

Approximate tail probabilities of the maximum of a chi-square field on multi-dimensional lattice points and their applications to detection of loci interactions

Satoshi Kuriki · Yoshiaki Harushima ·
Hironori Fujisawa · Nori Kurata

Received: 20 February 2012 / Revised: 1 April 2013 / Published online: 20 August 2013
© The Institute of Statistical Mathematics, Tokyo 2013

Abstract In this study, we define a chi-square random field on a multi-dimensional lattice points index set with a direct product covariance structure and consider the distribution of the maximum of this random field. We provide two approximate formulas for the upper tail probability of the distribution based on nonlinear renewal theory and an integral-geometric approach called the volume-of-tube method. This study is motivated by the detection problem of the interactive loci pairs which play an important role in forming biological species. The joint distribution of scan statistics for detecting the pairs is regarded as the chi-square random field above, and hence the multiplicity-adjusted p -value can be calculated using the proposed approximate formulas. By using these formulas, we examine the data of Mizuta, Harushima and Kurata (Proc Nat Acad Sci USA 107(47):20417–20422, 2010) who reported a new interactive loci pair of rice inter-subspecies.

Keywords Bateson–Dobzhansky–Muller model · Epistasis · Euler characteristic heuristic · Experimental crossing · Multiple testing · Nonlinear renewal theory · QTL analysis · Volume-of-tube method

S. Kuriki (✉) · H. Fujisawa
The Institute of Statistical Mathematics, 10-3 Midoricho, Tachikawa, Tokyo 190-8562, Japan
e-mail: kuriki@ism.ac.jp

H. Fujisawa
e-mail: fujisawa@ism.ac.jp

Y. Harushima · N. Kurata
National Institute of Genetics, Yata 1111, Mishima, Shizuoka 411-8540, Japan
e-mail: yharushi@lab.nig.ac.jp

N. Kurata
e-mail: nkurata@lab.nig.ac.jp

1 Introduction

1.1 Tests of multiplicity in detecting loci interactions

In genomic data analyses, genome scans for detecting loci that have some particular and interesting functions are often undertaken. These procedures are regarded as repeated statistical testings, and hence they are formalized as multiple testing procedures. In such multiple testings, one crucial point is how to adjust the multiplicity of tests. This is because the method of adjustment seriously affects the interpretation of the data analysis.

The detection of the interactive loci pairs assumed to exist in the Bateson–Dobzhansky–Muller (BDM) model, which motivates our study, is such a genome scan problem. In biological concept, “species” are defined as “groups of interbreeding natural populations which are reproductively isolated from other such groups” (Mayr 1942). The genetic mechanism for separating species is called reproductive isolation, which is observed as hybrid sterility or hybrid inviability between particular groups. The BDM model is a model for explaining the evolution of genetic incompatibility genes. More precisely, the BDM model assumes that there exist pairs of loci such that when the loci have particular genotypes, sterility or inviability occurs and hence a descendant is not produced (Dobzhansky 1951; Coyne and Orr 2004). In this paper, we refer to the interactive loci pair as the BDM pair.

The importance of studying such interactive pair loci is widely acknowledged. However, few studies have succeeded in identifying such pairs and in revealing the mechanism behind them. For the detection of BDM pairs, choosing two groups for crossing is crucial but difficult. If parents are genetically distant, then descendants cannot be produced. Conversely, if parents are too close, then sterility or inviability cannot be observed. The detection of a BDM pair of *Arabidopsis* intra-species by Bikard et al. (2009) and the detection of a BDM pair of rice inter-subspecies by Mizuta et al. (2010) are exceptionally successful studies.

The original purpose of this paper was to give an answer to a statistical problem that Mizuta et al. (2010) have faced during the course of their studies. Figure 1 is the contour plot depicting scan statistics for detecting BDM pairs in a second filial generation (F_2) population from two rice subspecies used by Mizuta et al. (2010). The horizontal and vertical axes represent loci positions in 12 chromosomes of rice. Each scan statistic is a chi-square statistic with 4 degrees of freedom, and the number of statistics is around 500,000. Because of the large number of tests, some adjustment for the multiplicity of tests is necessary. The Bonferroni adjustments are frequently used in multiple testing. However, in our case where the statistics are highly correlated with each other, the Bonferroni adjustment that is calculated without information of correlation would lead to very conservative results.

The multiplicity-adjusted p -value for correlated scan statistics is defined from the distribution of their maximum. For calculating this distribution, we require knowledge of the correlation structure or joint distribution. This structure can be determined from experimental design in the case of crossing experiments such as the detection problem of BDM pairs. In particular, when the number of statistics is large and when the correlation structure is systematic, we can consider a large number of scan statistics

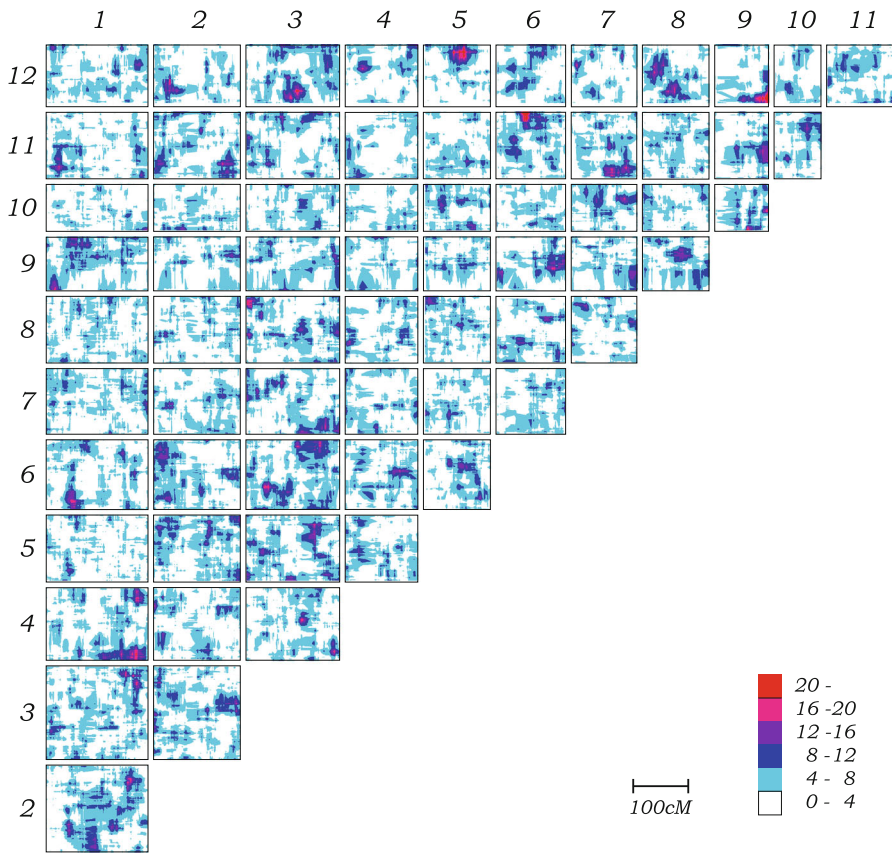


Fig. 1 Contour plot of chi-square statistics

as a random field and can obtain the distribution of the maximum. The distribution of the maximum of a random field (process) has been extensively studied. In this paper, the approaches we use are nonlinear renewal theory and the volume-of-tube method (tube method). The nonlinear renewal theory we use was developed by Woodroffe (1982) and Siegmund (1985, 1988). In this method, a random field is locally treated as a random walk, and the distribution of its maximum is obtained using sequential analysis. The volume-of-tube method is an integral-geometric approach for approximating the distribution of the maximum of a Gaussian random field through evaluating the volume of the index set (Sun 1993; Kuriki and Takemura 2001, 2009). Mathematically, this is equivalent to applying the Euler characteristic heuristic to a Gaussian field (Takemura and Kuriki 2002; Adler and Taylor 2007).

This paper is organized as follows. In Sect. 1.2, we explain the scan statistics for detecting BDM pairs. Under the null hypothesis that a BDM pair does not exist, we see that the joint distribution of the scan statistics is regarded asymptotically as a chi-square random field with a direct product covariance structure restricted on a lattice point index set. We also discuss other statistical problems that have the same stochas-

tic structure as the detection of BDM pairs in Sect. 1.3. In Sect. 2, we formalize this chi-square random field in a general setting, and provide approximate formulas for its maximum distribution using nonlinear renewal theory and the volume-of-tube method. Renewal theory assumes that the lattice points are equally spaced. This assumption may be unreasonable, because it implies that marker spacings are uniform. Hence, we use numerical comparisons to examine the difference between the randomly spaced case and the equally spaced case. The volume-of-tube method yields asymptotically conservative bounds by embedding the random field defined on a discrete set (i.e., unequally spaced lattice points) into a random field that has a continuous and piecewise smooth sample path. In Sect. 3, we analyze the data of Mizuta et al. (2010). They first screened the candidates of loci by analyzing datasets from two F_2 populations and reciprocal backcross (BC) populations, and finally succeeded in isolating causal genes of a BDM pair by positional cloning. We examine their data and confirm that their genetic finding about the BDM pair is significant from the viewpoint of multiple testing procedures. The proofs of Proposition 1, which describes the asymptotic correlation structure of the chi-square statistics for detecting interactive pairs, and the tail probability formulas in Theorems 1 and 2 are given in Sect. 4.

1.2 Scan statistics for the detection of interactive loci pairs

In this subsection, we explain the scan statistic for detecting BDM pairs and its asymptotic joint distribution for the case of the F_2 population dealt with by Mizuta et al. (2010).

We focus on the number of F_2 individuals that avoided such a fatal event and grew up. Each locus of an individual in the F_2 population produced by two strains A and B has the genotypes AA, BB, and AB. Abbreviating them to A, B, and H, respectively, the genotypes of loci 1 and 2 are cross-classified in Table 1. If this table shows some discrepancy against the independence of rows and columns, then the lack of individuals (sterility) is assumed to have happened when the loci pair has particular genotypes. Noting this, Mizuta et al. (2010) used the chi-square statistics for independence (Pearson's chi-square statistics) as scan statistics for detection. Similar scan statistics are used by Kao et al. (2010) in an F_1 spore population from an interspecies cross of yeast.

Let $T_{c_1c_2}(j_1, j_2)$ ($c_1 < c_2$) be the chi-square statistic calculated from the pair of the marker j_1 on chromosome c_1 and the marker j_2 on chromosome c_2 . The multiplicity-adjusted p -value can be obtained from the upper probability of the maximum of all chi-square statistics $\max_{c_1 < c_2} \max_{j_1, j_2} T_{c_1c_2}(j_1, j_2)$ under the null hypothesis H_0 that a BDM pair does not exist. The distribution of each statistic $T_{c_1c_2}(j_1, j_2)$ is approx-

Table 1 Cross table of genotypes in two loci (F_2)

Locus 1 \ Locus 2	A	B	H
A	n_{AA}	n_{AB}	n_{AH}
B	n_{BA}	n_{BB}	n_{BH}
H	n_{HA}	n_{HB}	n_{HH}

imated as the chi-square distribution with 4 degrees of freedom when the number n of individuals is large. However, these statistics are not independent and are highly correlated because of the linkage. Under the assumption of Haldane’s model (see e.g., [Siegmond and Yakir 2007](#), Sect. 5.6), which is the most standard model for linkage, the joint distribution under the null hypothesis H_0 is described in Proposition 1 below. The proof is given in Sect. 4.1.

Proposition 1 (a) Let d_{1j_1} (M : Morgan) be locations of markers j_1 ($= 1, \dots, m_1$) on a chromosome (chromosome 1, say). Let d_{2j_2} be locations of markers j_2 ($= 1, \dots, m_2$) on another chromosome (chromosome 2, say). Under the null hypothesis that a BDM pair does not exist, as the total sample size n goes to infinity, convergence in distribution

$$T_{12}(j_1, j_2) \Rightarrow Z_1(j_1, j_2)^2 + Z_2(j_1, j_2)^2 + Z_3(j_1, j_2)^2 + Z_4(j_1, j_2)^2 \quad (n \rightarrow \infty) \tag{1}$$

holds jointly for all (j_1, j_2) , where Z_1, \dots, Z_4 are independent, and for each k , the $Z_k(i_1, i_2)$ s are distributed according to the multivariate normal distribution with a marginal mean 0, a variance 1, and the following covariance structure:

$$\text{Cov}(Z_k(i_1, i_2), Z_k(j_1, j_2)) = e^{-\rho_{k1}|d_{1i_1}-d_{1j_1}|} \times e^{-\rho_{k2}|d_{2i_2}-d_{2j_2}|} \tag{2}$$

with

$$(\rho_{k1}, \rho_{k2}) = \begin{cases} (2, 2) & (k = 1), \\ (2, 4) & (k = 2), \\ (4, 2) & (k = 3), \\ (4, 4) & (k = 4). \end{cases} \tag{3}$$

(b) Under the null hypothesis that a BDM pair does not exist, $T_{c_1c_2}$ and $T_{c'_1c'_2}$ are asymptotically independently distributed unless $(c_1, c_2) = (c'_1, c'_2)$.

