and dependent segments

In this section we present an algorithm to find all maximal segments. In the process the concept of independent and dependent segments will be introduced.

Formally, we let $A_0 = 0$ and define the accumulated sums by

$$A_k = \max\{0, Z_k + A_{k-1}\}, \quad k \in \mathbb{K}. \tag{2.4}$$

If $\mathbb{K}$ is finite, we put $A_{|\mathbb{K}|+1} = 0$. Define the start points (s_{i0}) and termination points (t_{i0}) by

$$\begin{aligned} s_{i0} &= \min\{k \mid k > t_{i-1,0}, A_k > 0\}, \\ t_{i0} &= \min\{k \mid k \geq s_{i0}, A_{k+1} = 0\}, \end{aligned} \tag{2.5}$$

with $t_{00} = 0$. Let $\mathbb{I}$ be the set of indices i for which s_{i0}, hence also t_{i0}, is defined. The interval $S_i = [s_{i0}, t_{i0}]$, $i \in \mathbb{I}$, is called the i-th *section*. Only the last section can be infinite. By definition s_{i0} is the first time Z_k is positive after $t_{i-1,0}$, $A_k > 0$ for all $k \in S_i$ and $A_k = 0$ between sections. If $Z_k > 0$ for at least one k, then also $A_k > 0$ for at least one k, and there is at least one section. Otherwise there are none.

We further define the following:

$$\begin{aligned} Y_{i0} &= \max\{A_k \mid k \in S_i\}, \\ e_{i0} &= \min\{k \in S_i \mid A_k = Y_{i0}\}. \end{aligned} \tag{2.6}$$

The variable Y_{i0} is the maximum of the accumulated sums obtained in section S_i and e_{i0} is the index for which it is obtained for the first time. Here and elsewhere we allow the 'maximum' to be infinite. Recursively, define for $j > 0$,

$$\begin{aligned} s_{ij} &= \min\{k \in S_i \mid k > e_{i,j-1}, A_k > A_{k-1}\}, \\ t_{ij} &= \min\{k \in S_i \mid k \geq s_{ij}, A_{s_{ij}-1} \geq A_{k+1}\}, \\ Y_{ij} &= \max\{A_k - A_{s_{ij}-1} \mid k \in [s_{ij}, t_{ij}]\}, \\ e_{ij} &= \min\{k \in [s_{ij}, t_{ij}] \mid A_k - A_{s_{ij}-1} = Y_{ij}\}. \end{aligned} \tag{2.7}$$

For given i, the recursion stops the first time s_{ij} is not defined.

The intervals $[s_{ij}, e_{ij}]$ are by definition non-overlapping and between any such two intervals there is at least one point, that is, $s_{ij} > e_{i,j-1} + 1$, hence they cannot be adjacent. The main difference between (2.5) and (2.7) is that s_{ij} is the first time $Z_k = A_k - A_{k-1}$ is positive after $e_{i,j-1}$, whereas s_{i0} is the first time this happens after $t_{i-1,0}$.

It follows that $[s_{ij}, e_{ij}]$ is a segment (Definition 2.2): From (2.7) we have that the partial sums $U_{s_{ij},k}$ are positive. If $U_{k,e_{ij}}$ was non-positive for some k, the accumulated sum Y_{ij} would not be maximal, as required by (2.7). Hence, $[s_{ij}, e_{ij}]$ is a segment.

DEFINITION 2.8. *The first segment in section S_i, $i \in \mathbb{I}$, is called the* independent segment *of the section and denoted by J_{i0}. The remaining segments in S_i are numbered consecutively J_{ij}, $j \in \mathbb{D}_i$ (for some index set) and termed the* dependent segments *of section S_i.*

The independent and dependent segments $[s_{ij}, e_{ij}]$, indicated with red and blue bars, respectively, on the x-axis of Fig. 1A, comprise all the maximal segments in the motivating example. This holds in general:

PROPOSITION 2.9. *The maximal segments are precisely the segments J_{ij}, $i \in \mathbb{I}$, $j \in \{0\} \cup \mathbb{D}_i$ with score $U_{s_{ij},e_{ij}} = Y_{ij}$. The score of a dependent segment J_{ij} is dominated by the score of the independent segment J_{i0} in that $Y_{ij} \leq Y_{i0}$, $i \in \mathbb{I}$, $j \in \mathbb{D}_i$.*

We have implemented the recursions given in (2.5)-(2.7) in a program called LandScape (see Supplementary information). Overall this algorithm runs in time $O(|\mathbb{K}|^2)$, but in practice it is closer to $O(|\mathbb{K}|)$. The sections can be found in time $O(|\mathbb{K}|)$.

