

Local Average Likelihood Ratio Test Statistics with Applications in Genomics and Change-point Detection

Hock Peng Chan* and Nancy Ruonan Zhang†

Abstract

The scan test statistic is the supremum of a set of test statistics and is often used to resolve multiple comparisons in genomics and change-point detection problems. In models where the test statistics have complicated dependence structures, the p-values of scan test statistics are hard to compute and the difficulty is compounded when nuisance parameters are present. On the other hand, standard multiple comparison procedures that are based on independence of the test statistics do not work well on these problems when there is strong local dependence. We propose in this paper the use of average likelihood ratios (ALRs) to combine test statistics that are highly correlated before comparing them through multiple comparison strategies. Limit theorems of the ALRs are provided here for quick evaluation of the overall p-values.

KEY WORDS: Cluster detection, FDR control, limit theorems, multiple comparisons, p-values, scan statistics.

*Department of Statistics and Applied Probability, National University of Singapore, e-mail: stachp@nus.edu.sg

†Department of Statistics, Stanford University, California, Stanford, e-mail: nzhang@stat.stanford.edu

1 Introduction

In Siegmund (2001), the issue of whether information can be obtained from the “peak width” in a genome scan was discussed. Let there be n locations and let Z_j be the score for a change-point at location j . Under the null hypothesis of no change, each Z_j is standard normal and the correlation of two scores decays exponentially with distance. If there is a change-point at location j , then Z_j has positive mean. The scan test statistic $M = \max_{1 \leq j \leq n} Z_j$ is often used to determine if there exists a change-point. Because Z_j for j close to a change-point have smaller but non-negligible positive mean, it does seem as if this piece of information is not utilized in the scan test statistic. Siegmund proposed a Bayes-like test statistic

$$B = n^{-1} \sum_{j=1}^n \exp(\xi Z_j - \xi^2/2), \quad (1.1)$$

where $\xi > 0$ is known, and communicated that B has slightly more power compared to the scan test statistic.

In Section 2, we introduce the average likelihood ratio (ALR) test statistic, which is like the Bayes-like test statistic (1.1), except that ξ is now replaced by maximum likelihood estimates. Instead of considering averages of the likelihood ratios for all the scores as in (1.1), we partition them into blocks and compute an average for each block of scores. Each average is a local ALR test statistic that tells us whether the scores in that block all have mean zero. ALR test statistics have an interesting property in that their p-values have low dependence on the correlation structure between the individual test statistics when the p-values are small. This can be exploited to provide easy to compute p-values that do not depend on unknown parameters. These p-values from each block can then be compared using standard multiple comparison strategies. In contrast, the p-values of scan test statistics often require sophisticated calcu-

lations and even then, some regularity structure of the dataset is needed for the formulae to be tractable. In Section 3, we discuss how the local ALR test statistics can be applied on three selected change-point detection problems and show via numerical experiments that the blocking strategy improves the performance of multiple comparison procedures. Technical details are deferred to the appendices.

2 ALR test statistics and their limiting p-values

Let $\mathbf{Z} = (Z_1, \dots, Z_n)$ be partitioned into m blocks, with $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{in_i})$ the i th block of random variables. Hence $\sum_{i=1}^m n_i = n$. Let H_i be the null hypothesis that \mathbf{Z}_i is multivariate normal with

$$E(Z_{ij}) = 0 \text{ and } \text{Var}(Z_{ij}) = 1 \text{ for all } 1 \leq j \leq n_i.$$

No assumptions are made on the correlations of \mathbf{Z}_i under H_i . The limiting uniform p-value approximations in Lemma 1 below allows us to handle unknown correlation matrices when applying ALR test statistics.

We would like to test which of the null hypotheses $\{H_i : i = 1, \dots, m\}$ are true. The direct approach would be to apply multiple comparison strategies on these n random variables and reject H_i if Z_{ij} is declared to have non-zero mean for some $1 \leq j \leq n_i$. Two common strategies are Bonferroni's correction and false discovery rate (FDR) control. Bonferroni's correction does not work well when there is heavy positive dependence and FDR methods are also based on independence of random variables.

Lemma 1 provides us with a simple method for computing a combined p-value for each block. Our suggestion is to apply multiple comparison strategies on the m combined p-values rather than on the n random variables separately, so that dependencies within each block do not affect the performance of these strategies. We summarize the information in each block

using a local ALR test statistic but instead of assuming ξ known as in (1.1), we maximize the exponents $\xi Z_{ij} - \xi^2/2$ over ξ to obtain

$$A_i^{(2)} = n_i^{-1} \sum_{j=1}^{n_i} \exp\left(\frac{Z_{ij}^2}{2}\right) \text{ for } 1 \leq i \leq m, \quad (2.1)$$

the superscript (2) denoting a two-sided test. For a one-sided version, maximize the exponents over $\xi \geq 0$ to obtain

$$A_i^{(1)} = n_i^{-1} \sum_{j=1}^{n_i} \exp\left(\frac{Z_{ij+}^2}{2}\right) \text{ for } 1 \leq i \leq m, \quad (2.2)$$

where $Z_{ij+} = \max\{0, Z_{ij}\}$ denotes the positive part of Z_{ij} . The p-values of these ALR test statistics can be computed from the transformations

$$p_i = \frac{k}{2A_i^{(k)} \sqrt{\pi \log A_i^{(k)}}} \text{ for } k = 1, 2. \quad (2.3)$$

Let $n^* = \max_{1 \leq i \leq m} n_i$ and let $a_j \sim b_j$ if $\lim_j (a_j/b_j) = 1$.

LEMMA 1. *Assume that $\log n^* = o(|\log p^*|^{1/2})$ as $p^* \rightarrow 0$. Then*

$$P\{p_i \leq p^*\} \sim p^* \text{ uniformly over } 1 \leq i \leq m. \quad (2.4)$$

The ALR test statistic was proposed in Gangnon and Clayton (2001) and applied to a cluster detection problem in an epidemiological dataset. Like Siegmund (2001), they found the ALR test statistic to have slightly more power compared to the scan test statistic. Chan (2009) gave asymptotic p-values of these test statistics in the context of cluster detection. In Appendix A, we prove Lemma 1 for weighted ALR test statistics, extending Chan (2009).

When applying Bonferroni's correction on the combined p-values (2.1)-(2.2), we reject as false all H_i for which $p_i \leq \alpha/m$ where $\alpha > 0$ is a given overall significance level. FDR control is based on a procedure of Simes (1986) which was proposed independently in Seeger (1968), with credit given to Eklund (1963). Let $p_{(1)} \leq \dots \leq p_{(m)}$ be the sorted values of

p_1, \dots, p_m . Using Sime's procedure, we reject all H_i with the i^* smallest p-values, where i^* is the largest i satisfying $p_{(i)} \leq \alpha i/m$. No null hypothesis is rejected if $p_{(i)} > \alpha i/m$ for all i . Bonferroni's correction is a conservative procedure whereas Sime's procedure rejects at least one null hypothesis exactly α of the time when all the null hypotheses are true and the p-values compared are independent. Sime's procedure does not strictly control the family-wide error rate (FWER), in the sense that it is possible that the probability of at least one incorrect rejection exceeds α when one or more null hypotheses are false. However, when the number of null hypotheses is large, maintaining a strict FWER also means that weak signals may be missed.