This proposition does not tell us about marker pairs belonging to the same chromosome. When two markers are located on the same chromosome, the linkage affects the independence of the rows and columns in Table 1, and the chi-square statistic simply measures the effect of the linkage directly. Since this is irrelevant to the reproductive isolation, we ignore such pairs.

Based on the asymptotic distribution given by Proposition 1, we can evaluate the multiplicity-adjusted p -value (see 17). Actually, in our genetic application, the sample size n is large enough (more than 100, at least) and this asymptotic approximation works well (see Sect. 2.5). In this context, calculation of the upper probability of the maximum of a chi-square random field on lattice points is crucial. The primary theoretical purpose of this paper is to provide approximate formulas for upper tail probability in a more general setting.

1.3 Other examples

The covariance structure in Proposition 1 also appears in other scan statistics. We illustrate two examples briefly.

The first example is the detection of epistasis in quantitative trait loci (QTL) analysis. In QTL analysis for F_2 population, phenotype y and genotype z_j are observed for each individual, where j is the index of markers, and z_j takes the values A, B, and H. The following is a simple model of QTL analysis incorporating the effects of epistasis between a loci pair (j_1, j_2) :

$$y = \mu + \sum_j (\alpha_j v_j + \beta_j w_j) + \gamma_1 v_{j_1} v_{j_2} + \gamma_2 v_{j_1} w_{j_2} + \gamma_3 w_{j_1} v_{j_2} + \gamma_4 w_{j_1} w_{j_2} + \varepsilon,$$

where $v_j = 1 (z_j = A), = 0 (z_j = H), = -1 (z_j = B), w_j = 1 (z_j = A, B), = -1 (z_j = H)$, and ε is a Gaussian measurement error. The parameters $\gamma_1, \dots, \gamma_4$ represent the epistasis. For identifying the loci pair (j_1, j_2) , the scan statistic $U(j_1, j_2)$ defined as the likelihood ratio test (LRT) statistic for testing the null hypothesis of no epistasis $\gamma_1, \dots, \gamma_4 = 0$ is used. It is shown that the asymptotic joint distribution of $\{U(j_1, j_2)\}$ is the same as that of $\{T(j_1, j_2)\}$ in Proposition 1 when j_1 and j_2 are on different chromosomes, and the multiplicity-adjusted p -value can be obtained similarly.

The second example is the detection of a change-point in two-way ordered categorical data. For the cell probability $\{p_{ij}\}_{a \times b}$, Hirotsu (1997) assumed a log-linear model with a change-point at (i_0, j_0) :

$$\log p_{ij} = \alpha_i + \beta_j + \gamma \mathbb{1}(i \leq i_0, j \leq j_0),$$

where $\mathbb{1}(\cdot)$ is the indicator function, and define a scan statistic $V(i_0, j_0)$ as the LRT statistic for testing $\gamma = 0$. Under the null hypothesis, $\{V(i_0, j_0)\}_{i_0=1, \dots, a, j_0=1, \dots, b}$ is asymptotically equivalent to $\{Z_1(j_1, j_2)^2\}$ in Proposition 1 with $d_{1j} = \log \frac{P_j}{1-P_j}, d_{2j} = \log \frac{Q_j}{1-Q_j}, P_i = \sum_{k=1}^i \sum_{l=1}^b p_{kl}, Q_j = \sum_{k=1}^a \sum_{l=1}^j p_{kl}$, and multiplicity-adjusted p -value can be obtained in our framework.

2 Approximate tail probabilities

2.1 Chi-square random fields restricted on lattice points

In this section, as a generalization of the random field referred to in Proposition 1, we define a chi-square random field on a multi-dimensional index set with a product-type covariance structure such as (2), and consider the distribution of its maximum over a multi-dimensional lattice points.

For $k = 1, \dots, m$, let us consider a real-valued continuous Gaussian random field on \mathbb{R}^p that has the following moment structure:

$$E[Z_k(t)] = 0, \quad V[Z_k(t)] = 1, \quad \text{Cov}(Z_k(t), Z_k(t')) = R_k(t - t'),$$

where for $h = (h_1, \dots, h_p)$,

$$R_k(h) = \prod_{i=1}^p R_{ki}(h_i), \quad R_{ki}(h_i) = 1 - \rho_{ki}|h_i| + o(|h_i|) \text{ as } h_i \rightarrow 0, \tag{4}$$

and ρ_{ki} is a positive constant. In particular, when $R_{ki}(h_i) = e^{-\rho_{ki}|h_i|}$, this expression represents the direct product covariance structure of the stationary Ornstein–Uhlenbeck process. Z_1, \dots, Z_m are assumed to be independent. Moreover, define

$$Z(t) = (Z_1(t), \dots, Z_m(t)), \quad Y(t) = \|Z(t)\| = \sqrt{\sum_{k=1}^m Z_k(t)^2}. \tag{5}$$

$Y(t)^2$, $t = (t_1, \dots, t_p) \in \mathbb{R}^p$ is a chi-square random field whose marginal distribution is the chi-square distribution with m degrees of freedom.

For $i = 1, \dots, p$, let $0 = d_{i0} < d_{i1} < \dots < d_{in_i}$ be distinct points, and let $T_i = \{d_{i0} (= 0), d_{i1}, \dots, d_{in_i}\}$. Define a p -dimensional unequally spaced lattice point set

$$T = T_1 \times \dots \times T_p \subset \mathbb{R}^p.$$

In this section, we provide an approximate formula for the tail probability of the maximum of the chi-square random field Y restricted on the discrete set T :

$$P\left(\max_{t \in T} Y(t) \geq b\right) \text{ as } b \rightarrow \infty. \tag{6}$$

2.2 Approximations based on nonlinear renewal theory

In this subsection, we study large-deviation approximations for the distribution of the maximum (6) in the framework of the nonlinear renewal theory devised by Woodroffe (1982) and Siegmund (1988). The outline of this method is that we first prove that $\max_{t \in T} Y(t)$ can be approximated by the maximum of a suitably defined random walk when Y is large and the spacing of lattice is small. We then evaluate the distribution of its maximum with the help of sequential analysis.

A drawback of the method is that the index set T must be an equally spaced lattice point set. That is, for all i , the points $d_{i0} < \dots < d_{in_i}$ belonging to T_i are assumed to be equally spaced as

$$d_{i1} - d_{i0} = \dots = d_{in_i} - d_{in_i-1} (= D_i, \text{ say}).$$

If the spaces are not equal, the random walk in the limit does not approach the sum of identical distributions, and hence one cannot utilize the reproductivity in the sequential analysis. However, as we show in Sect. 2.4, in typical settings for genome analysis, the upper probability for the maximum on unequally spaced lattice points is

bounded above by that for the maximum on the equally spaced lattice (i.e., the latter gives a conservative bound for the former), and the difference between them is not substantial.

Define a bounded rectangle in \mathbb{R}^p by

$$\tilde{T} = \tilde{T}_1 \times \dots \times \tilde{T}_p \subset \mathbb{R}^p, \quad \tilde{T}_i = [0, d_{in_i}].$$

For

$$j = (j_1, \dots, j_p) \in \mathbb{Z}^p, \quad D = (D_1, \dots, D_p) \in \mathbb{R}^p, \tag{7}$$

we write $jD = (j_1 D_1, \dots, j_p D_p)$. Our problem is to approximate the distribution of the maximum on p -dimensional lattice points whose spacing in the i th coordinate is D_i as follows:

$$P\left(\max_{j \in J} Y(jD) \geq b\right), \quad J = \left\{j \in \mathbb{Z}^m \mid jD \in \tilde{T}\right\}, \quad \text{as } b \rightarrow \infty.$$

By using the approach of nonlinear renewal theory, we can obtain the following formula. The proof is given in Sect. 4.2.

Theorem 1 *As $b \rightarrow \infty$, $D_i \rightarrow 0$ such that $b\sqrt{D_i} \rightarrow c_i \in (0, \infty)$, $i = 1, \dots, p$,*

$$P\left(\max_{j \in J} Y(jD) \geq b\right) \sim \frac{|\tilde{T}|}{(2\pi)^{m/2}} b^{m+2p-2} e^{-b^2/2} \int_{\mathbb{S}^{m-1}} \prod_{i=1}^p \bar{\rho}_i v(b\sqrt{2\bar{\rho}_i D_i}) du, \tag{8}$$

where du is the volume element of the unit sphere \mathbb{S}^{m-1} in \mathbb{R}^m at $u = (u_1, \dots, u_m) \in \mathbb{S}^{m-1}$,

$$\bar{\rho}_i = \bar{\rho}_i(u) = \sum_{k=1}^m u_k^2 \rho_{ki}, \tag{9}$$

$|\tilde{T}|$ is the Lebesgue measure of \tilde{T} , and

$$v(x) = \begin{cases} 2x^{-2} \exp\left\{-2 \sum_{n=1}^{\infty} n^{-1} \Phi\left(-\frac{1}{2}x\sqrt{n}\right)\right\} & (x > 0), \\ 1 & (x = 0) \end{cases}$$

with $\Phi(\cdot)$, the cumulative distribution function of the standard normal distribution.

It is reported that the asymptotic setting where $D_i = O(b^{-2})$ as $b \rightarrow \infty$ assumed in Theorem 1 leads to good approximation formulas in QTL analysis when markers are dense (Dupuis and Siegmund 1999; Siegmund 2004; Siegmund and Yakir 2007).

Remark 1 The function $v(x)$ can be conveniently approximated by the following:

$$v(x) \approx \frac{(2/x)(\Phi(x/2) - 1/2)}{(x/2)\Phi(x/2) - \phi(x/2)}, \tag{10}$$

where $\phi(\cdot)$ is the density function of the standard normal distribution (Siegmund and Yakir 2007). We use this in numerical calculations presented in Sect. 2.4.

Remark 2 The upper tail probability of the maximum of a continuous chi random field Y over a continuous set \tilde{T} can be obtained by following Piterbarg (1996), Corollary 7.1 as follows:

$$P\left(\max_{t \in \tilde{T}} Y(t) \geq b\right) \sim \frac{|\tilde{T}|}{(2\pi)^{m/2}} b^{m+2p-2} e^{-b^2/2} \int_{\mathbb{S}^{m-1}} \prod_{i=1}^p \bar{\rho}_i(u) \, du \quad (b \rightarrow \infty). \tag{11}$$

This is coincident with the right-hand side of (8) with $c_i = 0$. Since $\max_{t \in T} Y(t) \leq \max_{t \in \tilde{T}} Y(t)$, (11) is an asymptotic upper bound for (6). This can be confirmed directly from the fact $v(x) \leq 1$.

Remark 3 The Bonferroni bound of the left-hand side of (8) is

$$P\left(\max_{j \in J} Y(jD) \geq b\right) \leq |J| P(\chi_m^2 \geq b^2),$$

where χ_m^2 is a chi-square random variable with m degrees of freedom. As $b \rightarrow \infty$, this Bonferroni bound is asymptotically evaluated as

$$\frac{|\tilde{T}|}{(2\pi)^{m/2}} b^{m+2p-2} e^{-b^2/2} \frac{1}{\prod_{i=1}^p (b^2 D_i)} \int_{\mathbb{S}^{m-1}} \, du. \tag{12}$$

Here, we used $|J| = |\tilde{T}| / \prod_{i=1}^p D_i$, $P(\chi_m^2 \geq b^2) \sim b^{m-2} e^{-b^2/2} / 2^{m/2-1} \Gamma(\frac{m}{2})$ and $\int_{\mathbb{S}^{m-1}} \, du = 2\pi^{m/2} / \Gamma(\frac{m}{2})$. The right-hand side of (8) is actually bounded above by (12) because of $v(x) \leq 2x^{-2}$.

2.3 Approximations based on the volume-of-tube method

In this subsection, we provide a conservative bound for the distribution of the maximum of a chi-square random field (6) by adopting an integral-geometric approach referred to as the volume-of-tube method or the Euler characteristic heuristic.