2.4 Reversing and extending the sequence

If $|\mathbb{K}| < \infty$, then the reversed sequence of random variables, $Z_k^r = Z_{|\mathbb{K}|-k+1}$, $k \in \mathbb{K}$, starting from the right running towards the left is well defined. By definition the maximal segments of the reversed sequence are the same as those of the forward sequence. The same is *not* true for the independent and dependent segments as they depend on the direction of the sequence, compare Fig. 1A and B.

If the sequence Z_k, $k \in \mathbb{K}$, is extended to the right then the maximal segments do not change, except possibly for section $S_{|\mathbb{I}|}$. If $A_{|\mathbb{K}|} \neq 0$, then the last section stops before the accumulated sum reaches zero. Adding more variables might therefore change the maximal segments of the last section. If $A_{|\mathbb{K}|} = 0$, then the original maximal segments remains when adding more variables and new maximal segments are appended. Similarly, if the sequence is extended to the left, the last section of the reversed sequence determines the maximal segments that might change. The start and end points of the sections in the forward and the reversed sequence are not identical. Hence, even though the maximal segments are unchanged we cannot identify the sections of the reversed sequence from the sections of the forward sequence (Fig. 1A and B).

3 EVALUATION OF SCORES

We are interested in the distribution of the score Y_{ij} of a typical maximal segment. Even if the Z_ks are independent random variables, the scores will in general not be independent. Only a few theoretical results are known about the distribution of Y_{ij}. These are primarily based on random walk theory. In typically applications, however, the assumptions necessary to apply random walk theory are not fulfilled and we have to resort to other methods. We propose two simulation-based strategies.

We start with two examples.

EXAMPLE 3.1. Assume as in the motivating example (Section 2.1) that a test is performed for each $k \in \mathbb{K}$ and let $\gamma \in (0, 1)$ be a threshold. Let $Z_k = 1$ if the k-th test is below the threshold and otherwise let $Z_k = -1$. That is, under the null hypothesis, $Z_k = 1$ with probability γ and $Z_k = -1$ with probability $1 - \gamma$. The expectation $E(Z_k) = 2\gamma - 1$ is negative for $\gamma < \frac{1}{2}$. We consider γ an input parameter to LandScape.

EXAMPLE 3.2. Let X_k be a positive variable, for example a p-value. Define $Z_k = \log(z_\gamma / X_k)$ for some $z_\gamma > 0$. If X_k is a uniform p-value, z_γ could be a common significance threshold for the tests, $z_\gamma = \gamma$. In that case $-\log(X_K)$ is an exponential variable with intensity 1 and $E(Z_k) = \log(\gamma) + 1$, which is negative for $\gamma < \exp(-1) \approx 0.367$. If X_k is a test statistic, z_γ could be the γ-quantile of X_k. For example, if X_k is $\chi^2(1)$-distributed and $\gamma = 0.05$, then the γ-quantile is $z_\gamma = 3.84$ and $Z_k = \log(3.84/X_k)$.

If $X_k = \gamma e^{-1}$ with probability γ and $X_k = \gamma e$ otherwise, then we retrieve the situation in Example 3.1.

3.1 Independent variables

Assume Z_k fulfils the condition

$$E(Z_k) < 0, \quad \text{and} \quad P(Z_k > 0) > 0. \tag{3.3}$$

That is, Z_k tends to be negative but it can take positive values. It follows from the law of large numbers, that if the Z_ks are independent with common distribution, U_{nm} will eventually hit zero as m becomes large. Examples 3.1 and 3.2 fulfil that $P(Z_k > 0)$ is positive. Further, $E(Z_k) < 0$ might or might not be fulfilled depending on the choice of threshold γ.

Let $M_0 = |\mathbb{I}|$ be the number of independent segments and $M_i = |\mathbb{D}_i|$ the number of dependent segments of section S_i.