We show in Theorems 1 and 2 below that Bonferroni's correction and Sime's procedure are applicable on the p-value approximations of Lemma 1. This is followed by a discussion on the role of Lemma 1 on FDR and positive FDR (pFDR) control in Section 2.1 and illustrative examples in Section 2.2.

THEOREM 1. *Assume that $\log n^* = o(|\log(\alpha/m)|^{1/2})$ as $\alpha/m \rightarrow 0$. Then*

$$P\{p_{(1)} \leq \alpha/m\} \leq (1 + o(1))\alpha.$$

THEOREM 2. *Let $\mathbf{Z}_1, \dots, \mathbf{Z}_m$ be independent. If there exists $\eta > 0$ such that $\log(n^* + 1) = o(\min\{\alpha^{-(1/2)+\eta}, |\log(\alpha/m)|^{1/2}\})$ as $m \rightarrow \infty$, then*

$$P\{p_{(i)} \leq \alpha i/m \text{ for some } 1 \leq i \leq m\} \sim \alpha.$$

Theorem 1 follows directly from Lemma 1 whereas additional non-trivial arguments, given in Appendix B, are needed to prove Theorem 2. For an example of a multiple comparison strategy which should not be applied on the p-value approximations of Lemma 1, we refer the

reader to Donoho and Jin (2004), where the second-level significance test statistic

$$\text{HC}_m^* = \max_{1 \leq i \leq \alpha m} \frac{\sqrt{m}[(i/m) - p_{(i)}]}{\sqrt{p_{(i)}(1 - p_{(i)})}}$$

was discussed. The test statistic HC_m^* is very sensitive to bias in the calculations of $p_{(i)}$ for large i whereas the approximation (2.4) is most accurate when p^* is small.

2.1 FDR and pFDR

In the seminal paper by Benjamini and Hochberg (1995), Sime's procedure was shown to achieve FDR control. Let R be the number of rejected hypotheses and let V of these be true null hypotheses. Then

$$\text{FDR} = E\left(\frac{V}{R} \mid R > 0\right)P\{R > 0\}.$$

Applying Sime's procedure ensures the control $\text{FDR} \leq \alpha$. Storey (2002, 2003) defined

$$\text{pFDR} = E\left(\frac{V}{R} \mid R > 0\right)$$

and gave the pFDR a nice Bayesian interpretation. Assume that the test statistics for the hypotheses are independent and identically distributed (i.i.d.). Let π_0 be the probability that a null hypothesis is true and $\pi_1 = 1 - \pi_0$. Let $[0, p^*]$ be the rejection region of the p-values. Then by Theorem 1 of Storey (2003),

$$\text{pFDR}(p^*) = \frac{\pi_0 P\{p \leq p^* \mid H_0 \text{ true}\}}{\pi_0 P\{p \leq p^* \mid H_0 \text{ true}\} + \pi_1 P\{p \leq p^* \mid H_0 \text{ false}\}}.$$

The idea then is to find p^* satisfying $\widehat{\text{pFDR}}(p^*) = \alpha$ using resampling methods, where $\widehat{\text{pFDR}}$ is an estimate of pFDR. If the root of $\widehat{\text{pFDR}}(p^*) = \alpha$ involves a small p^* , then we can easily apply Lemma 1. This occurs when α is small, π_1 is small or a combination of both. Storey (2002) showed that when π_1 is small, the FDR and pFDR procedures are similar. Storey, Taylor and Siegmund (2004) used novel martingale arguments to provide a more streamlined

proof of FDR control in Sime's procedure. These arguments can be adapted to incorporate Lemma 1 and provide us with the following extension.

THEOREM 3. *Let $\mathbf{Z}_1, \dots, \mathbf{Z}_n$ be independent and let π_0 be the probability that a null hypothesis is true. If $\log n^* = o(|\log \alpha|^{1/2})$ as $\alpha \rightarrow 0$ and (2.1)–(2.3) are used to compute p-values, then $\text{FDR} \sim \pi_0 \alpha$.*

Theoretical advances have been made in recent years to understand how and whether FDR methods work on dependent p-values. Benjamini and Yekutieli (2001) showed that a $\log n$ threshold adjustment ensures FDR control. Storey, Taylor and Siegmund (2004) showed that under weak dependence, as defined in their equation (7), pFDR is asymptotically controlled. Their dependence example and the dependence example in Storey (2003) involves strong local dependence and are suitable for the local ALR approach that we study here. These papers demonstrated that FDR control on dependent p-values is an important problem.

We take an approach which is quite different from the earlier papers. If we use a single ALR test statistic over all n hypotheses, then we are still adopting a FWER criterion that may be too strict in the situation of many weak signals. By blocking the hypotheses, computing the ALR for each block, and then applying the FDR criterion to ALR statistics across blocks, we deal simultaneously with local consolidation of information and global multiple testing of distantly related hypotheses.

2.2 Examples

We illustrate the limiting results here with two current applications in genomics and one classical example.

EXAMPLE 2.1. Consider the detection of single nucleotide mutations by high-throughput sequencing experiments, cf. Shendure et al. (2004) and Pop and Salzberg (2008), in particular

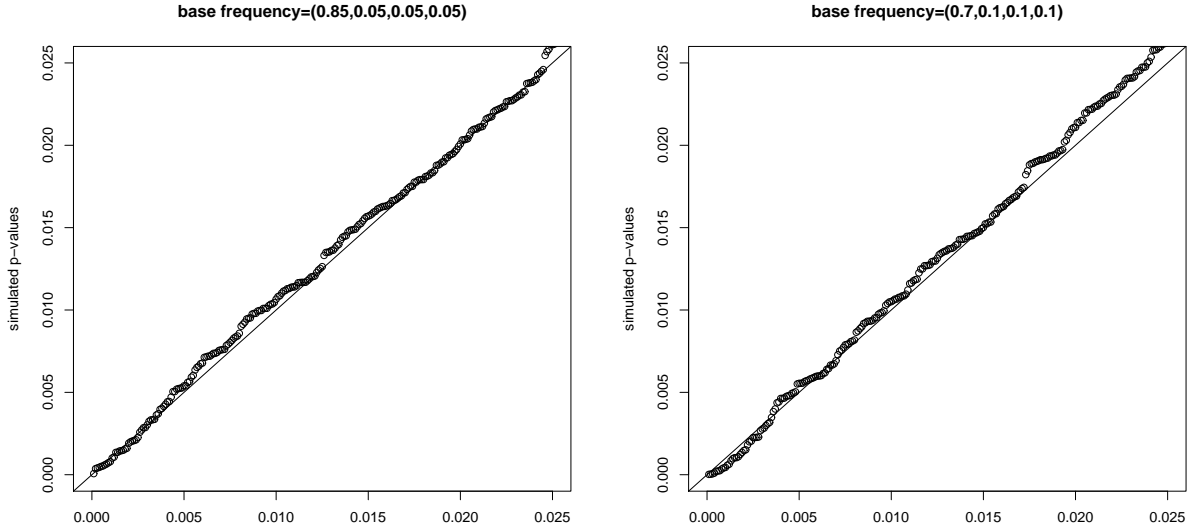


Figure 1: Plots of ordered simulated p-values at different base frequencies.

the case where the same genomic region in matched tumor and normal control samples are amplified and sequenced in parallel, see Dahl et al. (2007) and Zhang (2009). Let m be the number of base positions in the target region. We first focus on a fixed position i . The number of sequenced reads that cover the position is called the *sequencing depth*. Let N_i^C and N_i^T be the sequencing depth of the normal control and tumor samples respectively (at position i), and let Y_i^C and Y_i^T be four dimensional vectors representing the counts of the bases a, c, g and t in the control and tumor samples Then

$$Y_i^C \sim \text{Multinomial}(N_i^C, q_i^C) \text{ and } Y_i^T \sim \text{Multinomial}(N_i^T, q_i^T),$$

where $q_i^C = (q_{ia}^C, q_{ic}^C, q_{ig}^C, q_{it}^C)$ are the probabilities of observing each base in the control sample and $q_i^T = (q_{ia}^T, q_{ic}^T, q_{ig}^T, q_{it}^T)$ are the corresponding base probabilities in the tumor sample. Due to sequencing and mapping errors, we do not assume that q_i^C is concentrated on any one base, even when the cells in that sample are non-polymorphic at that position.