The volume-of-tube method approximates the distribution of the maximum of a Gaussian random field that has a continuous and piecewise smooth sample path. It is particularly useful when the marginal distribution (with a fixed index) is standard normal $N(0, 1)$. (See Sun 1993; Kuriki and Takemura 2001, 2009; Takemura and Kuriki 2002; Adler and Taylor 2007.) In order to apply the volume-of-tube method to our problem, we need to describe our problem in terms of a Gaussian random field with a continuous and piecewise smooth sample path.

First, we modify the Gaussian random field Z_k on a discrete set T to define a Gaussian random field \tilde{Z}_k on a continuous set \tilde{T} that has the following properties:

- (a) $Z_k(t) = \tilde{Z}_k(t)$ (if $t \in T$).
- (b) As a function of $t \in \tilde{T}$, $\tilde{Z}_k(t)$ is continuous and piecewise smooth.

Note that continuous processes with the covariance structures given by (4) do not satisfy (b). This is because the covariance function is not differentiable at $h = 0$, and hence the sample path is not differentiable everywhere.

Define a chi random field on the index set \tilde{T} by

$$\tilde{Y}(t) = \sqrt{\sum_{k=1}^m \tilde{Z}_k(t)^2}.$$

In addition, define a Gaussian random field on the index set $\tilde{T} \times \mathbb{S}^{m-1}$ by

$$\tilde{X}(t, u) = \sum_{k=1}^m u_k \tilde{Z}_k(t), \quad u = (u_1, \dots, u_m) \in \mathbb{S}^{m-1}.$$

Since $Y(t) = \tilde{Y}(t) = \max_{u \in \mathbb{S}^{m-1}} \tilde{X}(t, u)$ for $t \in T$, we can use the upper probability of $\max_{t \in \tilde{T}} Y(t) = \max_{(t,u) \in \tilde{T} \times \mathbb{S}^{m-1}} \tilde{X}(t, u)$ as a conservative bound for that of $\max_{t \in T} Y(t)$. Note that $\tilde{X}(t, u)$ with (t, u) fixed has a standard normal distribution.

Under the volume-of-tube method, the index set $\tilde{T} \times \mathbb{S}^{m-1}$ is regarded as a Riemannian manifold endowed with a metric of

$$g(t, u) = \text{Cov}(\nabla_{(t,u)} \tilde{X}(t, u), \nabla_{(t,u)} \tilde{X}(t, u)) \tag{13}$$

at (t, u) . When a positive definite metric can be defined by (13), approximate tail probability formulas can be obtained as asymptotic expansions involving geometric invariants measured by this metric. However, even when the index set contains singularities where the metric is not properly defined, if the volume $\text{Vol}(\tilde{T} \times \mathbb{S}^{m-1})$ of the index set can only be evaluated by integrals over regular sets, the leading-term formula given below applies (Takemura and Kuriki 2003). Note that the dimension of the index set is $\dim(\tilde{T} \times \mathbb{S}^{m-1}) = p + m - 1$.

$$P\left(\max_{t \in \tilde{T}} \tilde{Y}(t) \geq b\right) = P\left(\max_{(t,u) \in \tilde{T} \times \mathbb{S}^{m-1}} \tilde{X}(t, u) \geq b\right) \\ \sim \text{Vol}(\tilde{T} \times \mathbb{S}^{m-1}) \cdot \frac{2}{(2\pi)^{(p+m)/2}} b^{p+m-2} e^{-b^2/2} \quad (b \rightarrow \infty). \tag{14}$$

There is no unique way of constructing a \tilde{Z}_k satisfying (a) and (b) from Z_k . We construct \tilde{Z}_k by undertaking the following steps.

- (i) Dissect the p -dimensional rectangle whose vertices are flanking lattice points of T ,

$$[d_{1j_1-1}, d_{1j_1}] \times \cdots \times [d_{pj_p-1}, d_{pj_p}],$$

into $p!$ simplices.

- (ii) For each simplex, define \tilde{Z}_k over the simplex by linearly interpolating the values of Z_k at vertices and multiplying by a scalar so that the variance of \tilde{Z}_k at each point of the simplex is 1.

Details of the proof of the next theorem and details of how to construct \tilde{Z}_k are given in Sect. 4.3.

Theorem 2 *Let $D_{ij} = d_{ij} - d_{ij-1}$. As $b \rightarrow \infty$ and $\max D_{ij} \rightarrow 0$,*

$$P\left(\max_{t \in T} Y(t) \geq b\right) \leq P\left(\max_{t \in T} \tilde{Y}(t) \geq b\right) \sim \frac{2V}{(2\pi)^{(m+p)/2}} b^{m+p-2} e^{-b^2/2}, \quad (15)$$

where

$$V = 2^{p/2} \prod_{i=1}^p \sum_{j=1}^{n_i} \sqrt{D_{ij}} \int_{\mathbb{S}^{m-1}} \prod_{i=1}^p \sqrt{\bar{\rho}_i(u)} \, du,$$

and $\bar{\rho}_i(u)$ is defined in (9). In addition, du is the volume element of \mathbb{S}^{m-1} at u .

Remark 4 The polynomial factor b^{m+p-2} in (15) is smaller than b^{m+2p-2} in (8) and (11). However, this does not imply that (15) is a better bound than (8). As $\max D_{ij} \rightarrow 0$, $\sum_{j=1}^{n_i} D_{ij} = O(1)$, $\sum_{j=1}^{n_i} \sqrt{D_{ij}} \geq \sum_{j=1}^{n_i} D_{ij} / \sqrt{\max D_{ij}} \rightarrow \infty$, and hence $V \rightarrow \infty$. V is not of constant order.

Ninomiya (2004) provided a conservative bound for the upper probability of the maximum of a Gaussian random field on a 2-dimensional lattice with a product-type covariance structure (4) in detecting a change-point in two-way ordered categorical data. Rebař et al. (1994) also applied the volume-of-tube method to linkage analysis. He computed thresholds for the maximum log odds (LOD) score in the interval mapping method using Rice’s formula, which is essentially equivalent to the volume-of-tube method.

2.4 Numerical comparisons of proposed formulas

This and succeeding subsections are devoted to numerical studies. In this subsection, we make numerical comparisons of three approximations: the formula based on nonlinear renewal theory (Theorem 1); the conservative bound based on continuous processes (Remark 2); and the conservative bound based on the volume-of-tube method (Theorem 2). The Bonferroni method (Remark 3) is also included as a reference. Mindful of the problem of detecting the interactive loci pairs (BDM pairs), as explained in Sect. 1, we set the parameters as follows: The dimension of the index set is $p = 2$, the chi-square degrees of freedom is $m = 4$ and 1, (ρ_{k1}, ρ_{k2}) ($k = 1, 2, 3, 4$) are in (3), $n_1 = n_2 = 50, 100, 200$, $D_{1j} = D_{2j} \equiv 0.2/100, 1/100, 5/100$ (equally spaced), $(D_{1j})_{j \geq 1} = (D_{2j})_{j \geq 1} = (0.5, 1, 0.5, 1, 3, 0.5, 1, 0.5, 1, 1, \dots)/100$ (repeat the cycle with period 10) (pattern I), $(D_{1j})_{j \geq 1} = (D_{2j})_{j \geq 1} = (0.5, 0.5, 3, 0.5, 0.5, \dots)/100$ (repeat the cycle with period 5) (pattern II), $T = [0, 1]^2$. Note that the length 1/100 corresponds to 1 cM on a chromosome.

Let $U = (U_1, \dots, U_m)$ be a random vector with a uniform distribution on the unit sphere \mathbb{S}^{m-1} in \mathbb{R}^m . An integral over \mathbb{S}^{m-1} with respect to the volume element du can be replaced by the expectation $\int_{\mathbb{S}^{m-1}} f(u)du = \text{Vol}(\mathbb{S}^{m-1}) E[f(U)]$, $\text{Vol}(\mathbb{S}^{m-1}) = 2\pi^{m/2}/\Gamma(m/2)$. In particular, we use the following for $m = 4$ and (ρ_{k1}, ρ_{k2}) given in (3):

$$E \left[\prod_{i=1}^2 \bar{\rho}_i(U) \right] = \frac{\prod_{i=1}^2 (\sum_{k=1}^m \rho_{ki}) + 2 \sum_{k=1}^m \rho_{k1}\rho_{k2}}{m(m+2)} = 9,$$

$$E \left[\prod_{i=1}^2 \sqrt{\bar{\rho}_i(U)} \right] \doteq 2.971.$$

Moreover, we use the approximation (10) in calculating the special function $\nu(x)$.

Figures 2, 3, 4 illustrate the comparisons among three approximate formulas as well as empirical distributions of Monte Carlo simulations with 10,000 iterations for the probability $P(\max_{t \in T} Y(t)^2 \geq b^2)$. Random numbers are generated from the following spatial autoregressive model: For $k = 1, \dots, m$, $i = 0, 1, \dots, n_1 (= 100)$, $j = 0, 1, \dots, n_2 (= 100)$, let $\varepsilon_k(i, j)$ be independent standard normal distributed random variables. Generate $Z_k(i, j)$ sequentially according to

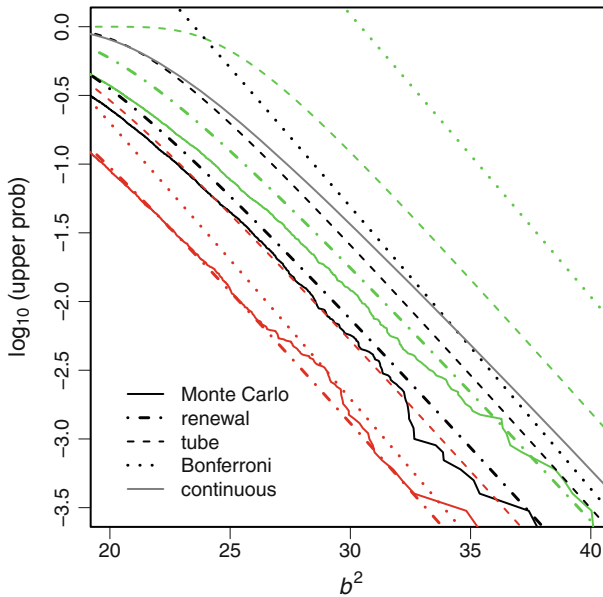


Fig. 2 Comparisons of upper probability formulas (equally spaced case). Degrees of freedom $m = 4$, $\tilde{T} = [0, 1]^2$, $D_{ij} \equiv 0.05$ (red), 0.01 (black), 0.002 (green). Continuous approximation is in gray. Monte Carlo simulations were based on 10,000 iterations

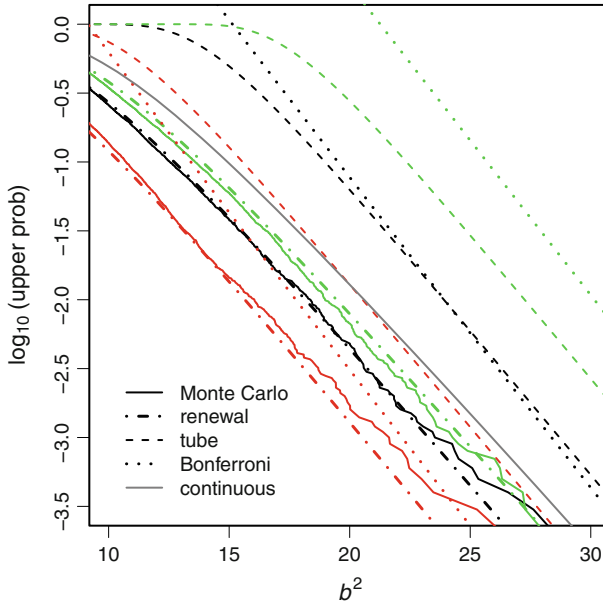


Fig. 3 Comparisons of upper probability formulas (equally spaced case). Degrees of freedom $m = 1$, $\tilde{T} = [0, 1]^2$, $D_{ij} \equiv 0.05$ (red), 0.01 (black), 0.002 (green). Continuous approximation is in gray. Monte Carlo simulations were based on 10,000 iterations