THEOREM 3.4. *Assume that Z_k, $k \in \mathbb{K}$, are independent random variables with common distribution, fulfilling condition (3.3). Assume $|\mathbb{K}| = \infty$, then the following holds:*

1. *$M_0 = \infty$ and $\sum_{i=1}^{M_0} M_i = \infty$, but M_i, $i > 0$, is finite with probability one.*
2. *The distribution of Y_{ij}, $j \geq 0$, does not depend on i.*
3. *Let $\widetilde{\mathbb{I}} \subseteq \mathbb{I}$ be a finite index set. The distribution of the scores factorizes over sections as*

$$P(Y_{ij} \leq x_{ij}, j \in \widetilde{\mathbb{D}}_i, i \in \widetilde{\mathbb{I}}) = \prod_{i=1}^{|\widetilde{\mathbb{I}}|} P(Y_{ij} \leq x_{ij}, j \in \widetilde{\mathbb{D}}_i),$$

where $\widetilde{\mathbb{D}}_i = \{0\} \cup \mathbb{D}_i$ and x_{ij} is a real number. In particular, the scores of the independent segments are independent random variables.

For any value of $|\mathbb{K}|$, we have

4. *$P(Y_{ij} \leq Y_{i0}, j \in \mathbb{D}_i) = 1$ for any $i \in \mathbb{I}$.*

The last property is general and does not require any of the assumptions of the theorem. It follows from Definition 2.8 as noted in Proposition 2.9.

To state the distributional results we need one further assumption. Assume, in addition to the assumptions of Theorem 3.4, that there exists a number $\lambda > 0$, such that

$$E(e^{\lambda Z_k}) = 1 \quad \text{and} \quad E(Z_k e^{\lambda Z_k}) < \infty. \tag{3.5}$$

Hereafter let λ be such a number. If Z_k only takes a *finite* number of values then the existence of a λ fulfilling condition (3.5) follows from condition (3.3) (see the Supplementary information).

A random variable X is said to be a lattice variable if there is $\delta > 0$, such that X takes values in $\mathbb{Z}_\delta = \{\delta j | j \in \mathbb{Z}\}$, where $\mathbb{Z}$ is the set of integers. Given a lattice variable with values in $\mathbb{Z}_\delta$, δ is assumed to be chosen as large as possible. If X is not a lattice variable, then X is said to be a non-lattice variable. Karlin and Dembo (1992) (see also Iglehart (1972)) prove that if Z_k is a lattice variable and $|\mathbb{K}| \gg i$, then for large integers y,

$$P(Y_{i0} \geq \delta y) \approx C e^{-\delta \lambda y} \tag{3.6}$$

for some real valued constant C. Thus, the tail distribution of Y_{i0} is approximately geometric. By applying the result to the reversed sequence we obtain an approximate distribution of the score of other independent segments. This applies in particular to the last segment, $i = M_0$, which is the first independent segment in the reversed sequence.

If Z_k is a non-lattice variable a similar result holds, however, the tail distribution of Y_{i0} is now approximately exponential (Karlin and Dembo, 1992),

$$P(Y_{i0} \geq y) \approx C e^{-\lambda y}, \tag{3.7}$$

where y is a large real number and C a constant. The constant is characterised by the distribution of the partial sums, U_{nm}, in the lattice as well as the non-lattice case. In general, it cannot be worked out explicitly, but must be found by approximation or simulation.

Assigning p-values based on the theoretical approximations above is termed A0. To correct for multiple testing we follow the procedure outlined in Section 3.2.1.

EXAMPLE 3.8 (Example 3.1, continued). Here Z_k is a lattice variable with $\delta = 1$ and $\lambda = \log(1-\gamma) - \log(\gamma)$, $\gamma < 0.5$. It follows that the asymptotic tail distribution is $P(Y_{i0} \geq y) \approx C e^{-\lambda y}$ for large integers y. In this case, $C \approx 1 - e^{-\lambda}$ (Karlin and Dembo, 1992), hence Y_{i0} is approximately geometric Geo(p) with $p = 1 - e^{-\lambda}$.