Under the null hypothesis H_i of no mutation, $q_i^C = q_i^T$. Under the alternative hypotheses,

the correct base in the control sample, say base a , is mutated to one of the other three bases in a fraction of cells in the tumor sample. If for example, a is mutated to g , then $q_i^T = (q_{ia}^C - r, q_{ic}^C, q_{ig}^C + r, q_{it}^C)$ for some $r > 0$ if we assume for simplicity that experimental error rates are constant across the different bases. Similar reasoning can be applied when a is mutated to c or t . We shall let H_{ij} be the hypothesis that the mutated base is j . For these types of experiments, we can assume that the sequencing error is low enough for the correct base to be determined accurately by b_i , the most frequent base occurring in the control samples at position i . Let $\hat{q}_i^C = Y_i^C/N_i^C$ and $\hat{q}_i^T = Y_i^T/N_i^T$ be the estimated base probabilities of the control and tumor samples respectively. Define $v(x, y) = x(1 - x) + y(1 - y) + 2xy$ and let

$$Z_{ij} = \frac{(\hat{q}_{ib_i}^C - \hat{q}_{ij}^C) - (\hat{q}_{ib_i}^T - \hat{q}_{ij}^T)}{\sqrt{\frac{v(\hat{q}_{ib_i}^C, \hat{q}_{ij}^C)}{N_i^C} + \frac{v(\hat{q}_{ib_i}^T, \hat{q}_{ij}^T)}{N_i^T}}}, \quad j \neq b_i.$$

Then Z_{ij} , which quantifies the evidence for mutation of the original base b_i at position i to base j for $j \neq b_i$, is approximatedly standard normal under H_i when N_i^T and N_i^C are large. The one-sided ALR test statistic for a mutation at position i is

$$A_i^{(1)} = \frac{1}{3} \sum_{j \neq b_i} \exp\left(\frac{Z_{ij}^2}{2}\right) \quad (2.5)$$

and the corresponding scan test statistic is $M_i = \max_{j \neq b_i} Z_{ij}$. In this example, we treat the three statistics $\{Z_{ij} : j \neq b_i\}$ at each sequenced position as one block, compute the average likelihood ratio statistic for each block, and applying existing multiple testing procedures across blocks.

The correlation between Z_{ij} for $j \neq b_i$, and thus the threshold level M_i for a given significance level, depend on the underlying q_i^C . In practice, q_i^C is unknown and varies with i . Using \hat{q}_i^C to estimate q_i^C is a potentially dangerous exercise for p-value computation of the scan test statistic since under H_i , M_i is large precisely when \hat{q}_i^C is a poor estimate of q_i^C or when \hat{q}_i^T is a poor estimate of q_i^T . Thus, relying on Monte Carlo simulations to determine threshold levels for the scan test statistic is not feasible. The threshold level for the ALR test

statistic, on the other hand, is almost invariant to the correlation structure for small p-values. In practice, the number of nucleotides can run into the millions and only very small p-values at each position would result in a conclusion of significance, even when FDR control is used in place of a stricter FWER.

We plot the p-value estimates p_i of the ALR test statistics, see (2.3), using 10,000 simulation runs each at base frequencies $q_i^C = q_i^T = (.7, .1, .1, .1)$ and $q_i^C = q_i^T = (.85, .05, .05, .05)$, with $N_i^C = N_i^T = 100$. Figure 1 shows that the ordered simulated p-values match well with the uniform distribution for $0 \leq \text{p-value} \leq .025$, as expected from Lemma 1.

An advantage of the ALR test statistic over the scan test statistic is its ability to aggregate information across related hypotheses in a very natural manner. The nucleotide bases a and g are purines and are found in evolutionary data to be more likely to mutate to each other than to the pyrimidines c and t . Similarly, pyrimidines are more likely to mutate to each other than to purines. Assume for the sake of argument that the base a has an 80% chance of mutating to g and 10% each to c and t . Then instead of (2.5), we can consider the weighted ALR test statistic

$$A_i^{(1)} = 0.8 \exp(Z_{ig}^2/2) + 0.1 \exp(Z_{ic}^2/2) + 0.1 \exp(Z_{it}^2/2), \quad (2.6)$$

where a is assumed to be the correct base at position i . By Lemma 1' in Appendix A, the transformations (2.3) are asymptotic p-values for (2.6). In Figure 2, we see that the weighted ALR test statistic has slightly more power compared to the scan test statistic over a wide range of mutation rates, at a significance level of 1%.

EXAMPLE 2.2. Needleman et al. (1979) studied the effects of low lead exposures by analyzing 35 different end-points in two groups of children. The end-points were divided into three families and the individual p-values are given in Table 1. By (2.1) and (2.3), the combined p-values for the three families are $p_1 = .01$, $p_2 = .1$ and $p_3 = .005$. Using Bonferroni's correction at $\alpha = .05$ and $m = 3$, significance would be declared for both "teacher's behavioral

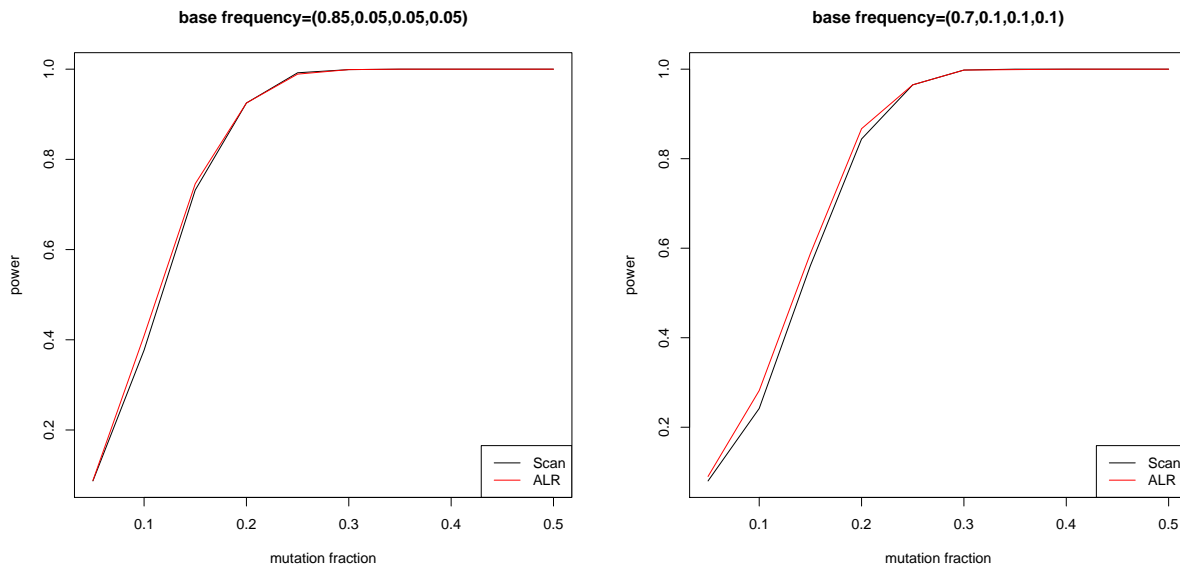


Figure 2: Power curves of the scan and weighted ALR test statistics.