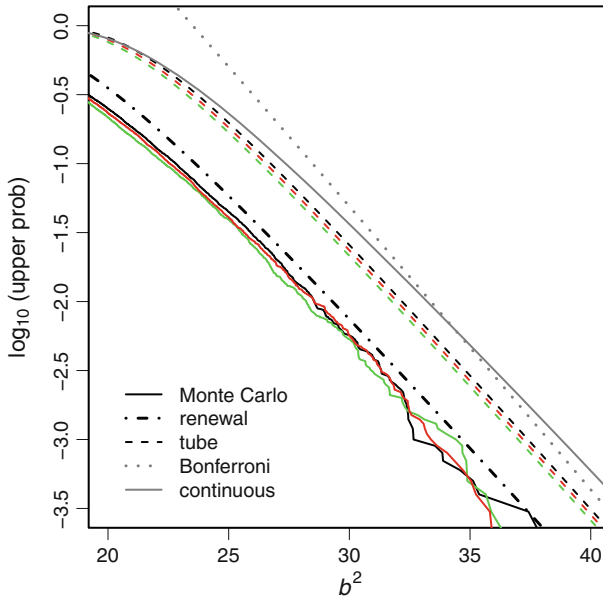


Fig. 4 Comparisons of upper probability formulas (unequally spaced case). Degrees of freedom $m = 4$, $\tilde{T} = [0, 1]^2$, $D_{ij} \equiv 0.01$ (black), pattern I: $(D_{ij})_{j \geq 1} = (0.5, 1, 0.5, 1, 3, 0.5, 1, 0.5, 1, 1, \dots)/100$ (red), pattern II: $(D_{ij})_{j \geq 1} = (0.5, 0.5, 3, 0.5, 0.5, \dots)/100$ (green). Continuous approximation and the Bonferroni bound are in gray. Monte Carlo simulations were based on 10,000 iterations

$$\left\{ \begin{array}{l} Z_k(0, 0) = \varepsilon_k(0, 0), \\ Z_k(i, 0) = \alpha_k(i)Z_k(i-1, 0) + \sqrt{1 - \alpha_k(i)^2} \varepsilon_k(i, 0) \quad (i \geq 1), \\ Z_k(0, j) = \beta_k(j)Z_k(0, j-1) + \sqrt{1 - \beta_k(j)^2} \varepsilon_k(0, j) \quad (j \geq 1), \\ Z_k(i, j) = \alpha_k(i)Z_k(i-1, j) + \beta_k(j)Z_k(i, j-1) \\ \quad - \alpha_k(i)\beta_k(j)Z_k(i-1, j-1) \\ \quad + \sqrt{1 - \alpha_k(i)^2}\sqrt{1 - \beta_k(j)^2} \varepsilon_k(i, j) \quad (i, j \geq 1), \end{array} \right. \quad (16)$$

where

$$\alpha_k(i) = e^{-\rho_{k1}D_{1i}}, \quad \beta_k(j) = e^{-\rho_{k2}D_{2j}}.$$

Then,

$$\max_{i, j \geq 0} Y(i, j)^2 = \max_{i, j \geq 0} \sum_{k=1}^4 Z_k(i, j)^2$$

is obtained. In these figures, the transformed upper probabilities of the three approximate formulas using the transformation $x \mapsto 1 - e^{-x}$ are depicted. This map is adopted by Dupuis and Siegmund (1999), (9), to restrict the maximum p -value to less than 1 without altering the asymptotic behaviors of the tail probabilities.

Figures 2 and 3 show that the formula based on nonlinear renewal theory approximates the tail probabilities well in wide ranges of the marker spacing, length of chromosomes. In particular, the case where the degree m of freedom is 1 shows greater accuracy than when $m = 4$. We conclude that the asymptotic setting where $D_i = O(b^{-2})$ ($b \rightarrow \infty$) assumed in Theorem 1 fits to our genetic applications where the marker spacings are fairly small. On the other hand, the formulas based on the volume-of-tube method and the continuous process yield upper bounds for the upper probabilities. Neither of these two methods is superior to the other. The Bonferroni method is always most conservative.

Figure 4 shows that the statistics for unequally spaced sampling are slightly below those for equally spaced sampling. This suggests that the formulas for equally spaced lattice lead to conservative p -value estimators when the sampling spaces are unequal.

2.5 Adequacy of asymptotic approximation

Throughout the paper, our arguments rely on the asymptotic approximation of Pearson's statistics to chi-square statistics. For a single contingency table, it is said that this approximation works well practically if expected cell frequencies are greater than 5 (Agresti 2002, Section 3.2.1). The sample size in our application is large enough, and this criterion holds for each loci pair table in Table 1. However, we need to be careful since we are coping with a joint distribution of many tables. Figure 5 depicts the upper probabilities of the statistics in both cases where the sample size n is finite and infinite by Monte Carlo simulations. The setting of experiments is the same as in Fig. 2 with $D_{ij} \equiv 0.01$. The curve for $n = \infty$ is the same as in Fig. 2. The curves for $n < \infty$ are

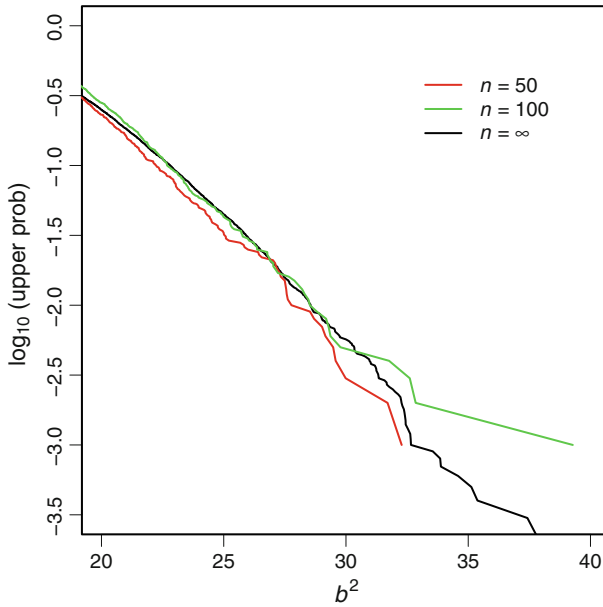


Fig. 5 Tail probabilities when n is finite and infinite. Degrees of freedom $m = 4$, $\tilde{T} = [0, 1]^2$, $D_{ij} \equiv 0.01$. Numbers of iterations were 1,000 ($n < \infty$), 10,000 ($n = \infty$)

estimated by Monte Carlo simulations with 1,000 replications. For the case $n < \infty$, we first generate the sequences of genotypes $\epsilon_i^{(t)}$, $\delta_i^{(t)}$, $\tilde{\epsilon}_j^{(t)}$, $\tilde{\delta}_j^{(t)}$ by means of Markov property (19), calculate T_{ij} by (20), and then take the maximum $\max_{i,j} T_{ij}$. Figure 5 suggests that asymptotic approximation based on chi-square distribution is practically enough even when $n = 50$.

3 Detection of interactive loci pairs

3.1 Data analysis for the F_2 population

As we explained in Sect. 1, Mizuta et al. (2010) conducted a genome scan of all pairs of marker loci of F_2 individuals of rice using chi-square statistics for independence. In this section, we reexamine the data from the viewpoint of multiple testings.

Rice has 12 chromosomes and their total length is around 1,600 cM. Two strains of rice used to produce the F_2 population are Nipponbare and Kasalath. Nipponbare is a short-grained rice in *japonica* variety, and Kasalath is a long-grained rice in *indica* variety. These two types have contrasting characteristics, and hence are used often in QTL analysis. By using Kasalath pollen, the F_1 population was produced. The F_2 is an offspring resulting from the self-pollination of F_1 individuals. The data comprise genotypes of 994 codominant markers at different locations covering the whole genome for $n = 186$ individuals of the F_2 population (Harushima et al. 1998).

Figure 1 is a contour plot of chi-square statistics calculated from all $\binom{994}{2} \doteq 500,000$ marker pairs. Because of linkage, the statistics are highly positively correlated, and large values tend to appear in neighborhoods of the “high peak”. (As stated in Sect. 1, marker pairs on the same chromosome take large values. Because these values simply measure the linkage, we ignore them.)

Table 2 shows the highest 20 peaks that do not seem to be caused by the linkage effect. The maximum chi-square statistic,

$$\max_{1 \leq c_1 < c_2 \leq 12} \max_{j_1, j_2} T_{c_1 c_2}(j_1, j_2) = 33.6,$$

is observed between markers on chromosomes 9 and 12. This corresponds to a p -value of 0.9×10^{-6} for a chi-square distribution with 4 degrees of freedom, which is highly significant if we do not take the multiplicity of tests into account. However, because of the high number of observed statistics (around 500,000), some adjustment for multiplicity is required. The Bonferroni-adjusted p -value for the maximum value is $0.9 \times 10^{-6} \times 500,000 = 0.45$. However, this is conservative because the Bonferroni adjustment does not take into account the highly positive correlations.

Table 2 The largest 20 chi-square values

No.	Marker	Chr	(cM)	Marker	Chr	(cM)	Chi-square T^a
1	R1683	9	94.1	S10637A	12	13.4	33.6 (2.9)
2	P130	6	54.0	S12886	11	116.1	33.2 (7.1)
3	V163	5	71.1	S11447	12	95.9	26.2 (1.2)
4	S2074	9	57.4	S10906	10	2.0	23.8 (7.2)
5	P60	3	92.1	S2572	12	26.5	23.3 (3.1)
6	Y5714L	1	69.1	R3203	1	160.0	21.7 (3.9)
7	S1046	1	161.9	C946	4	10.4	20.9 (2.9)
8	V10A	3	2.5	V133	8	107.0	20.7 (6.1)
9	C191A	1	141.9	C1219	3	157.1	20.6 (1.7)
10	P61	1	181.7	R2965	10	2.3	20.5 (5.9)
11	S11214	1	45.6	S1520	6	15.2	20.0 (21.1)
12	G55	3	34.4	P126	6	39.6	19.8 (7.6)
13	S1046	1	161.9	G267	4	111.2	19.8 (4.3)
14	R3192	1	26.9	C922A	1	121.0	19.7 (3.0)
15	R19	3	98.2	G7004	4	72.3	19.5 (9.3)
16	P60	3	92.1	C1424	6	112.1	19.3 (3.8)
17	R2625	1	155.3	S851	3	150.1	19.2 (2.3)
18	C506	9	93.0	Y1053R	10	34.6	19.1 (3.8)
19	S10879	9	94.4	C496	11	30.3	19.0 (2.8)
20	C2523S	7	8.8	S2545	12	72.5	19.0 (1.7)

^a Values in parentheses are chi-square T s in the second experiment

When we consider a particular chromosome pair, say (c_1, c_2) , the statistics $T_{c_1c_2}(j_1, j_2)$ ($j_1 = 1, \dots, n_{c_1}, j_2 = 1, \dots, n_{c_2}$) have the correlation structure described in Proposition 1 (a). Hence, the asymptotic null distribution of the maximum for pairs on the chromosome pair (c_1, c_2) can be evaluated. Furthermore, noting Proposition 1 (b), which states that statistics on the different pairs of chromosomes are asymptotically independent, we can evaluate the multiplicity-adjusted p -values for the maximum statistics over whole chromosomes as follows:

$$\begin{aligned}
 p\text{-value} &= F\left(\max_{1 \leq c_1 < c_2 \leq 12} \max_{j_1, j_2} T_{c_1c_2}(j_1, j_2)\right), \\
 F(x) &= 1 - \prod_{1 \leq c_1 < c_2 \leq 12} \left\{1 - P\left(\max_{t_1 \in T_{c_1}, t_2 \in T_{c_2}} Y(t_1, t_2)^2 \geq x\right)\right\}, \tag{17}
 \end{aligned}$$

where Y is a chi random field defined in (5) with $p = 2, m = 4$, and ρ_{ki} in (3). The locations (M) of markers on chromosome i are denoted by $T_i = \{d_{i0}, \dots, d_{in_i}\}$.

The multiplicity-adjusted p -value (17) for the maximum chi-square of 33.6 was estimated as 0.068 (Monte Carlo), 0.104 (renewal theory), and 0.240 (tube method). In applying Theorem 1, we substituted the average of the marker spacing on chromosome i for D_i . All of the peaks listed in Table 2 were not significant at 5%.

In the Monte Carlo method, random variables were generated from the recurrence relations in (16). Computational time was 14 days and 8 h for 10,000 iterations using a supercomputer SGI Altix3700 and the R language.

Remark 5 In QTL analysis, permutation tests are commonly used for estimating the null distribution of the maximum LOD scores (Churchill and Doerge 1994). For our problem, we can propose the procedure described below. The data set of the genotypes of all individuals is denoted by \mathcal{D} . Let Π be the set of all permutations of individual numbers. Repeat steps (i)–(ii).

- (i) Choose a permutation π from Π at random. Let \mathcal{D}_π be the data set \mathcal{D} with their individual numbers relabeled by the permutation π .
- (ii) Make cross-classified tables between all markers of \mathcal{D} and all markers of \mathcal{D}_π by their genotypes (i.e., in Table 1, locus 1 is taken from \mathcal{D} , and locus 2 is taken from \mathcal{D}_π), calculate the chi-square statistics from the tables, and find their maximum.