EXAMPLE 3.9 (Example 3.2, continued). Here $Z_k = \log(z_\gamma / X_k)$ and λ fulfils $1 = E(e^{\lambda Z_k}) = z_\gamma^\lambda E(e^{-\lambda \log(X_k)})$. Let X_k be a uniform p-value and $z_\gamma = \gamma \in (0,1)$. It follows that

$$\gamma^\lambda E(e^{-\lambda \log(X_k)}) = \gamma^\lambda \int_0^\infty e^{(\lambda-1)x} dx = \frac{\gamma^\lambda}{1-\lambda} = 1.$$

Since $0 < \gamma < 1$, there is a unique $0 < \lambda < 1$ fulfilling

$$\log(\gamma) = \frac{\log(1-\lambda)}{\lambda} \quad \text{such that} \quad E(e^{\lambda Z_k}) = 1.$$

The tail distribution of Y_{i0} is approximately an exponential distribution $\lambda e^{-\lambda y}$ and $C = \lambda$.

3.2 Dependent variables

The assumption that Z_k, $k \in \mathbb{K}$, are independent random variables is very restrictive. The results in the previous section can be shown to hold also if Z_k is controlled by a Hidden Markov Model (Karlin and Dembo, 1992), which broadens the scope of applications. However, this might still be too restrictive, for example in association mapping, where the variables Z_k rarely are equally spaced along chromosomes and there might be higher order dependencies among the variables. In addition, only independent segments can be assigned a p-value. Here we present two resampling-based approaches to remedy the theoretical shortcomings.

3.2.1 Approach A1. The score of each maximal segment is evaluated against the distribution of the score of a randomly chosen maximal segment. Thus, we disregard the positional information encoded in the indices of the score Y_{ij}. Clearly, for this approach to make sense, we must require some homogeneity in distribution across the sequence, for example, that $(Z_1, \ldots, Z_{|\mathbb{K}|})$ forms a stationary sequence, that is, the distribution of a subsequence $(Z_k, Z_{k+1}, \ldots, Z_{k+j})$ does not depend on k. In particular, the correlation between Z_k and Z_{k+j} only depends on j.

To evaluate the significance of a score, apply a resampling procedure to obtain B resampled data sets, each containing at least one segment (by rejecting samples with no segments): $(Z^b_1, \ldots, Z^b_{|\mathbb{K}|})$, $b = 1, \ldots, B$. Potential procedures include bootstrap and permutation methods, tailored to the concrete data set. For each resampling we find the maximal segments.

Let Y be the score of a randomly chosen maximal segment. The score Y is equal to Y_{ij} with probability $1/M$, if there are $M = M_0 + \sum_{i'=1}^{M_0} M_{i'}$ maximal segments and $1 \le i \le M_0, j \le M_i$. Hence, the distribution of Y might be approximated by,

$$P(Y \ge y) \approx \frac{1}{B+1} \sum_{b=0}^{B} \sum_{i=1}^{M_0^b} \sum_{j=1}^{M_i^b} \frac{1(Y_{ij}^b \ge y)}{M^b}, \qquad (3.10)$$

where $M^b = M_0^b + \sum_{i=1}^{M_0^b} M_i^b$ is the number of maximal segments in the b-th resampled data set with $b = 0$ denoting the original (not resampled) sample, and $1(\cdot)$ is the indicator function taking the value one if the condition in the parenthesis is fulfilled and zero otherwise. The precision of the approximation depends on B to the order $1/\sqrt{B}$, which is a consequence of the central limit theorem.

To control the total number of type I errors at level α, we test the M scores at threshold $\alpha/E(M)$. Under mild assumptions, this correction controls the expected number of type I errors as well as the family-wise error at level α in the weak sense. See proof of this in the Supplementary information. The expectation $E(M)$ can be approximated from the resampled data, $E(M) \approx \frac{1}{B+1} \sum_{b=0}^{B} M^b$.

In Example 3.1 and 3.2, $E(M) \le \gamma|\mathbb{K}|$. The bound would be achieved if all significant tests gave rise to a maximal segment. Hence we expect the correction factor to be at least γ times smaller than the standard Bonferroni correction, $|\mathbb{K}|$, but it is likely to be much smaller in practical applications.

The same correction can be applied to the theoretical approach discussed in Section 3.1 with $E(M)$ replaced by $E(M_0)$. If the constant C is known, as in the two examples, then the approximate distributions (3.6)-(3.7) are fully characterised. In that case the observed M_0 might be used as a crude estimate of $E(M_0)$ and the need for simulation becomes redundant.