<u>Family</u>	<u>P-values</u>
1. Teacher's behavioral ratings	.003, .05, .05, .14, .08, .01, .04, .01, .05, .003, .003
2. Score of Wechsler Intelligence scale of children (revised)	.04, .05, .03, .49, .08, .36, .03, .38, .15, .90, .37, .54
3. Verbal processing and reaction times	.002, .03, .07, .37, .90, .42, .05, .04, .32, .001, .001, .01

Table 1: P-values of 35 end-points in Needleman et al. (1979).

ratings” and “verbal processing and reaction times”. Conversely, if Bonferroni’s correction is used for all 35 end-points, only the p-values of .001 in “verbal processing and reaction times” would have been declared significant. The conclusions in Needleman et al. (1979) are thus justified without resorting to the use of a separate $\alpha = .05$ significance level for each family of p-values, a procedure criticized in Westfall and Young (1993). This example was recounted in Benjamini and Yekutieli (2001), where they showed that the first and third families were significant when using FDR control on all 35 end-points.

EXAMPLE 2.3. Scientists are often interested in differential gene expression between groups of samples, for example between normal cell lines and cell lines containing mutations. Subramanian et al. (2005) proposed the concept of gene-set enrichment analyses for assessing the significance of pre-defined gene-sets rather than individual genes. Let there be n genes in the study. A gene-set enrichment analysis starts with a collection of gene-sets $\{S_1, \dots, S_m\}$, where $S_i \subset \{1, \dots, n\}$ for $i = 1, \dots, m$. For each gene set S_i , a computed score statistic Z_i assesses the difference in aggregate expression of the genes in S_i between the groups of samples. Different methods for computing Z_i and assessing their significance have been proposed, for example Efron and Tibshirani (2008), Newton et al. (2006), Nobel and Wright (2005), Pavlidis et al. (2002), and Rahnenfhrer et al. (2004). After either direct standardization of Z_i or transformation of its achieved significance value, we can treat Z_i as standard normal when there is no differential gene expression in S_i .

The dataset we analyze comes from a study of p53 factor mutation in the NCI-60 collection of cancer cell lines, see Subramanian et al. (2005). There are $\ell = 10,100$ genes assayed using microarrays from 50 cell lines, of which 17 cell lines are normal and 33 cell lines contain the mutated form of the p53 factor. The goal is to find gene pathways (gene-sets) that are affected by p53 mutation. We considered a collection of 522 gene pathways developed by Subramanian et al. (2005). This is the C2 “functional” collection in their Molecular Signature Database.

The Z_i 's are computed using the GSA software of Efron and Tibshirani (2008). There is substantial overlap between the gene-sets, and thus the complex correlation structure among the Z_i 's, which depend on the overlap between the pairs of gene sets, should not be ignored.

We do not partition the scores into blocks but rather use this example to illustrate a stepwise forward selection method for identifying the list of significant gene pathways when the p-value of the ALR test statistic falls below a given significance level. We initialize with an empty list and execute the following two steps recursively.

1. Compute the ALR test statistic using all Z_j not in the list.
2. If its p-value falls below the significance level, add the gene-set with the largest $|Z_j|$ that is used for computing the ALR test statistic in step 1 into the list and return to step 1. Otherwise, end the recursion.

Using Bonferroni's correction for all n scores, two significant gene-sets, labeled as the "p53" pathway and the "hsp27" pathway, were identified. By using ALR test statistics and the stepwise forward selection method, we were able to identify two additional gene pathways. They are the "p53 hypoxia" pathway which is obviously related to p53 status, and the "SA G1 and S phases" pathway which contains 15 genes that regulate the checkpoint between the G1 and S phase of cell division. It is well known that p53 governs the G1/S transition in response to DNA damage, see Bartek and Lukas (2001) and Momand et al. (2000).

3 Applications in change-point detection

One of the cornerstones of large sample theory in statistics is the limiting chi-square distribution of $2 \log(\text{GLR})$, where GLR is the generalized likelihood ratio test statistic for testing the null hypothesis that a parameter of interest is equal to a target value, possibly in the

presence of nuisance parameters, see Begun et al. (1983), Hajek (1972), LeCam (1972) and Murphy and van der Vaart (2000). Unlike the combined p-values of Fisher (1932) for n independent tests, which is powerful for detecting many weak signals, the ALR test statistic is more inclined to declare significance when there are a few strong signals. It is recommended in situations where scan test statistics are suitable for signal identification, and is useful when the test statistics can be divided into related families, as illustrated in Example 2.2. Tail probability formulae of the maximum of GLR test statistics have been developed by Lai and Siegmund (1977) and Woodroffe (1976) using renewal theory and random walk theory, and have been successfully applied in many research areas, see Chan, Tu and Zhang (2009) for an overview of the history. However in complicated or irregular models, the asymptotic constants in these formulae can be difficult to evaluate. If Monte Carlo methods are used, then we run into difficulties when nuisance parameters are present. In Sections 3.1–3.3, we elaborate upon three important change-point detection models and we show how Lemma 1 can be used to provide p-values that do not depend on nuisance parameters. The advantage of using blocks of highly correlated random variables to compute ALR test statistics is illustrated with the help of numerical experiments in Section 3.4.

3.1 Cluster detection in epidemiological datasets

Let there be a large number of subjects located in a given domain D . Let \mathbf{t}_k be the location vector of the k th subject, Y_k an indicator of whether the subject has contracted a specified environmental disease and $\mathbf{x}_k = (x_{k1}, \dots, x_{kr})$ a covariate vector of potential influencing factors, for example age or gender. Let $p_{\mathbf{x}}(\theta, \beta)$ be the probability that a person with covariate \mathbf{x} contracts the disease, where θ is a location dependent parameter and β a nuisance parameter

vector common to all subjects. For example under the logistic model,

$$p_{\mathbf{x}}(\theta, \beta) = \frac{\exp\left(\theta + \sum_{j=1}^r \beta_j x_j\right)}{1 + \exp\left(\theta + \sum_{j=1}^r \beta_j x_j\right)}.$$