The null distribution of the maximum chi-square statistics can be estimated as the empirical distribution of the maxima obtained in (ii).

However, the method referred to in Remark 5 requires at least as much computational time as that required for Monte Carlo.

Moreover, Mizuta et al. (2010) performed additional genome scan searches for another F_2 population of a similar sample size. The chi-square statistics corresponding to the peaks detected in the initial experiment are listed in the last column of Table 2. Except for peak No. 11, all other peaks in Table 2 showed low values of the chi-square statistics in the second scan.

3.2 Data analysis for the BC population

Furthermore, Mizuta et al. (2010) carried out an additional experiment using the reciprocal BC population to Nipponbare. This experiment can distinguish where the interaction occurs, i.e., male gametophyte, female gametophyte, or zygote. They selected 159 markers including those exhibiting large chi-square values in the F_2 data analysis and examined the genotypes of all pairs of these selected markers in the BC populations.

Compared with the F_2 , the types of BDM pairs that can be detected from the BC population are limited. On the other hand, the detection power (the power function of test) for detectable pairs is expected to be higher.

The BC population is the experimental crossing population produced by crossing strain A with the F_1 made from strains A and B. Note that there is some arbitrariness about whether the F_1 is used as the maternal parent or pollen parent. The set of two BC populations corresponding to these two cases is called the reciprocal BC. Only genotype AB is observed in the F_1 population. Two types of genotypes, AA and AB, are observed in the BC population. We abbreviate these two genotypes to A and H, respectively. The genotypes of two loci 1 and 2 are cross-classified as shown in Table 3. The chi-square statistic for independence obtained from this table has an asymptotic chi-square distribution with 1 degree of freedom under the null hypothesis that there exists no BDM pair.

The 2×2 table showing the maximum value of the chi-square statistics is given in Table 3 (in parentheses). The maximum value is 39.6, which was observed between chromosomes 1 and 6 in the BC population with the F_1 pollen parent. The sample size was $n = 235$. This is the loci pair listed as No. 11 in Table 2. In another BC population with the F_1 maternal parent, no significant peak was observed.

In order to obtain the multiplicity-adjusted p -value for this maximum value, we need the joint distribution of the chi-square statistics. In the BC case, we can prove a proposition similar to Proposition 1: Part (a) of Proposition 1 holds if convergence in law (1) is replaced with the convergence

$$T_{12}(j_1, i_2) \Rightarrow Z_1(j_1, j_2)^2 \quad (n \rightarrow \infty).$$

Part (b) of Proposition 1 holds as it is.

The multiplicity-adjusted p -value is 2.86×10^{-6} (renewal theory) and 1.57×10^{-5} (tube method). In either case, it is highly significant. This suggests that this pair is a candidate of the BDM pair that we are seeking for and that the selection occurred in male gametophyte, pollen. Actually, Mizuta et al. (2010) confirmed that the male gametophyte selection of the unbearable genotype combination of the true

Table 3 Cross table of genotypes in two loci (BC)

Locus 1 (Chr 6 S1520) \ Locus 2 (Chr 1 S11214)	A	H
A (Nipponbare)	n_{AA} (75)	n_{AH} (13)
H	n_{HA} (64)	n_{HH} (83)

The table attaining at the maximum chi-square is shown in parentheses

BDM pair occurred through failure of pollen germination, and the reciprocal disruption of duplicated genes in the two strains caused the BDM incompatibility. Note that no other significant peaks were detected.

Finally, we discuss why the interaction was not detected in the F_2 but was in the BC. As explained in Sect. 4.1 (see Lemma 1 and succeeding descriptions), the chi-square statistic with 4 degrees of freedom obtained from Table 1 can be asymptotically decomposed into four chi-square components each with 1 degree of freedom. One of the four components corresponds to the chi-square statistic obtained from Table 3. However, in producing the BC population, there is some arbitrariness about whether F_1 is used as mother or father, and both cases are assumed to be included in the F_2 population each with a probability 1/2. Since the sample sizes for the F_2 and BC data were similar (around 200), no other significant component except for the one component with 1 degree of freedom was detected in Table 3 (in parentheses). It is convincing that the chi-square statistic of 20.0 (Table 2, No. 11) in the F_2 is almost half of that of 39.6 in the BC population (pollen parent is F_1). In conclusion, although the chi-square statistic with 4 degrees of freedom obtained from F_2 has statistical power in many directions, larger sample size was needed to detect the BDM pair.

4 Proofs

4.1 Proof of Proposition 1

First, we provide asymptotic presentations of chi-square statistics for independence when the independent model is true. Let $X = (x_{ij})_{a \times b}$ ($x_{..} = n$) be a contingency table distributed as a multinomial distribution with the cell probability $(p_{ij})_{a \times b}$ ($p_{..} = 1$). Here, we apply the convention that the summation with respect to an index is denoted by “.”. The chi-square statistic for the hypothesis of independence $H_0 : p_{ij} = p_{i.} \cdot p_{.j}$ is denoted by

$$T = T(X) = \sum_{i,j} \frac{(x_{ij} - x_{i.} \cdot x_{.j} / n)^2}{x_{i.} \cdot x_{.j} / n}.$$

The proofs of the following lemmas are easy and omitted.

Lemma 1 For a 3×3 table $X = (x_{ij})_{1 \leq i, j \leq 3}$, define four 2×2 tables:

$$X_1 = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix}, \quad X_2 = \begin{pmatrix} x_{11} + x_{12} & x_{13} \\ x_{21} + x_{22} & x_{23} \end{pmatrix}, \quad X_3 = \begin{pmatrix} x_{11} + x_{21} & x_{12} + x_{22} \\ x_{31} & x_{32} \end{pmatrix},$$

$$X_4 = \begin{pmatrix} x_{11} + x_{12} + x_{21} + x_{22} & x_{13} + x_{23} \\ x_{31} + x_{32} & x_{33} \end{pmatrix}.$$

Under H_0 , four statistics $T(X_1), T(X_2), T(X_3), T(X_4)$ are asymptotically distributed according to the independent chi-square distributions with 1 degree of freedom, and it holds that

$$T(X) = T(X_1) + T(X_2) + T(X_3) + T(X_4) + O_p(n^{-1/2}).$$

Lemma 2 For a 2×2 table, $X = (x_{ij})_{1 \leq i, j \leq 2}$ with the cell probability $(p_{ij})_{1 \leq i, j \leq 2}$,

$$T(X) = \frac{1}{n} \left(\sum_{i,j=1}^2 (-1)^{i+j} \sqrt{\frac{p_{3-i, 3-j}}{p_i \cdot p_j}} x_{ij} \right)^2 + O_p(n^{-1/2}) \tag{18}$$

holds under H_0 .

For the F_2 individuals $t = 1, \dots, n$ made from two strains A and B, by cross-classifying the genotypes of marker i ($i = 1, \dots, m$) on chromosome 1 and marker j ($j = 1, \dots, \tilde{m}$) on chromosome 2, we have the 3×3 tables represented by Table 1. Let T_{ij} be the chi-square statistic obtained from the table for marker pair (i, j) .

For individual t , let $\epsilon_i^{(t)}$ be the genotype of locus i on chromosome 1 inherited from its mother, and let $\delta_i^{(t)}$ be that from its father. Let $\tilde{\epsilon}_j^{(t)}$ be the genotype of locus j on chromosome 2 inherited from its mother, and let $\tilde{\delta}_j^{(t)}$ be that from its father. We let

$$\epsilon_i^{(t)}, \delta_i^{(t)}, \tilde{\epsilon}_j^{(t)}, \tilde{\delta}_j^{(t)} = \begin{cases} 1 & \text{(from strain A),} \\ -1 & \text{(from strain B).} \end{cases}$$

Then, the $4n$ random vectors $(\epsilon_1^{(t)}, \dots, \epsilon_m^{(t)})$, $(\delta_1^{(t)}, \dots, \delta_m^{(t)})$, $(\tilde{\epsilon}_1^{(t)}, \dots, \tilde{\epsilon}_{\tilde{m}}^{(t)})$, $(\tilde{\delta}_1^{(t)}, \dots, \tilde{\delta}_{\tilde{m}}^{(t)})$, $t = 1, \dots, n$ are independent of each other, and all elements take the value ± 1 with probabilities $1/2$ and $1/2$ satisfying a Markov property

$$P(\epsilon_{i+1}^{(t)} = \pm \epsilon_i^{(t)} \mid \epsilon_i^{(t)}) = P(\delta_{i+1}^{(t)} = \pm \delta_i^{(t)} \mid \delta_i^{(t)}) = \frac{1}{2}(1 \pm e^{-2d_{i,i+1}}). \tag{19}$$

$\tilde{\epsilon}_j^{(t)}$ and $\tilde{\delta}_j^{(t)}$ have the same Markov structure with $d_{i,i+1}$ replaced by $\tilde{d}_{j,j+1}$. Here, the genetic distance between markers i and i' on chromosome 1 is denoted by $d_{ii'}$ (M), and the genetic distance between markers j and j' on chromosome 2 is denoted by $\tilde{d}_{jj'}$ (M). This assumption of linkage is called Haldane’s model. From this model, it is easy to derive the correlation structures

$$E[\epsilon_i^{(t)} \epsilon_{i'}^{(t)}] = E[\delta_i^{(t)} \delta_{i'}^{(t)}] = e^{-2d_{ii'}}, \quad E[\tilde{\epsilon}_j^{(t)} \tilde{\epsilon}_{j'}^{(t)}] = E[\tilde{\delta}_j^{(t)} \tilde{\delta}_{j'}^{(t)}] = e^{-2\tilde{d}_{jj'}}.$$

Using this notation, the 3×3 table represented by Table 1 can be rewritten as

$$\begin{pmatrix} n_{AA} & n_{AB} & n_{AH} \\ n_{BA} & n_{BB} & n_{BH} \\ n_{HA} & n_{HB} & n_{HH} \end{pmatrix} = \sum_{t=1}^n \begin{pmatrix} \frac{1}{4}(1 + \epsilon_i^{(t)})(1 + \delta_i^{(t)}) \\ \frac{1}{4}(1 - \epsilon_i^{(t)})(1 - \delta_i^{(t)}) \\ \frac{1}{2}(1 - \epsilon_i^{(t)}\delta_i^{(t)}) \end{pmatrix} \times \begin{pmatrix} \frac{1}{4}(1 + \tilde{\epsilon}_j^{(t)})(1 + \tilde{\delta}_j^{(t)}) & \frac{1}{4}(1 - \tilde{\epsilon}_j^{(t)})(1 - \tilde{\delta}_j^{(t)}) & \frac{1}{2}(1 - \tilde{\epsilon}_j^{(t)}\tilde{\delta}_j^{(t)}) \end{pmatrix}. \tag{20}$$

In order to derive the joint distribution of the chi-square statistics T_{ij} , we decompose the 3×3 table into four 2×2 tables (i)–(iv) according to Lemma 1.