3.2.2 Approach A2. Denote by $Y(J)$ the score assigned to each maximal segment J. We suggest the following test probability to evaluate the score of a maximal segment:

$$P_J(y) = P\big(\exists J' \text{ max'l seg} : J' \cap J \ne \emptyset \text{ and } Y(J') \ge y \,\big|\, \exists J' \text{ max'l seg} : J' \cap J \ne \emptyset\big) \qquad (3.11)$$

where $J' \cap J$ is the intersection of the stochastic maximal segment J' and the observed maximal segment J. The probability measures how often there exists a maximal segment overlapping the original segment that has a score larger than y. A similar alternative, called A3, is discussed in the supplement.

To estimate the probability $P_J(y)$, we apply a resampling procedure to obtain B resampled data sets, $(Z^b_1, \ldots, Z^b_{|\mathbb{K}|})$, $b = 1, \ldots, B$, and calculate the resampled scores $Y^b(k)$ for each position in each resampled data set. The probability $P_J(y)$ can be approximated from the resampled maximal segments *that overlap the original segment*:

$$P_J(y) \approx \frac{1}{\hat{B}_J + 1} \sum_{b=0}^{B} 1\big(\exists J^b \colon J^b \cap J \ne \emptyset \text{ and } Y(J^b) \ge y\big), \qquad (3.12)$$

where $b = 0$ denotes the original (not resampled) sample, $\hat{B}_J$ is the (stochastic) number of times there is a resampled maximal segment overlapping the original segment, and J^b is any maximal segment in resampling b that overlaps the observed segment J. This makes the approach computationally demanding because $\hat{B}_J$ might be small even when B is large.

The precision of the approximation depends on B as well as the number of maximal segments obtained in each resampled data set. In general (3.12) requires a higher B than (3.10) to obtain the same precision as we cannot use all resampled segments to assess the significance of an individual segment, but only those resampled segments that overlap with it. The right side of (3.12) converges as $1/\sqrt{\hat{B}_J}$, which follows from the central limit theorem.

If the (resampled) value of $P_J(y)$ is less than $\alpha/E(M)$, then the score is declared significant at level α, potentially using the resampled expectation found under A1 instead of $E(M)$. In the supplementary information we argue that this and the alternative procedure A3, control the expected number of type I errors in the weak sense, similarly to A1. The simulation results in next section indicate this as well.

4 APPLICATIONS

4.1 Simulated data example

We use an auto-regressive process $AR(1)$ as model for a stationary sequence of variables,

$$X_1 \sim N(0, \sigma^2), \quad \text{and} \quad X_k = \phi X_{k-1} + e_k, \quad k \ge 2,$$

where $\sigma^2 = \frac{1}{1-\phi^2}$, $0 \le \phi < 1$ and $e_k \sim N(0,1)$. It follows that

$$X_k \sim N(0, \sigma^2), \quad \text{and} \quad \text{cor}(X_k, X_{k+l}) = \phi^l.$$

For a simulated sequence, x_k, $k \in \mathbb{K}$, we define corresponding p-values $p_k = P(X_k > x_k)$. To simulate a region with generally reduced p-values, we choose a window W of size w randomly and multiply the p-values in the window by a factor f, $0 < f < 1$, such that if $f \approx 1$ the effect is small and if $f \approx 0$ the effect is large.

We follow Example 3.2 with the parameter $\gamma = 0.05, 0.1, 0.2$ and 0.3, and apply the three assessment strategies A0, A1 and A2, as well as Bonferroni correction $\alpha/|\mathbb{K}|$ to a series of simulated data sets with different w and f. We address family-wise type I error rate (FWER) under the null hypothesis ($f = 1$), and the power and FWER under various alternative hypotheses ($f < 1$) using the global significance level $\alpha = 0.05$. For each scenario, 1000 sequences $\{Z_k\}_{k\in\mathbb{K}}$ were generated by simulation. The power is estimated as the proportion of the 1000 simulations where a maximal segment overlapping W is significant, while the FWER is measured as the proportion of simulations where a maximal segment not overlapping W is significant. In Fig. 2 results are shown for $|\mathbb{K}| = 300$, $\phi = 0.4$, and $w = 20$, but results for $|\mathbb{K}| = 300$ and 600, $\phi = 0$, 0.2 and 0.4, and $w = 10$, 20 and 40 (also for A3) can be found in the supplement, where the simulation procedure is explained in greater detail as well.