Let θ_k be the location dependent parameter of the k th subject. In spatial cluster detection, we are interested in finding a set C such that $\theta_k = \theta_C + \gamma_C$ whenever $\mathbf{t}_k \in C$ and $\theta_k = \theta_C$ whenever $\mathbf{t}_k \notin C$, with θ_C unknown and $\gamma_C \neq 0$. Let $L_C(\theta, \beta) = \prod_{\mathbf{t}_k \in B} \{[p_{\mathbf{x}_k}(\theta, \beta)]^{Y_k} [1 - p_{\mathbf{x}_k}(\theta, \beta)]^{1-Y_k}\}$ be the likelihood function for a given set C . Then the GLR test statistic for testing the null hypothesis that $\gamma_C = 0$ is

$$\text{GLR}(C) = \frac{\sup_{\theta_C, \gamma_C, \beta} \{L_C(\theta_C + \gamma_C, \beta) L_{D \setminus C}(\theta_C, \beta)\}}{\sup_{\theta, \beta} L_D(\theta, \beta)}. \quad (3.1)$$

Let C_1, \dots, C_n be the candidate spatial sets under study. The scan test statistic $M = \max_{1 \leq j \leq n} \text{GLR}_j$, where $\text{GLR}_j = \text{GLR}(C_j)$, considers the worst-case scenario. The weighted ALR test statistic

$$A^{(2)} = \sum_{j=1}^n w_j \text{GLR}_j, \quad \text{where } \sum_{j=1}^n w_j = 1, \quad (3.2)$$

was proposed in Gangnon and Clayton (2001), but covariates were not taken into account there. They were thus able to use randomized permutation tests to compute p-values of $A^{(2)}$. For one-sided versions of (3.2), restrict the supremum in the numerator of (3.1) to $\gamma_C \geq 0$.

By the theory of empirical processes, see for example Pollard (1984) Chapter 7 or Shorack and Wellner (1986), if there are ℓ subjects and $(\mathbf{t}_1, \mathbf{x}_1), \dots, (\mathbf{t}_\ell, \mathbf{x}_\ell)$ are i.i.d., then under simple regularity conditions, there exists a multivariate normal $\mathbf{Z} = (Z_1, \dots, Z_n)$ such that

$$(2 \log \text{GLR}_1, \dots, 2 \log \text{GLR}_n) \Rightarrow (Z_1^2, \dots, Z_n^2) \text{ as } \ell \rightarrow \infty. \quad (3.3)$$

Let $A_{\mathbf{Z}}^{(2)} = \sum_{i=1}^n w_i e^{Z_i^2/2}$. In Appendix A, we prove Lemma 1', a more general version of Lemma 1 which shows that the transformations (2.3) are p-values computations for $A_{\mathbf{Z}}^{(2)}$. Since $|2 \log(A_{\mathbf{Z}}^{(2)}/A^{(2)})| \leq \max_{1 \leq j \leq n} |2 \log \text{GLR}_j - Z_j^2|$, by (3.3), the transformations (2.3) are also p-value computations for $A^{(2)}$. Hence we can avoid Monte Carlo simulations when computing

p-values of $A^{(2)}$ and estimation of nuisance parameters is also not required. In Chan (2009), numerical studies on real and simulated datasets were done using global ALR test statistics. However if the domain is large, instead of computing one global test statistic, we can compute a large number of local ALR test statistics, one for each sub-divided region, and use FDR multiple comparison strategies to increase signal detection powers.

3.2 Genome scans in linkage analysis

We revisit here briefly the problem of assessing the statistical significance of whole genome scans in genetic linkage studies. The reader can refer to the excellent monograph by Siegmund and Yakir (2007) for a comprehensive treatment. We focus on a fixed chromosome pair of an organism containing part of her genetic code.

In linkage studies, we have markers that are ordered along the chromosome at positions $\zeta_1 < \dots < \zeta_n$. At each marker i , we compute a test statistic Z_i , sometimes called the *lod score*, which quantifies the evidence for proximity of marker i to the phenotype locus, that is, a location that affects an observable physical trait of interest. The vector $\mathbf{Z} = (Z_1, \dots, Z_n)$ is usually computed from a set of pedigrees or family trees, assuming a probabilistic model for recombinations along the chromosome sequence. Under the null hypothesis that none of the markers is linked with the phenotype, \mathbf{Z} is multivariate Gaussian with standard normal marginal distributions. Large values of Z_j is evidence for linkage. In genome scans, the scan statistic $M = \max_{1 \leq j \leq n} Z_j$ is often used. Under certain simplifying assumptions, including constant recombination rate, there exists a known $\beta > 0$ such that

$$\text{Cov}(Z_j, Z_k) = \exp(-\beta|\zeta_j - \zeta_k|),$$

in which case the scores follow a limiting Ornstein-Uhlenbeck process. In Siegmund and Yakir (2003), delicate calculations are used to obtain p-values of the scan statistic under the

additional assumption that the markers are equally spaced.

As suggested by Siegmund (2001), an ALR-like test statistic may improve the power of linkage tests due to the information in the markers surrounding the peak. Moreover, the p-value approximation for the ALR test statistic does not depend on the correlation structure in \mathbf{Z} , and hence is theoretically justified for cases where the uniform marker assumption or the constant recombination rate assumption is violated. If ALR test statistics are used in place of scan statistics to measure significance, knowledge and calculations of the covariance structure of the scores is not needed. It is thus a simpler and more uniformly applicable alternative to scan statistics for linkage analysis.

3.3 Genetic association studies

In genetic association studies, the goal is also to find markers that are linked to the disease trait, see Risch and Merikangas (1996). However, unlike in linkage analysis, the scores are computed based on direct association of the marker genotype to the observed phenotype in a large population. With large samples and dense markers, association studies have more power than linkage studies and can potentially identify multiple-linked loci. However, the correlation between the scores in association studies do not have a simple monotone dependence on the spacing between markers, and is often unknown or unreliably estimated. Since there are potentially many more true signals that can be captured by an association study, false discovery rates are a reasonable and popular approach to multiple hypothesis testing in this problem.

Even though the dependence of correlation between association scores on genomic distance is erratic within short ranges, it is generally true that there is no correlation between scores spaced very far apart on a chromosome. Hence, instead of applying FDR to individual markers, we can partition the genome into blocks, compute the ALR value for each block, and apply

ρ	$n = 5$			$n = 10$		
	Bonferroni	Sime's	ALR	Bonferroni	Sime's	ALR
0	4.96	5.07	4.46	4.86	4.98	4.38
.3	4.68	4.85	4.58	4.49	4.72	4.65
.6	3.85	4.23	4.64	3.46	3.93	4.87
.9	2.36	3.32	4.46	1.75	2.87	4.62

Table 2: 100 times (simulated) Type I errors. Their standard errors do not exceed .07.

FDR to these values. Blocking breaks down the dependence, thus making the existing FDR approaches more suitable. In association analysis, signals usually appear in peaks that have a width. Thus the ALR test statistics, by aggregating across markers within each block, can potentially improve the power of the genome scan. We are currently applying these ideas on a real dataset, and have obtained some positive preliminary results.

3.4 Numerical experiments

In Example 3.1, we compare the power of Sime's procedure against the ALR test statistic for various levels of dependence among the random variables. This is followed in Example 3.2 on the application of FDR control on change-point detection problems, investigating the effect of combining highly correlated adjacent random variables using local ALR test statistics.