- (i) Table $\begin{pmatrix} n_{AA} & n_{AB} \\ n_{BA} & n_{BB} \end{pmatrix}$. The sum of the expected frequencies is $n/4$. From (18), the corresponding chi-square statistic has the asymptotic representation

$$T_{1,ij} = \frac{1}{n/4} (n_{AA} - n_{AB} - n_{BA} + n_{BB})^2 + O_p(n^{-1/2})$$

$$= \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n z_{1,ij}^{(t)} \right)^2 + O_p(n^{-1/2}), \quad z_{1,ij}^{(t)} = (\epsilon_i^{(t)} + \delta_i^{(t)})(\tilde{\epsilon}_j^{(t)} + \tilde{\delta}_j^{(t)})/2.$$

- (ii) Table $\begin{pmatrix} n_{AA} + n_{AB} & n_{AH} \\ n_{BA} + n_{BB} & n_{BH} \end{pmatrix}$. The sum of the expected frequencies is $n/2$. The corresponding chi-square statistic has the asymptotic representation

$$T_{2,ij} = \frac{1}{n/2} ((n_{AA} + n_{AB}) - n_{AH} - (n_{BA} + n_{BB}) + n_{BH})^2 + O_p(n^{-1/2})$$

$$= \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n z_{2,ij}^{(t)} \right)^2 + O_p(n^{-1/2}), \quad z_{2,ij}^{(t)} = (\epsilon_i^{(t)} + \delta_i^{(t)})(\tilde{\epsilon}_j^{(t)} \tilde{\delta}_j^{(t)})/\sqrt{2}.$$

- (iii) Table $\begin{pmatrix} n_{AA} + n_{BA} & n_{AB} + n_{BB} \\ n_{HA} & n_{HB} \end{pmatrix}$. The sum of the expected frequencies is $n/2$. The corresponding chi-square statistic has the asymptotic representation

$$T_{3,ij} = \frac{1}{n/2} ((n_{AA} + n_{BA}) - (n_{AB} + n_{BB}) - n_{HA} + n_{HB})^2 + O_p(n^{-1/2})$$

$$= \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n z_{3,ij}^{(t)} \right)^2 + O_p(n^{-1/2}), \quad z_{3,ij}^{(t)} = (\epsilon_i^{(t)} \delta_i^{(t)})(\tilde{\epsilon}_j^{(t)} + \tilde{\delta}_j^{(t)})/\sqrt{2}.$$

- (iv) Table $\begin{pmatrix} n_{AA} + n_{AB} + n_{BA} + n_{BB} & n_{AH} + n_{BH} \\ n_{HA} + n_{HB} & n_{HH} \end{pmatrix}$. The sum of the expected frequencies is n . The corresponding chi-square statistic has the asymptotic representation

$$T_{4,ij} = \frac{1}{n} ((n_{AA} + n_{AB} + n_{BA} + n_{BB}) - (n_{AH} + n_{BH}) - (n_{HA} + n_{HB}) + n_{HH})^2$$

$$+ O_p(n^{-1/2})$$

$$= \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n z_{4,ij}^{(t)} \right)^2 + O_p(n^{-1/2}), \quad z_{4,ij}^{(t)} = \epsilon_i^{(t)} \delta_i^{(t)} \tilde{\epsilon}_j^{(t)} \tilde{\delta}_j^{(t)}.$$

$z_{k,ij}^{(t)}$ ($k = 1, 2, 3, 4$) has a mean 0 and a covariance structure

$$\begin{aligned}
 E[z_{1,ij}^{(t)}z_{1,i'j'}^{(t)}] &= E[(\epsilon_i^{(t)} + \delta_i^{(t)})(\epsilon_{i'}^{(t)} + \delta_{i'}^{(t)})E[(\tilde{\epsilon}_j^{(t)} + \tilde{\delta}_j^{(t)})(\tilde{\epsilon}_{j'}^{(t)} + \tilde{\delta}_{j'}^{(t)})]/4 \\
 &= e^{-2d_{ii'}}e^{-2\tilde{d}_{jj'}}, \\
 E[z_{2,ij}^{(t)}z_{2,i'j'}^{(t)}] &= E[(\epsilon_i^{(t)} + \delta_i^{(t)})(\epsilon_{i'}^{(t)} + \delta_{i'}^{(t)})E[(\tilde{\epsilon}_j^{(t)}\tilde{\delta}_j^{(t)})(\tilde{\epsilon}_{j'}^{(t)}\tilde{\delta}_{j'}^{(t)})]/2 \\
 &= e^{-2d_{ii'}}e^{-4\tilde{d}_{jj'}}, \\
 E[z_{3,ij}^{(t)}z_{3,i'j'}^{(t)}] &= E[(\epsilon_i^{(t)}\delta_i^{(t)})(\epsilon_{i'}^{(t)}\delta_{i'}^{(t)})E[(\tilde{\epsilon}_j^{(t)} + \tilde{\delta}_j^{(t)})(\tilde{\epsilon}_{j'}^{(t)} + \tilde{\delta}_{j'}^{(t)})]/2 \\
 &= e^{-4d_{ii'}}e^{-2\tilde{d}_{jj'}}, \\
 E[z_{4,ij}^{(t)}z_{4,i'j'}^{(t)}] &= E[(\epsilon_i^{(t)}\delta_i^{(t)})(\epsilon_{i'}^{(t)}\delta_{i'}^{(t)})E[(\tilde{\epsilon}_j^{(t)}\tilde{\delta}_j^{(t)})(\tilde{\epsilon}_{j'}^{(t)}\tilde{\delta}_{j'}^{(t)})] = e^{-4d_{ii'}}e^{-4\tilde{d}_{jj'}}, \\
 E[z_{k,ij}^{(t)}z_{k',i'j'}^{(t)}] &= 0 \quad (k \neq k').
 \end{aligned}$$

Part (a) of Proposition 1 follows from the central limit theorem and the continuous mapping theorem.

When markers i and i' are on different chromosomes, or markers j and j' are on different chromosomes, we can let $d_{ii'} = \infty$ or $\tilde{d}_{jj'} = \infty$. In each case, $E[z_{k,ij}^{(t)}z_{k',i'j'}^{(t)}] = 0$ for all k and k' . This implies that the statistics T_{ij} and $T_{i'j'}$ are made from random variables whose limiting distributions are independent Gaussian, and hence, part (b) of Proposition 1 follows. \square

4.2 Proof of Theorem 1

The proof is divided into three parts. Section 4.2.1 provides an outline of the proof without proving a key relation (23). In Sect. 4.2.2, it is shown that the chi field $Y(t)$ restricted on lattice points is approximated by a suitably defined random walk, and that the maximum of $Y(t)$ can be approximated by the maximum of the corresponding random walk (26 and 28). Then, (23) is proved using an identity of Laplace transform provided in Sect. 4.2.3. Differently from change-point problems dealt with in previous work, the random field $Y(t)$ has a general dimensional index set and general degrees of freedom. We thereby need to introduce a random walk on a general dimensional index set, and an integral on a general dimensional unit sphere.

4.2.1 Proof of (8)

By arranging the index set J in the lexicographic order, we can let $j^0 = (j_1^0, \dots, j_d^0) \in J$ be the first point such that the random field $Y(jD)$ takes a value of at least b . Let

$$\begin{aligned}
 J^0(j^0) &= \{j \in J \mid j_1 > j_1^0, \\
 &\text{or } j_1 = j_1^0, j_2 > j_2^0, \\
 &\text{or } \dots, \\
 &\text{or } j_1 = j_1^0, \dots, j_{d-1} = j_{d-1}^0, j_d > j_d^0\}.
 \end{aligned}$$

Let \mathbb{S}^{m-1} be the unit sphere in \mathbb{R}^m . Let du be its volume element at $u \in \mathbb{S}^{m-1}$. Let $dy = (y, y + dy)$.

The event $\{\max_{j \in J} Y(jD) \geq b\}$ is exclusively divided by the value of $j^0 \in J$ (see e.g., Dupuis and Siegmund 2000, (15)) as

$$\begin{aligned}
 &P\left(\max_{j \in J} Y(jD) \geq b\right) \\
 &= \sum_{j^0 \in J} P\left(\max_{j \in J^0(j^0)} Y(jD) < b, Y(j^0D) \geq b\right) \\
 &= \int_{\mathbb{S}^{m-1}} \sum_{j^0 \in J} P\left(\max_{j \in J^0(j^0)} Y(jD) < b, Y(j^0D) \geq b, \frac{Z(j^0D)}{Y(j^0D)} \in du\right) \\
 &= \int_{y>b} \int_{\mathbb{S}^{m-1}} \sum_{j^0 \in J} P\left(\max_{j \in J^0(j^0)} Y(jD) < b, Y(j^0D) \in dy, \frac{Z(j^0D)}{Y(j^0D)} \in du\right) \\
 &= \int_{y>b} \int_{\mathbb{S}^{m-1}} \sum_{j^0 \in J} P\left(\max_{j \in J^0(j^0)} Y(jD) < b \mid Z(j^0D) = yu\right) \\
 &\quad \times P\left(Y(j^0D) \in dy, \frac{Z(j^0D)}{Y(j^0D)} \in du\right) \\
 &= \int_{x>0} \int_{\mathbb{S}^{m-1}} \sum_{j^0 \in J} P\left(\max_{j \in J^0(j^0)} Y(jD) < b \mid Z(j^0D) = yu\right) \\
 &\quad \times P\left(Y(j^0D) \in \left(b + \frac{(x, x + dx)}{b}\right), \frac{Z(j^0D)}{Y(j^0D)} \in du\right). \tag{21}
 \end{aligned}$$

In the last expression, we made change of variable $y = b + x/b$.

For fixed j^0 , $Z_k(j^0D) \sim N_m(0, I_m)$, and hence $Y(j^0D) \sim \chi_m$ and $Z(j^0D)/Y(j^0D) \sim \text{Unif}(\mathbb{S}^{m-1})$ are independent. Therefore,

$$\begin{aligned}
 &P\left(Y(j^0D) \in \left(b + \frac{(x, x + dx)}{b}\right), \frac{Z(j^0D)}{Y(j^0D)} \in du\right) \\
 &= P\left(Y(j^0D)^2 \in \left((b + x/b)^2, (b + x/b)^2 + 2dx\right)\right) \times \frac{du}{\text{Vol}(\mathbb{S}^{m-1})} \\
 &= \frac{2}{2^{m/2}\Gamma(m/2)} b^{m-2} e^{-b^2/2} e^{-x} dx \times \frac{du}{\text{Vol}(\mathbb{S}^{m-1})}. \tag{22}
 \end{aligned}$$

Moreover, as shown later,

$$\int_{x>0} P\left(\max_{j \in J^0(j^0)} Y(jD) < b \mid Z(j^0D) = yu\right) dx \sim \prod_i \bar{\rho}_i c_i^2 \nu(c_i \sqrt{2\bar{\rho}_i}) \tag{23}$$

($y = b + x/b$, $\bar{\rho}_i = \bar{\rho}_i(u)$ is in (9)).

By substituting (22) and (23) into (21) and noting that $\prod_i D_i \sum_{j_0 \in J} \sim \int_{\tilde{T}} \prod_i dt_i = |\tilde{T}|$, $\text{Vol}(\mathbb{S}^{m-1}) = 2\pi^{m/2} / \Gamma(m/2)$, we obtain

$$P\left(\max_{j \in J} Y(jD) \geq b\right) \sim \frac{|\tilde{T}|}{\prod_i D_i} \times \frac{1}{(2\pi)^{m/2}} b^{m-2} e^{-b^2/2} \int_{\mathbb{S}^{m-1}} du \prod_i \tilde{\rho}_i c_i^2 v(c_i \sqrt{2\tilde{\rho}_i}).$$

This means (8).

4.2.2 Proof of (23)

We use the large-deviation approach developed by Siegmund (1988). See also Kim and Siegmund (1989).

Suppose that t is fixed. Under a conditional probability measure given $Z(t) = (Z_k(t))_{1 \leq k \leq m} = \xi = (\xi_k)_{1 \leq k \leq m}$, the \mathbb{R}^m valued random field $Z(t+h) = (Z_k(t+h))_{1 \leq k \leq m}$ with the index $h = (h_i)_{1 \leq i \leq p}$ is a Gaussian random field with a mean of

$$E[Z_k(t+h) | \xi] = R_k(h)\xi_k,$$

and a covariance function of

$$\text{Cov}(Z_k(t+h), Z_{k'}(t+h') | \xi) = \begin{cases} R_k(h-h') - R_k(h)R_k(h') & (k = k'), \\ 0 & (k \neq k'). \end{cases}$$

When h_i is small, these moments can be rewritten as

$$E[Z_k(t+h) | \xi] = \xi_k - \xi_k \sum_{i=1}^p \rho_{ki} |h_i| + \xi_k o(|h|),$$

$$\text{Cov}(Z_k(t+h), Z_k(t+h') | \xi) = \sum_{i=1}^p \rho_{ki} (|h_i| + |h'_i| - |h_i - h'_i|) + o(|h|).$$

We consider asymptotics where

$$h_i \rightarrow 0, \quad \|\xi\| \rightarrow \infty \quad \text{such that} \quad \xi_k / \|\xi\| = u_k, \quad \|\xi\| \sqrt{|h_i|} = O(1).$$

Since $Z_k(t+h) = \xi_k + O(\sqrt{|h|}) = \xi_k(1 + O(|h|))$, we have

$$Y(t+h) = \sqrt{\sum_{k=1}^m Z_k(t+h)^2} = \|\xi\| \sqrt{1 + \frac{\sum_k (Z_k(t+h)^2 - \xi_k^2)}{\|\xi\|^2}}$$

$$\begin{aligned}
 &= \|\xi\| \left\{ 1 + \sum_k \frac{\xi_k(Z_k(t+h) - \xi_k)}{\|\xi\|^2} (1 + O(|h|)) + O(|h|^2) \right\} \\
 &= \|\xi\| + \frac{1}{\|\xi\|} \sum_k \xi_k(Z_k(t+h) - \xi_k)(1 + O(|h|)).
 \end{aligned}$$

In this expression, we used

$$Z_k(t+h)^2 - \xi_k^2 = 2\xi_k(Z_k(t+h) - \xi_k)(1 + O(|h|)) = O(1)$$

and $\xi_k(Z_k(t+h) - \xi_k)/\|\xi\|^2 = O(|h|)$. Next, consider a conditional random field with the index h defined by $\|\xi\| \{Y(t+h) - \|\xi\|\} \Big|_{Z(t)=\xi}$. The leading terms of the mean and covariance function of this field are shown to be

$$- \sum_k \|\xi\|^2 u_k^2 \sum_i \rho_{ki} |h_i|, \quad \sum_k \|\xi\|^2 u_k^2 \sum_i \rho_{ki} (|h_i| + |h'_i| - |h_i - h'_i|), \quad (24)$$

respectively.