Overall, the FWER is well maintained for A1 and A2 in all 72 scenarios examined (see the Supplementary information and Fig. 2). As a matter of fact it appears FWER is well below the desired 0.05 indicating that A1 and A2 may be conservative. What is more,

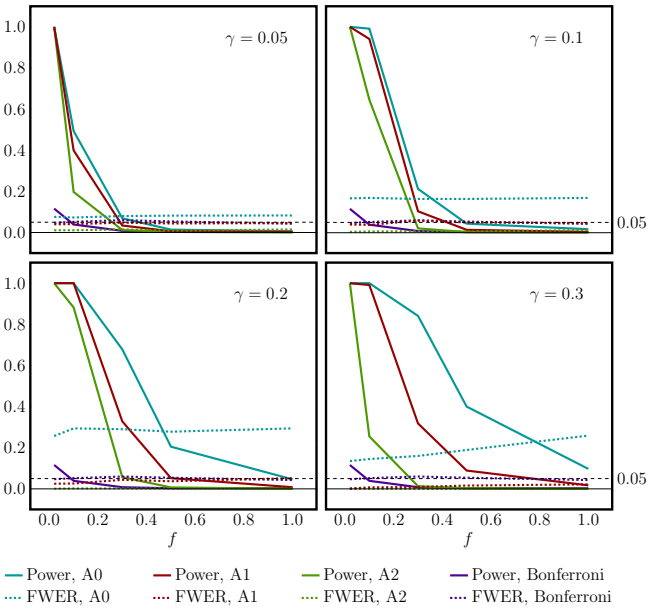

Figure 2. Power and FWER at four different thresholds, γ. Data was generated with dependency parameter $\phi = 0.4$, window size $w = 20$, and size of index set $|\mathbb{K}| = 300$.

it indicates that the approaches provide strong control of FWER. The simulations also show that as expected, A0 is highly anti-conservative in the presence of dependency, $\phi > 0$, and thus, the apparently higher power is misleading. Also as expected, the Bonferroni correction has much less power than A1 and A2.

4.2 Real data: example 1

In a large multi-stage schizophrenia genome-wide association study 108 genome-wide significant loci were identified (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). We reanalysed the 108 regions using LandScape on SNP data from the Molecular Genetics of Schizophrenia study (MGS) (Sanders *et al.*, 2010). The MGS data consists of 2 679 cases and 2 484 controls. After applying standard quality control, we extracted all SNPs in the 108 regions from the MGS data. Subsequently excluding 7 regions with no SNPs and the remaining 101 regions (range: 1-175 SNPs, mean 28.3 SNPs) were tested for association using logistic regression. LandScape was then applied to $Z_k = \log(\gamma/X_k)$ for $\gamma = 0.05, 0.1, 0.2, 0.3$, X_k being the p-values for SNP-wise association. A2 was employed with a total of 10 000 permutations of case-control status and a significance level of $\alpha = 0.05$.

We used a wide range of γ-values to illustrate the method. Further, a high value of γ allows the signal from each individual SNP to be weak which might well be the case in schizophrenia. The results, together with the results from a single-marker Bonferroni corrected analysis, are summarised in Fig. 3. Overall the results show that as γ increases from 0.05 to 0.30, more significant maximal segments are discovered (14, 21, 26, 29, resp.), compared to 19 identified using Bonferroni corrected p-values.

4.3 Real data: example 2

As a second real data example we examined the DNA copy number data from Carvalho *et al.* (2009). It consists of array CGH (comparative genomic hybridisation) data from 68 colon tumours, 33 of these are classified as adenomas and 35 as carcinomas. In total we have copy number intensities (X_k) for 4 046 markers in the genome. We applied LandScape using $Z_k = \log(\gamma/X_k)$ for $\gamma = 0.01, 0.05$, and A0, A1, A2 with a total of 10 000 permutations (adenoma/carcinoma status is permuted). A standard t-test X_k was used to test for differences in copy number intensity between the two groups. A significance level of $\alpha = 0.05$ was applied.

We only seek to identify the most important abnormalities. Hence we put γ lower than in the previous example. Differences in copy numbers are expected priorly and often cover an entire chromosome, see e.g. Jasmine *et al.* (2012). In particular chromosome 8, 13 and 20 typically have long-ranged chromosomal abnormalities in colon cancer, and we focussed on these chromosomes.