EXAMPLE 3.1. Let $\mathbf{Z} = (Z_1, \dots, Z_n)$ be multivariate normal with $\text{Var}(Z_j) = 1$ for all j and $\text{Cov}(Z_i, Z_j) = \rho$ for $i \neq j$. Let the ALR test statistic and its computed p-value be

$$A = n^{-1} \sum_{j=1}^n \exp\left(\frac{Z_j^2}{2}\right) \text{ and } p = \frac{1}{A\sqrt{\pi \log A}}$$

respectively. Under the null hypothesis, $E(Z_j) = 0$ for all j . The actual Type I error probabilities of Bonferroni's correction, Sime's procedure and the ALR test statistic are simulated

ρ	μ	<u>5 non-zero</u>			<u>10 non-zero</u>		
		Bonferroni	Sime's	ALR	Bonferroni	Sime's	ALR
0	.5	7.8	8.0	7.3	10.4	10.8	10.2
	1	18.8	19.5	19.2	30.2	32.0	33.8
	1.5	40.7	42.7	44.2	63.2	67.3	73.2
.3	.5	7.0	7.4	7.4	8.9	9.5	10.1
	1	16.1	16.9	17.4	23.5	25.2	27.8
	1.5	33.8	35.6	37.1	48.0	51.3	56.0
.6	.5	5.3	5.9	7.0	6.6	7.5	9.2
	1	12.3	13.3	14.9	16.8	18.8	22.5
	1.5	26.8	28.7	30.8	34.7	38.1	43.6
.9	.5	2.9	4.1	6.2	3.2	5.0	7.4
	1	7.6	9.7	12.4	8.6	12.3	16.8
	1.5	17.1	20.5	24.0	19.5	25.9	32.7

Table 3: Power in terms of percentage rejection. Here $n = 10$ and the standard error in each entry does not exceed .2.

	<u>Sensitivity</u>		<u>Selectivity</u>	
	Blocking	Direct	Blocking	Direct
1. $n = 1000, \mu_{45} = 7$.439	.422	.849	.793
2. $n = 1000, \mu_{45} = \mu_{145} = 7$.479	.487	.899	.833
3. $n = 10,000, \mu_{45} = \dots = \mu_{49} = 2$.495	.493	.928	.904

Table 4: Comparing the sensitivity and selectivity of two methods of FDR control.

using 100,000 independent copies of \mathbf{Z} generated under the null hypothesis, at nominal significance level $\alpha = .05$. We then simulated the powers of these procedures, with some of the random variables given mean $\mu > 0$, using 100,000 independent copies of \mathbf{Z} for each situation. While the larger power of the ALR test statistic over Sime’s procedure at $\rho = .6$ and $.9$ can be attributed to Sime’s procedure being more conservative at these level of dependencies, see Tables 2 and 3, we also observed many instances, at $\rho = 0$ and $.3$, in which the ALR test statistic is more powerful despite having a lower actual Type I error probability. This suggests that the ALR is more effective in combining individual p-values for signal detection, at least for these types of dependencies.

EXAMPLE 3.2. Let

$$Z_j = \frac{X_j + \dots + X_{j+k-1}}{\sqrt{k}} \text{ for } j = 1, \dots, n,$$

where $X_j, 1 \leq j \leq n + k - 1$, are i.i.d. unit variance normal random variables. We start-off with direct application of FDR control at $\alpha = .05$ on $\mathbf{Z} = (Z_1, \dots, Z_n)$, to detect for random variables Z_j with non-zero means. We label this method “direct”.

We then partitioned the random variables into m blocks of $n^* = n/m$ random variables, with $\mathbf{Z}_i = (Z_{(i-1)n^*+1}, \dots, Z_{in^*})$ the random variables in the i th block. Compute p_i from \mathbf{Z}_i using (2.1) and (2.3) and apply FDR control at $\alpha = .05$ on p_1, \dots, p_m . We label this method

“blocking”. Using the direct method, the hypothesis

$$H_i : E(Z_j) = 0 \text{ for all } (i-1)n^* + 1 \leq j \leq in^*,$$

is rejected if Z_j is declared significant for some $(i-1)n^* + 1 \leq j \leq in^*$. The blocking method rejects H_i if p_i is declared to be significantly small using FDR control on p_1, \dots, p_m .

We compare the

$$\text{sensitivity} = \frac{\# \text{ correct rejections}}{\# \text{ false null hypotheses}} \text{ and } \text{selectivity} = \frac{\# \text{ correct rejections}}{\# \text{ rejections}},$$

of the two procedures by generating 100,000 copies of \mathbf{Z} , at $k = 5$, $n^* = 10$ and at various configurations of n and $\mu_r = E(X_r)$, see Table 4. All μ_r not specified in Table 4 are taken to be 0. We see here that blocking highly correlated adjacent test statistics improves the sensitivity and selectivity of FDR control.

ACKNOWLEDGEMENTS

We would like to thank Josee Dupuis, David Siegmund, and Rob Tibshirani for insightful discussions.

APPENDIX A: PROOF OF LEMMA 1

We begin with a sketch of the proof of Lemma 1 for $k = 1$. This is followed by a rigorous proof of Lemma 1', an extended version of Lemma 1, for weighted ALR test statistics. Let $\mathbf{Z} = (Z_1, \dots, Z_n)$. Showing Lemma 1 for $k = 1$ is equivalent to showing that

$$P\{A^{(1)} \geq c\} \sim \frac{1}{2c\sqrt{\pi \log c}} \text{ whenever } \log n = o(\mu), \tag{A.1}$$

where $\mu = \sqrt{2 \log c}$. Let $M = \max_{1 \leq i \leq n} Z_i$. If $\log n = o(\mu)$, then there exists $\epsilon > 0$ such that

$$\epsilon \rightarrow 0 \text{ and } ne^{-\epsilon\mu} = o(1) \text{ as } \mu \rightarrow \infty. \tag{A.2}$$

Then

$$P\{M > \mu + \epsilon\} \leq nP\{Z_1 > \mu + \epsilon\} = O(n\mu^{-1}e^{-\mu^2/2-\epsilon\mu}) \quad (\text{A.3})$$

is negligible compared to the right-hand side of (A.1). Since $A^{(1)} \leq e^{M^2/2} < c$ whenever $M < \mu$, it suffices to show (A.1) with the probability restricted to the event $\Gamma = \{\mu \leq M \leq \mu + \epsilon\}$.

Let $B = n^{-1} \sum_{i=1}^n e^{\mu Z_i - \mu^2/2}$. If for some i , $|Z_i - \mu| \leq \epsilon$, then by (A.2),

$$\frac{e^{Z_i^2/2}}{e^{\mu Z_i - \mu^2/2}} = e^{(Z_i - \mu)^2/2} \leq e^{\epsilon^2/2} \rightarrow 1. \quad (\text{A.4})$$

Under Γ , the contributions of $e^{Z_i^2/2}$ and $e^{\mu Z_i - \mu^2/2}$ to $A^{(1)}$ and B respectively are negligible when $Z_i < \mu - \epsilon$ and hence by (A.4),

$$A^{(1)} \sim B \quad (\text{under } \Gamma). \quad (\text{A.5})$$

Let Q_i be the probability measure satisfying $(dQ_i/dP)(\mathbf{Z}) = e^{\mu Z_i - \mu^2/2}$ and $Q = n^{-1} \sum_{i=1}^n Q_i$. Then $B = (dQ/dP)(\mathbf{Z})$. We shall analyze the asymptotic behavior of B under probability measure Q . Let $\rho_{ij} = \text{Cov}(Z_i, Z_j)$ and fixing i , express

$$Z_j = \rho_{ij} Z_i + (1 - \rho_{ij})^{1/2} X_{ij}, \quad (\text{A.6})$$

with $\mathbf{X}_i = (X_{i1}, \dots, X_{in})$ independent of Z_i . If we condition on \mathbf{X}_i , then Z_i becomes the only random component under Q_i . By (A.6), we have the expansion

$$\log B = \mu Z_i - \mu^2/2 + W_i(Z_i), \quad (\text{A.7})$$

where $W_i(z) = \log n^{-1} + \log(1 + \sum_{j \neq i} e^{\mu(\rho_{ij}-1)z + \mu(1-\rho_{ij})^{1/2} X_{ij}})$. Let

$$\Omega_i = \{\mathbf{X}_i : |X_{ij}| \leq (\epsilon\mu)^{1/2} \text{ for all } 1 \leq j \leq n\}. \quad (\text{A.8})$$