From now on, let $t = j^0D$ and $h = (j - j^0)D$ in the multi-index notation of (7), and consider the following (finite dimensional) joint distribution under the condition that $Z(j^0D) = \xi$:

$$b\{Y(jD) - \|\xi\|\} \Big|_{Z(j^0D)=\xi}, \quad j = (j_1, \dots, j_p) \in J \subset \mathbb{Z}^p. \quad (25)$$

When

$$\|\xi\|, b \rightarrow \infty, \quad D_i \rightarrow 0 \quad \text{such that} \quad \|\xi\| \sim b, \quad b\sqrt{D_i} \rightarrow c_i \in (0, \infty),$$

from (24), the limit of the conditional mean is

$$- \sum_k u_k^2 \sum_i \rho_{ki} c_i^2 |j_i| = - \sum_i \bar{\rho}_i c_i^2 |j_i|$$

with $\bar{\rho}_i = \bar{\rho}_i(u)$ defined in (9), and the limit of the covariance between $b\{Y(jD) - \|\xi\|\}$ and $b\{Y(j'D) - \|\xi\|\}$ ($j' = (j'_1, \dots, j'_p)$) is

$$\begin{aligned}
 &\sum_k u_k^2 \sum_i \rho_{ki} c_i^2 (|j_i| + |j'_i| - |j_i - j'_i|) \\
 &= \sum_i \bar{\rho}_{ki} c_i^2 (|j_i| + |j'_i| - |j_i - j'_i|) \\
 &= \begin{cases} 2 \sum_i \bar{\rho}_{ki} c_i^2 \min(|j_i|, |j'_i|) & (j_i \text{ and } j'_i \text{ have the same sign}), \\ 0 & (\text{otherwise}). \end{cases}
 \end{aligned}$$

Since the limit becomes Gaussian again, the limiting distribution of (25) is equivalent to the distribution of

$$\sum_{i=1}^p (S_{ij_i}^+ + S_{ij_i}^-), \quad j = (j_1, \dots, j_p) \in J,$$

where

$$S_{it}^+ = \begin{cases} X_{i1} + \dots + X_{it} & (t > 0), \\ 0 & (\text{otherwise}), \end{cases}$$

$$S_{it}^- = \begin{cases} X_{i,-1} + \dots + X_{i,t} & (t < 0), \\ 0 & (\text{otherwise}), \end{cases}$$

with $X_{it} \sim N(-\bar{\rho}_i c_i^2, 2\bar{\rho}_i c_i^2)$ ($i = 1, \dots, p, t \in \mathbb{Z}$) being independent Gaussian random variables.

Summarizing the discussion above, we have proved that for $y = \|\xi\| = b + x/b \sim b$,

$$\begin{aligned} &P\left(\max_{j \in J^0(j^0)} Y(jD) < b \mid Z(j^0D) = yu\right) \\ &= P\left(\max_{j \in J^0(j^0)} b\{Y(jD) - \xi\} < -x \mid Z(j^0D) = \xi\right) \\ &\sim P\left(\max_{j \in J^0(j^0)} \sum_{i=1}^p S_{i,j_i} < -x\right). \end{aligned} \tag{26}$$

In what follows, let $j := j - j^0$ for simplicity. $j \in J^0(j^0)$ is rewritten as $j \in J^0(0)$. Let

$$M_i^+ = \max_{j>0} S_{ij}, \quad M_i^- = \max_{j\leq 0} S_{ij}.$$

Since

$$\max_{j \in J^0(0)} = \max \left[\max_{j_1>0, j_2, \dots, j_p \in \mathbb{Z}}, \max_{j_1=0, j_2>0, j_3, \dots, j_p \in \mathbb{Z}}, \dots, \max_{j_1=j_2=\dots=j_{p-1}=0, j_p>0} \right],$$

the event

$$\max_{j \in J^0(0)} \sum_{i=1}^p S_{i,j_i} < -x \tag{27}$$

is equivalent to the event that all of the following inequalities hold:

$$\begin{aligned} M_1^+ + \max\{M_2^+, M_2^-\} + \max\{M_3^+, M_3^-\} + \dots + \max\{M_p^+, M_p^-\} &< -x, \\ M_2^+ + \max\{M_3^+, M_3^-\} + \dots + \max\{M_p^+, M_p^-\} &< -x, \\ &\dots \\ M_p^+ &< -x. \end{aligned}$$

Since $M_p^- \geq 0$, if both

$$M_i^+ + \max\{M_{i+1}^+, M_{i+1}^-\} + \dots + \max\{M_{p-1}^+, M_{p-1}^-\} + M_p^- < -x$$

and $M_p^+ < -x$ hold, then

$$\begin{aligned} M_i^+ + \max\{M_{i+1}^+, M_{i+1}^-\} + \dots + \max\{M_{p-1}^+, M_{p-1}^-\} + M_p^+ \\ < -x - M_p^- + M_p^+ \\ < -2x < -x \end{aligned}$$

holds. This implies that

$$M_i^+ + \max\{M_{i+1}^+, M_{i+1}^-\} + \dots + \max\{M_{p-1}^+, M_{p-1}^-\} + \max\{M_p^+, M_p^-\} < -x.$$

Therefore, (27) is equivalent to the event that all of the following hold:

$$\begin{aligned} M_1^+ + \max\{M_2^+, M_2^-\} + \dots + \max\{M_{p-1}^+, M_{p-1}^-\} + M_p^- < -x, \\ M_2^+ + \dots + \max\{M_{p-1}^+, M_{p-1}^-\} + M_p^- < -x, \\ \dots \\ M_p^+ < -x. \end{aligned}$$

Repeating this argument reveals that (27) is equivalent to the event that all of the following inequalities hold:

$$\begin{aligned} M_1^+ + M_2^- + M_3^- + \dots + M_p^- < -x, \\ M_2^+ + M_3^- + \dots + M_p^- < -x, \\ \dots \\ M_p^+ < -x. \end{aligned}$$

That is,

$$\begin{aligned} (26) &\sim P\left(M_i^+ + M_{i+1}^- + \dots + M_p^- < -x, 1 \leq i \leq p\right) \\ &= P\left(\max_{1 \leq i \leq p} (M_i^+ + M_{i+1}^- + \dots + M_p^-) < -x\right). \end{aligned} \tag{28}$$

Because the mean μ_i and variance σ_i^2 of X_{ik} satisfy

$$\frac{-\mu_i}{\sigma_i^2} = \frac{-\bar{\rho}_i c_i^2}{2\bar{\rho}_i c_i^2} \equiv -\frac{1}{2},$$

it follows for any $p \geq 1$ that

$$\begin{aligned} & \int_0^\infty e^{-x} P\left(\max_{1 \leq i \leq p} (M_i^+ + M_{i+1}^- + \dots + M_p^-) < -x\right) dx \\ &= \prod_{i=1}^m \mu_i \nu(\mu_i / \sigma_i) = \prod_{i=1}^m \rho_i c_i^2 \nu(c_i \sqrt{2\rho_i}). \end{aligned} \tag{29}$$

A proof is given below. Combining (26), (28) and (29) yields (23).

4.2.3 Proof of (29)

Note that $M_1^+, M_1^-, \dots, M_p^+, M_p^-$ are all independent. A proof of $p = 1$ is given by Siegmund (1992), Lemma 19. For $p \geq 2$, from the integration by parts essentially proved by Siegmund (1992), Proposition 24, we have

$$\begin{aligned} & \text{RHS of (29)} \\ &= \int_0^\infty e^{-x} P\left(\max_{1 \leq i \leq p} (M_i^+ + M_{i+1}^- + \dots + M_p^-) < -x\right) dx \\ &= \int_0^\infty e^{-x} P\left(\max_{1 \leq i \leq p-1} (M_i^+ + M_{i+1}^- + \dots + M_p^-) < -x\right) P\left(M_p^+ < -x\right) dx \\ &= \mu_p \nu(2\mu_p / \sigma_p) \int_0^\infty e^{-x} P\left(\max_{1 \leq i \leq p-1} (M_i^+ + M_{i+1}^- + \dots + M_{p-1}^-) < -x\right) dx. \end{aligned}$$

The proof follows from mathematical induction. □

4.3 Proof of Theorem 2

4.3.1 Random fields defined by triangulation

First, we discuss in detail the construction of \tilde{Z}_k by triangulation of index set. It is well known that a p -dimensional cube $[0, 1]^p$ can be dissected into congruent $p!$ simplices. For example, let Π_p be the set of all permutations of $\{1, \dots, p\}$, and for each $\pi \in \Pi_p$ let

$$S_\pi = \{(x_1, \dots, x_p) \in [0, 1]^p \mid x_{\pi(1)} \geq \dots \geq x_{\pi(p)}\}.$$

Then, $[0, 1]^p = \bigcup_{\pi \in \Pi_p} S_\pi$, and S_π and $S_{\pi'}$ ($\pi \neq \pi'$) do not share any interior point. We dissect the p -dimensional rectangle whose vertices are flanking lattice points

$$[d_{1j_1-1}, d_{1j_1}] \times \dots \times [d_{pj_p-1}, d_{pj_p}]$$

into $p!$ simplices according to the same rule. Let $e_i \in \mathbb{R}^p$ be a vector whose elements are all 0 except for the i th element of the value 1. Write

$$t_0 = (t_{1j_1-1}, \dots, t_{pj_p-1}), \quad D_i = D_{ij_i} = t_{ij_i} - t_{ij_i-1} \quad (i = 1, \dots, p)$$

for simplicity. Then, one of the resulting simplices produced by the dissection is

$$\text{conv}\left\{t_0 + \sum_{l=1}^i D_l e_l \mid i = 0, 1, \dots, p\right\}. \tag{30}$$

Let

$$\xi = (\xi_0, \dots, \xi_p), \quad \xi_i = Z_k\left(t + \sum_{l=1}^i D_l e_l\right)$$

be the values of the random field Z_k at the $p + 1$ vertices of the simplex (30). This is a Gaussian random vector with a mean 0 and a covariance matrix

$$\Sigma = \begin{pmatrix} 1 & \tau_1 & \tau_1\tau_2 & \cdots & \tau_1\tau_2\tau_3 \cdots \tau_p \\ & 1 & \tau_2 & \cdots & \tau_2\tau_3 \cdots \tau_p \\ & & 1 & \cdots & \tau_3 \cdots \tau_p \\ & & & \ddots & \vdots \\ & & & & 1 \end{pmatrix}_{(p+1) \times (p+1)}, \tag{31}$$

where $\tau_i = \text{Cov}(Z_k(t), Z_k(t + D_i e_i)) = R_{ki}(D_i)$. (Although ξ and τ_i depend on k , we omit the index k for simplicity.) We can define the random field \tilde{Z}_k by interpolating the random vector ξ into the simplex (30). To be precise, by the affine bijection map from the canonical p -dimensional simplex

$$\Delta^p = \text{conv}\{0, e_1, \dots, e_p\} = \left\{s \in \mathbb{R}^p \mid 0 \leq s_i, \sum_i s_i \leq 1\right\}$$

to the simplex (30), we can introduce a parameter (local coordinates) $s = (s_i)$ into (30), and define a Gaussian random field by

$$\tilde{Z}_k(s) = \frac{(1 - \sum_i s) \xi_0 + \sum_i s_i \xi_i}{\sigma(s)},$$

where

$$\sigma(s) = \sqrt{\varphi(s)^\top \Sigma \varphi(s)}, \quad \varphi(s) = \left(1 - \sum_i s_i, s_1, \dots, s_p\right)^\top$$

is the normalizing constant so that the variance of $\tilde{Z}_k(s)$ is 1.