Approach			#Segments	
A0	A1	A2	$\gamma = 0.01$	$\gamma = 0.05$
−	−	−	165	319
−	+	−	71	18
+	+	−	6	14
+	+	+	5	7
+	−	−	0	1

Table 1. The number of maximal segments that are significant at $\alpha = 0.05$ (+) or not (−) when applying A0, A1, A2 for $\gamma = 0.01, 0.05$, using data from Carvalho *et al.* (2009). For both values of γ, the number of significant segments is low compared to the total number of maximal segments, and those that are significant under A2 form a subset of those that are significant under A0 and A1.

We evaluated maximal segments using A0, A1 and A2. In general the results are in agreement with each other, but A1 is less strict than A2 (Table 1). For $\gamma = 0.01$, we identified one significant segment (spanning 36 markers) out of 37 segments on chromosome 8. The number of maximal segments only drop to 34 for $\gamma = 0.05$, still with one being significant (spanning 68 markers, including the previous 36). By inspection, the intensities are higher in the carcinomas than in the adenomas, corresponding to a chromosomal amplification in carcinoma of the entire p-arm. On chromosome 13, the entire chromosome (279 markers) is significant for $\gamma = 0.05$, while none of 28 maximal segments were significant for $\gamma = 0.01$. Likewise on chromosome 20, a maximal segment covering essentially the whole chromosome (192 markers out of 227) was significantly identified with $\gamma = 0.01$ (out of 4 maximal segments). For $\gamma = 0.05$, the maximal segment increased to 199 markers. These findings are in agreement with the literature (Jasmine *et al.*, 2012).

Either of the other approaches identifies more significant maximal segments. This is likely because the local correlation structure is not properly taken into account, implying that in regions with high correlation between markers, maximal segments appear artificially more significant than they actually are. Since A0 and A1 are computationally more efficient than A2, the results suggest that A0 and/or A1 might be used for screening and only those segments that are significant under A0 and/or A1 will be evaluated under A2.

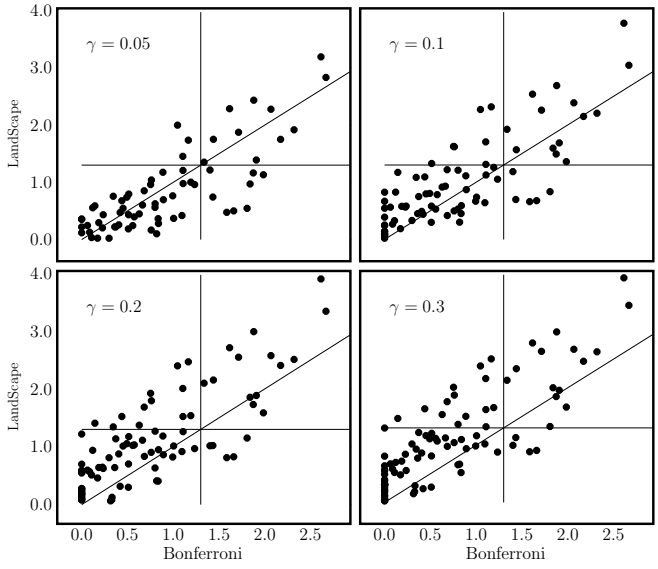

Figure 3. Analysis of 108 MGS regions: LandScape analysis using A2 with $\gamma = 0.05, 0.1, 0.2, 0.3$, and Bonferroni corrected single marker tests. For each region, the lowest Bonferroni corrected p-value is plotted against the p-value (A2) of the most significant maximal segment (if any) on $-\log_{10}(p)$ scale. The number of regions for which no maximal segments are identified (hence not showed in the figure) are 32, 22, 12, 6, resp. and the Bonferroni corrected p-values are non-significant in all these cases. Thin black lines indicate the diagonal ($x = y$) and 0.05 vertically and horizontally.

Bonferroni corrected analysis finds 6 significant markers in the region on chromosome 8 identified with $\gamma = 0.05$, 7 on chromosome 13, and 174 on chromosome 20, where the difference between the two groups is most visible.

In Meijer *et al.* (2015), they analyse the same data using a hierarchically structured test procedure to correct for multiple testing. They find 79 regions on chromosome 13 for which they reject the combined null hypothesis. These regions are of variable length and some are overlapping. At least 19 null hypotheses (positions) in the 79 regions must be non-true in order to explain their findings (Meijer *et al.*, 2015).