In the complete proof of Lemma 1', we show that the complement of Ω_i has negligible probability and for all $\mathbf{X}_i \in \Omega_i$, $|W_i(\mu)| \leq \mu\epsilon$ with $\epsilon > 0$ satisfying (A.2). Moreover under Ω_i , $W_i(z)$ varies slowly over $\mu - \epsilon \leq z \leq \mu + \epsilon$. This is because for ρ_{ij} close to 1,

the variation of $e^{\mu(\rho_{ij}-1)z+\mu(1-\rho_{ij})^{1/2}X_{ij}}$ is small whereas for ρ_{ij} significantly smaller than 1, $e^{\mu(\rho_{ij}-1)z+\mu(1-\rho_{ij})^{1/2}X_{ij}}$ provides negligible contribution to $W_i(z)$.

Let us treat $W_i(z)$ informally as a constant over $z \in [\mu - \epsilon, \mu + \epsilon]$ and denote it by $W_i(\mu)$. Then by (A.7), $Z_i = \mu - \mu^{-1}W_i(\mu)$ is the root of $B = c(= e^{\mu^2/2})$. Since $Z_i \sim N(\mu, 1)$ under Q_i ,

$$\begin{aligned}
P\{B \geq c | \mathbf{X}_i\} &= n^{-1} \sum_{i=1}^n E_{Q_i} E(B^{-1} \mathbf{I}_{\{B \geq c\}}) \\
&\sim n^{-1} \sum_{i=1}^n \int_{\mu - \mu^{-1}W_i(\mu)}^{\infty} e^{-\mu z + \mu^2/2 - W_i(\mu)} \left(\frac{1}{\sqrt{2\pi}} e^{-(z-\mu)^2/2} \right) dz \\
&= n^{-1} \sum_{i=1}^n e^{-W_i(\mu)} \int_{\mu - \mu^{-1}W_i(\mu)}^{\infty} \left(\frac{1}{\sqrt{2\pi}} e^{-z^2/2} \right) dz \\
&\sim n^{-1} \sum_{i=1}^n \frac{e^{-W_i(\mu)}}{\mu} \left(\frac{1}{\sqrt{2\pi}} e^{-[\mu - \mu^{-1}W_i(\mu)]^2/2} \right) \sim \frac{e^{-\mu^2/2}}{\mu\sqrt{2\pi}},
\end{aligned} \tag{A.9}$$

and (A.1) follows by taking expectation of (A.9) over $\mathbf{X}_i \in \Omega_i$ and substituting B by $A^{(1)}$, see (A.5).

Consider more generally weighted ALR test statistics

$$A^{(1)} = \sum_{j=1}^n w_j e^{Z_j^2/2} \text{ and } A^{(2)} = \sum_{j=1}^n w_j e^{Z_j^2/2}, \text{ with } \sum_{j=1}^n w_j = 1,$$

and p-values p given by the transformations (2.3). Let $w^* = \min_j w_j$.

LEMMA 1'. *Assume that $\log n = o(|\log p^*|^{1/2})$ and $|\log w^*| = o(|\log p^*|)$ as $p^* \rightarrow 0$. Then $P\{p \leq p^*\} \sim p^*$.*

Let $B = \sum_{j=1}^n w_j e^{\mu Z_j - \mu^2/2}$, where $\mu = (2 \log c)^{1/2} + o(1)$ as $c \rightarrow \infty$. Then (A.7) holds with

$$W_i(z) = \log \left(\sum_{j=1}^n w_j e^{-a_{ij} + Y_{ij} - \mu(z-\mu)(1-\rho_{ij})} \right), \tag{A.10}$$

$$\text{where } a_{ij} = \mu^2(1 - \rho_{ij}) \text{ and } Y_{ij} = \mu(1 - \rho_{ij}^2)^{1/2} X_{ij}.$$

Let Ω_i be defined in (A.8) with $\epsilon > 0$ satisfying

$$\epsilon \rightarrow 0 \text{ with } \epsilon\mu \rightarrow \infty \text{ and } n(\epsilon\mu)^{1/2} e^{-\epsilon\mu/2} = o(\mu^{-1}) \text{ as } \mu \rightarrow \infty.$$

LEMMA 2. For fixed \mathbf{X}_i , W_i is monotone decreasing and for any $\beta > 0$,

$$\kappa_{\beta, \mu} = \sup_{z \geq \mu - \epsilon, \mathbf{X}_i \in \Omega_i \text{ for all } 1 \leq i \leq n} \{W_i(z) - W_i(z + \beta\mu^{-1})\} \rightarrow 0 \text{ as } \mu \rightarrow \infty.$$

PROOF. The monotone decreasing property of W_i follows directly from (A.10). Let γ be positive constants satisfying $\gamma \rightarrow 0$ and $c^{-1}|\log w^*| = o(\gamma)$ as $\mu \rightarrow \infty$. Then under Ω_i ,

$$\begin{aligned} & e^{W_i(z)} - e^{W_i(z + \beta\mu^{-1})} \\ & \leq e^{-\gamma\mu(\mu - \epsilon)/\beta + \epsilon^{1/2}\mu^{3/2}} + (1 - e^{-\gamma}) \sum_{j=1}^n w_j e^{-a_{ij} + Y_{ij} - \mu(z - \mu)(1 - \rho_{ij})} \mathbf{1}_{\{1 - \rho_{ij} \leq \gamma/\beta\}} \\ & \leq o(w^*) + (1 - e^{-\gamma})e^{W_i(z)}. \end{aligned}$$

By (A.10), $e^{W_i(z)} \geq w_i \geq w^*$ and hence Lemma 2 holds. \square

LEMMA 3. $P\{B \geq c\} \sim (2c)^{-1}(\pi \log c)^{-1/2}$ as $c \rightarrow \infty$.