4.3.2 Volume of the index set of the chi-square random fields

The volume of the index set $\tilde{T} \times \mathbb{S}^{m-1}$ can be obtained by summing up the volumes of the index sets $\Delta^p \times \mathbb{S}^{m-1}$ for the Gaussian random fields

$$\tilde{X}(s, u) = \sum_{k=1}^m u_k \tilde{Z}_k(s), \quad (s, u) \in \Delta^p \times \mathbb{S}^{m-1}.$$

Let $u = u(\theta_a)$ be a local coordinate of \mathbb{S}^{m-1} . Partial derivatives with respect to s_i and θ_a are denoted by ∂_i and ∂_a , respectively. The covariance matrix of

$$\partial_i \tilde{X}(s, u) = \sum_{k=1}^m u_k \partial_i \tilde{Z}_k(s), \quad \partial_a \tilde{X}(s, u) = \sum_{k=1}^m \partial_a u_k \tilde{Z}_k(s)$$

is

$$\begin{pmatrix} \sum_{k=1}^m u_k^2 g_{k,ij}(s) & 0 \\ 0 & \bar{g}_{ab}(u) \end{pmatrix},$$

where

$$g_{k,ij}(s) = E[\partial_i \tilde{Z}_k(s) \partial_j \tilde{Z}_k(s)], \quad \bar{g}_{ab}(u) = \sum_{k=1}^m \partial_a u_k \partial_b u_k.$$

Hence, the volume of the index manifold $\Delta^p \times \mathbb{S}^{m-1}$ is

$$\text{Vol}(\Delta^p \times \mathbb{S}^{m-1}) = \int_{\Delta^p \times \mathbb{S}^{m-1}} C(s, u),$$

where

$$C(s, u) = \det \left(\sum_{k=1}^m u_k^2 g_{k,ij}(s) \right)^{1/2} \prod_i ds_i du, \quad du = \det(\bar{g}_{ab}(u))^{1/2} \prod_a d\theta_a$$

is the volume element.

We consider the case where $D_i \sim 0$, or equivalently $\tau_i \sim 1$, in Σ (31). Let J be the $(p + 1) \times (p + 1)$ matrix whose elements are all 1. Then,

$$\Sigma = J - \Sigma_1 + O(\max |1 - \tau_i|^2),$$

where Σ_1 is a symmetric matrix such that

$$(\Sigma_1)_{ii} = 0, \quad (\Sigma_1)_{ij} = \sum_{l=i}^{j-1} (1 - \tau_l) \quad (i < j).$$

By using the covariance function

$$\tilde{r}_k(s, s') = \text{Cov}(\tilde{Z}_k(s), \tilde{Z}_k(s')) = \frac{\varphi(s)^\top \Sigma \varphi(s')}{\sqrt{\varphi(s)^\top \Sigma \varphi(s) \cdot \varphi(s')^\top \Sigma \varphi(s')}},$$

the metric of the index set Δ^p is represented as

$$g_k(s) = (g_{k,ij}(s))_{1 \leq i, j \leq d}, \quad g_{k,ij}(s) = \left. \frac{\partial^2 \tilde{r}_k(s, s')}{\partial s_i \partial s'_j} \right|_{s'=s}.$$

Simple calculations yield

$$g_{k,ij} = \frac{\varphi_i^\top \Sigma \varphi_j}{\varphi^\top \Sigma \varphi} - \frac{(\varphi_i^\top \Sigma \varphi)(\varphi_j^\top \Sigma \varphi)}{(\varphi^\top \Sigma \varphi)^2},$$

$$\varphi_i = \frac{\partial \varphi(s)}{\partial s_i} = (-1, \underbrace{0, \dots, 0}_{i-1}, 1, \underbrace{0, \dots, 0}_{p-i})^\top.$$

Abbreviating $O(\max |1 - \tau_i|)$ as O yields

$$\begin{aligned} \varphi^\top \Sigma \varphi &= \varphi^\top J \varphi + O = 1 + O, & \varphi^\top \Sigma \varphi_j &= \varphi^\top J \varphi_j + O = O, \\ \varphi_i^\top \Sigma \varphi_j &= \varphi_i^\top J \varphi_j - \varphi_i^\top \Sigma_1 \varphi_j + O^2 = -\varphi_i^\top \Sigma_1 \varphi_j + O^2 \\ &= -(\Sigma_1)_{11} + (\Sigma_1)_{i+1,1} + (\Sigma_1)_{1,j+1} - (\Sigma_1)_{i+1,j+1} + O^2 \\ &= \begin{cases} \sum_{l=1}^i (1 - \tau_l) + \sum_{l=1}^j (1 - \tau_l) - \sum_{l=i+1}^j (1 - \tau_l) + O^2 & (i < j), \\ 2 \sum_{l=1}^i (1 - \tau_l) + O^2 & (i = j) \end{cases} \\ &= 2 \sum_{l=1}^i (1 - \tau_l) + O^2 \quad (i \leq j), \end{aligned}$$

and

$$g_{k,ij} = \left\{ 2 \sum_{l=1}^{\min(i,j)} (1 - \tau_l) \right\} (1 + O(\max |1 - \tau_i|)).$$

By substituting $\tau_i = 1 - \rho_{ki} D_i + o(D_i)$, we obtain

$$g_{k,ij} = \left(2 \sum_{l=1}^{\min(i,j)} \rho_{kl} D_l \right) (1 + o(1)) \quad (\max D_i \rightarrow 0).$$

Some simple calculations yield

$$\begin{aligned} \det \left(\sum_{k=1}^m u_k^2 g_{k,ij}(s) \right)^{1/2} &= \det \left(2 \sum_{l=1}^{\min(i,j)} \left(\sum_{k=1}^m u_k^2 \rho_{kl} \right) D_l \right)^{1/2} (1 + o(1)) \\ &= 2^{p/2} \prod_{i=1}^p D_i^{1/2} \prod_{i=1}^p \bar{\rho}_i(u)^{1/2} (1 + o(1)), \end{aligned}$$

where $\bar{\rho}_i(u)$ is defined in (9). Combined with

$$\int_{\Delta^p} \prod_i ds_i = \int_{0 \leq s_i, \sum s_i \leq 1} \prod_i ds_i = \frac{1}{p!},$$

we obtain the volume of the index set $\Delta^p \times \mathbb{S}^{m-1}$ as

$$\frac{2^{p/2} C}{p!} \prod_{i=1}^p D_i^{1/2} (1 + o(1)), \quad C = \int_{\mathbb{S}^{m-1}} \prod_{i=1}^p \bar{\rho}_i(u)^{1/2} du. \tag{32}$$

By letting $D_i := D_{ij_i}$, and summing up (32) with respect to $j_i = 1, \dots, n_i$ ($i = 1, \dots, p$), we can show that the volume of $\tilde{T} \times \mathbb{S}^{m-1}$ is

$$\text{Vol}(\tilde{T} \times \mathbb{S}^{m-1}) = 2^{p/2} C \prod_{i=1}^p \left(\sum_{j=1}^{n_i} D_{ij}^{1/2} \right) (1 + o(1)).$$

By substituting this into (14), we obtain the tube formula (15) for the probability $P(\max_{t \in \tilde{T}} \tilde{Y}(t) \geq b)$. □

Acknowledgments The authors are grateful to the Associate Editor and an anonymous referee for constructive comments and suggestions. They also thank David Siegmund and I-Ping Tu for helpful comments, Matt Shenton for careful proofreading, and Yuki Hasebe for help in preparing figures. This work was supported by the Systems Genetics Project of the Research Organization of Information and Systems.

References

Adler, R. J., Taylor, J. E. (2007). *Random fields and their geometry*. New York: Springer.
 Agresti, A. (2002). *Categorical data analysis*. 2nd edn. New Jersey: Wiley-Interscience.
 Bikard, D., Patel, D., Le Mett e, C., Giorgi, V., Camilleri, C., Bennett, M.J., Loudet, O. (2009). Divergent evolution of duplicate genes leads to genetic incompatibilities within *A. thaliana*. *Science*, 323(5914), 623–626.
 Churchill, G.A., Doerge, R.W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics*, 138(3), 963–971.
 Coyne, J. A., Orr, H. A. (2004). *Speciation*. Sunderland: Sinauer Associates.
 Dobzhansky, T. (1951). *Genetics and the origin of species*. 3rd edn, revised. New York: Columbia University Press.

- Dupuis, J., Siegmund, D. (1999). Statistical methods for mapping quantitative trait loci from a dense set of markers. *Genetics*, 151(1), 373–386.
- Dupuis, J., Siegmund, D. (2000). Boundary crossing probabilities in linkage analysis. In F. T. Bruss and L. Le Cam (Eds.) *Game theory, optimal stopping, probability and statistics: Papers in honor of Thomas S. Ferguson* (pp. 141–152). IMS Lecture Notes— Monograph Series (vol. 35), Beachwood: IMS.
- Harushima, Y., Yano, M., Shomura, A., Sato, M., Shimano, T., Kuboki, Y., Yamamoto, T., Lin, S. Y., Antonio, B. A., Parco, A., Kajiya, H., Huang, N., Yamamoto, K., Nagamura, Y., Kurata, N., Khush, G. S., Sasaki, T. (1998). A high-density rice genetic linkage map with 2275 markers using a single F_2 population. *Genetics*, 148(1), 479–494. Data are available at <http://rgp.dna.affrc.go.jp/pub/geneticmap98/>.
- Hirotsu, C. (1997). Two-way change point model and its application. *Australian Journal of Statistics*, 39(2), 205–218.
- Kao, K. C., Schwartz, K., Sherlock, G. (2010). A genome-wide analysis reveals no nuclear Dobzhansky-Muller pairs of determinants of speciation between *S. cerevisiae* and *S. paradoxus*, but suggests more complex incompatibilities. *PLoS Genetics*, 6, e1001038.
- Kim, H.-J., Siegmund, D. (1989). The likelihood ratio test for a change-point in simple linear regression. *Biometrika*, 76(3), 409–423.
- Kuriki, S., Takemura, A. (2001). Tail probabilities of the maxima of multilinear forms and their applications. *The Annals of Statistics*, 29(2), 328–371.
- Kuriki, S., Takemura, A. (2009). Volume of tubes and distribution of the maxima of Gaussian random fields. *Selected papers on probability and statistics*. American Mathematical Society Translations Series 2 (pp. 25–48). Rhode Island: AMS.
- Mayr, E. (1942). *Systematics and the origin of species from the viewpoint of a zoologist*. New York: Columbia University Press.
- Mizuta, Y., Harushima, Y., Kurata, N. (2010). Rice pollen hybrid incompatibility caused by reciprocal gene loss of duplicated genes. *Proceedings of the National Academy of Sciences of the United States of America*, 107(47), 20417–20422.
- Ninomiya, Y. (2004). Construction of conservative testing for change-point problems in two-dimensional random fields. *Journal of Multivariate Analysis*, 89(2), 219–242.
- Piterbarg, V. I. (1996). *Asymptotic methods in the theory of Gaussian processes and fields*. Translations of Mathematical Monographs (vol. 148). Rhode Island: AMS.
- Rebaï, A., Goffinet, B., Mangin, B. (1994). Approximate thresholds of interval mapping tests for QTL detection. *Genetics*, 138(1), 235–240.
- Siegmund, D. (1985). *Sequential analysis: tests and confidence intervals*. New York: Springer.
- Siegmund, D. (1988). Approximate tail probabilities for the maxima of some random fields. *The Annals of Probability*, 16(2), 487–501.
- Siegmund, D. O. (1992). Tail approximations for maxima of random fields. In L. H. Y. Chen, K. P. Choi, K. Hu, J.-H. Lou (Eds.) *Probability theory* (pp. 147–158). Berlin: Walter de Gruyter.
- Siegmund, D. (2004). Model selection in irregular problems: Applications to mapping quantitative trait loci. *Biometrika*, 91(4), 785–800.
- Siegmund, D., Yakir, B. (2007). *The statistics of gene mapping*. New York: Springer.
- Sun, J. (1993). Tail probabilities of the maxima of Gaussian random fields. *The Annals of Probability*, 21(1), 34–71.
- Takemura, A., Kuriki, S. (2002). On the equivalence of the tube and Euler characteristic methods for the distribution of the maximum of Gaussian fields over piecewise smooth domains. *The Annals of Applied Probability*, 12(2), 768–796.
- Takemura, A., Kuriki, S. (2003). Tail probability via tube formula when critical radius is zero. *Bernoulli*, 9(3), 535–558.
- Woodroffe, M. (1982). *Nonlinear renewal theory in sequential analysis*. CBMS-NSF Regional Conference Series in Applied Mathematics (vol. 39). Philadelphia: SIAM.