5 DISCUSSION

We have developed a method to aggregate sequentially ordered statistics and provided different means to assess the significance of the aggregated scores, that is, the scores of maximal segments. By combining the power of neighbouring variables the method may in some situations have a higher statistical power than the individual tests. This is indeed shown in real data examples, where the aggregated scores might be significant without the individual p-values being significant. Thus, our method can be a useful supplement to standard procedures relying on evaluation of test statistics individually.

In a sense, one can consider the aggregated score a smoothing of the individual test statistics (or p-values) as the values are 'smoothed' with the values of the surrounding positions. The smoothed value of position k is the score $Y(k)$ of the maximal segment containing that position.

Our method assigns p-values to maximal segments or positions. These we have to correct for multiple testing but now each of them borrow from their neighbours in contrast to the original p-values that are based on individual tests. Here we have applied a simple Bonferroni procedure for multiple testing, which still yield a potential power boost as the expected number of test that is adjusted for is considerably lower than the original number of tests.

In its current form, the method when applied to genetics is blind to the physical distance between markers, and it is ignorant of functional units and relationships. By way of evaluation, it does take into account the dependencies due to linkage disequilibrium and therefore in this respect, the physical distance. However, it might put SNPs together in the same segment that are not functionally or physically close, because they are subsequent SNPs in the index list. Adding weighting by distance or "functional awareness" is something we may do in the future.

ACKNOWLEDGEMENTS

The study was supported by grants from the Danish Strategic Research Council (2101-07-0059), the Lundbeck Foundation, Denmark, and the Danish Cancer Society.

REFERENCES

Carvalho, B., Postma, C., Mongera, S., Hopmans, E., Diskin, S., van de Wiel, M. A., van Criekinge, W., Thas, O., Matthäi, A., Cuesta, M. A., Terhaar sive Droste, J. S., Craanen, M., Schröck, E., Ylstra, B., and Meijer, G. A. (2009). Multiple putative oncogenes at the chromosome 20q amplicon contribute to colorectal adenoma to carcinoma progression. *Gut*, **58**(1), 79–89.

Cheverud, J. M. (2001). A simple correction for multiple comparisons in interval mapping genome scans. *Heredity*, **87**(1), 52–58.

Fisher, R. A. (1932). *Statistical methods for research workers*. Oliver and Boyd.

Gladwin, T. E., Derks, E. M., Risk, G., Rietschel, M., Mattheisen, M., Breuer, R., Schulze, T. G., Nöthen, M. N., Levinson, D., Shi, J., *et al.* (2012). Segment-wise genome-wide association analysis identifies a candidate region associated with schizophrenia in three independent samples. *PLoSONE*, **7**(6), e38828.

Hendricks, A. E., Dupuis, J., Logue, M. W., Myers, R. H., and Lunetta, K. L. (2014). Correction for multiple testing in a gene region. *Europ. J. Hum. Genet.*, **22**, 414–418.

Iglehart, E. (1972). Extreme values in GIG1 queue. *Ann. Math. Stat.*, **43**(2), 627–635.

Jasmine, F., Rahaman, R., Dodsworth, C., Roy, S., Paul, R., Raza, M., Paul-Brutus, R., Kamal, M., Ahsan, H., and Kibriya, H. G. (2012). A genome-wide study of cytogenetic changes in colorectal cancer using snp microarrays: Opportunities for future personalized treatment. *PLoSONE*, **7**(2), e31968.

Karlin, S. and Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. U S A*, **87**(6), 2264–2268.

Karlin, S. and Dembo, A. (1992). Limit distributions of maximal segmental score among markov-dependent partial sums. *Adv. Appl. Prob.*, **24**(1), 113–140.

Meijer, R. J., Krebs, T. J. P., and Goeman, J. J. (2015). A region-based multiple testing method for hypotheses ordered in space or time. *Stat. Appl. Genet. Mol. Biol.*, **14**(1), 1–19.

Naus, J. (1982). Approximations for distributions of scan statistics. *J. Amer. Stat. Assoc.*, **77**(377), 177–183.

Nyholt, D. R. (2004). A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am. J. Hum. Genet.*, **74**(4), 765–769.

Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, **511**(7510), 421–427.

Sanders, A. R., Levinson, D. F., Duan, J., Dennis, J. M., Li, R., Kendler, K. S., Rice, J. P., Shi, J., Mowry, B. J., Amin, F., *et al.* (2010). The internet-based mgs2 control sample: self report of mental illness. *The American journal of psychiatry*, **167**(7), 854–865.