PROOF. Let E_i denote expectation with respect to Q_i . Since $\log n = o(\mu)$, there exists $\gamma \rightarrow 0$ such that $\sum_{i: w_i \geq e^{-\gamma\mu}} w_i \rightarrow 1$ as $\mu \rightarrow \infty$. Let $L_i = \mu Z_i - \mu^2/2$ and $W_i = W_i(Z_i)$. By (A.7),

$$P\{B \geq c\} = \sum_{i=1}^n w_i E_i(e^{-(L_i + W_i)} \mathbf{1}_{\{L_i + W_i \geq \log c\}}). \quad (\text{A.11})$$

Under Q_i , $L_i \sim N(\mu^2/2, \mu^2)$ has normal density f_μ satisfying

$$\begin{aligned} \sup_{y \in \mathbf{R}} f_\mu(y) &= (2\pi\mu^2)^{-1/2}, \\ f_\mu(y) &\sim (2\pi\mu^2)^{-1/2} \text{ uniformly over } |y - \mu^2/2| \leq \xi \text{ for some } \xi = o(\mu). \end{aligned} \quad (\text{A.12})$$

If $\mathbf{X}_i \in \Omega_i$ and $Z_i < \mu - \epsilon$, then by Cauchy's inequality, for all $j \neq i$,

$$Z_j \leq \sup_{|\rho| \leq 1} \{\rho(\mu - \epsilon) + (1 - \rho^2)^{1/2}(\epsilon\mu)^{1/2}\} = \{(\mu - \epsilon)^2 + \epsilon\mu\}^{1/2} < \mu - \eta\mu^{-1},$$

for some $\eta \rightarrow \infty$ as $\mu \rightarrow \infty$ and it follows that $B < c$ for all large μ . If $\mathbf{X}_i \in \Omega_i$ and $Z_i \geq \mu - \epsilon$, then

$$-a_{ij} + Y_{ij} - b(Z_i - b)(1 - \rho_{ij}) \leq -\mu(\mu - \epsilon) + \sup_{|\rho| \leq 1} \{\rho\mu(\mu - \epsilon) + (1 - \rho^2)^{1/2}(\epsilon\mu^3)^{1/2}\}$$

$$= -\mu(\mu - \epsilon) + \{\mu^2(\mu - \epsilon)^2 + \epsilon\mu^3\}^{1/2} \leq \epsilon\mu,$$

and by (A.10), $\log w_i \leq W_i(Z_i) \leq \epsilon\mu$.

Let $g_i(z) = \mu z - \mu^2 + W_i(z)$ and let z_μ be the smallest root of $g_i(z) = \log c + \kappa_{\beta,\mu}$. By Lemma 2 and (A.12), if $|\log w_i| \leq \gamma\mu$ and $\mathbf{X}_i \in \Omega_i$, then for any $\beta > 0$,

$$\begin{aligned} E_i(e^{-(L_i+W_i)} \mathbf{1}_{\{L_i+W_i \geq \log c\}} | \mathbf{X}_i) &\geq E_i\left\{e^{-W_i(z_\mu)} \int_{\mu z_\mu - \mu^2}^{\mu z_\mu - \mu^2 + \beta} e^{-y} f_\mu(y) dy\right\} \\ &\sim (2\pi\mu^2)^{-1/2} c^{-1} e^{-\kappa_{\beta,\mu}} (1 - e^{-\beta}) \text{ as } c \rightarrow \infty. \end{aligned} \quad (\text{A.13})$$

Let \tilde{z}_μ be the smallest root of $g_i(z) = \log c$. Then

$$\begin{aligned} &E_i(e^{-(L_i+W_i)} \mathbf{1}_{\{L_i+W_i \geq \log c\}}) \\ &\leq \sum_{i=0}^{\infty} E_i\left\{e^{-W_i(\tilde{z}_\mu + j\mu^{-1})} \int_{\mu\tilde{z}_\mu - \mu^2 + j}^{\mu\tilde{z}_\mu - \mu^2 + j + 1} e^{-y} f_\mu(y) dy\right\} \\ &\sim (2\pi\mu^2)^{-1/2} c^{-1} \text{ as } c \rightarrow \infty. \end{aligned} \quad (\text{A.14})$$

By (A.11) and (A.13)–(A.14) with $\beta \rightarrow \infty$, and noting that by (A.2), $E_i(B^{-1} \mathbf{1}_{\{B \geq c\}} \cap \Omega_i^c) \leq c^{-1}[1 - Q_i(\Omega_i)] = o(c^{-1}(\log c)^{-1/2})$, Lemma 3 holds. \square

PROOF OF LEMMA 1'. By Lemma 3 and (A.3), to show (A.1), it suffices to show that

$$\sup_{\mu \leq M < \mu + \epsilon} |A^{(1)} - B| \rightarrow 0 \text{ as } \mu \rightarrow \infty. \quad (\text{A.15})$$

Define $J = \{j : \mu - \epsilon \leq Z_j \leq \mu + \epsilon\}$. Then

$$e^{Z_{j+}^2/2} = e^{\mu Z_j - \mu^2/2 + o(1)} \text{ uniformly over } j \in J \text{ and } \sum_{j \notin J} w_j e^{Z_{j+}^2/2} = o(e^{\mu^2/2}).$$

Since $A^{(1)} \geq B$ for all $\mu \geq 0$, (A.15) holds for $k = 1$.

To show Lemma 1 for $k = 2$, approximate $A^{(2)}/2$ by

$$B = \sum_{j=1}^n w_j (e^{Z_{j+}^2/2} + e^{Z_{j-}^2/2})/2,$$

where $Z_{j-} = \max\{-Z_j, 0\}$, and apply (A.1) on the $2n$ random variables $\pm Z_1, \dots, \pm Z_n$. \square

APPENDIX B: PROOF OF THEOREM 2

We shall prove Theorem 2 for $k = 1$, then indicate how it can be extended to $k = 2$.

LEMMA 4. *Let $\epsilon > 0$. Then*

$$P\{A_i^{(1)} \geq c\} \leq \frac{1}{2c\sqrt{\pi \log c}} + \frac{\{(2\epsilon)^{-1}\sqrt{2 \log(cn_i)} + 2\}e^{\epsilon^2/2}}{c}. \quad (\text{B.1})$$

PROOF. For notational simplicity, we shall omit the subscripts i in $A_i^{(1)}$, n_i and Z_{ij} . Let $\mu = \sqrt{2 \log(cn)}$ and $M = \max_{1 \leq j \leq n} Z_j$. Then

$$P\{M \geq \mu\} \leq nP\{|Z_1| \geq \mu\} \leq \frac{n}{\mu\sqrt{2\pi}}e^{-\mu^2/2} \leq \frac{1}{2c\sqrt{\pi \log c}}. \quad (\text{B.2})$$

Let $\ell_0 = \lfloor (\mu + \epsilon)/(2\epsilon) \rfloor + 1$, where $\lfloor \cdot \rfloor$ denotes the greatest integer function. Let $\mu_\ell = \mu - (2\ell - 1)\epsilon$ for $1 \leq \ell \leq \ell_0 - 1$ and $\mu_{\ell_0} = 0$. Define

$$Y = (n\ell_0)^{-1} \sum_{j=1}^n \sum_{\ell=1}^{\ell_0} e^{\mu_\ell Z_j - \mu_\ell^2/2}. \quad (\text{B.3})$$

If $M < \mu$, then for all j , there exists $\ell (= \ell_j)$ such that $|Z_{j+} - \mu_\ell| \leq \epsilon$. If $Z_{j+} = 0$, select in particular $\ell = \ell_0$. Then

$$\frac{e^{Z_{j+}^2/2}}{e^{\mu_\ell Z_j - \mu_\ell^2/2}} = e^{(Z_{j+} - \mu_\ell)^2/2} \leq e^{\epsilon^2/2}. \quad (\text{B.4})$$

By (2.1), (B.3) and (B.4), $A^{(1)} \leq \ell_0 e^{\epsilon^2/2} Y$ and hence

$$P\{A^{(1)} \geq c\} \leq P\{M \geq \mu\} + P\{Y \geq \ell_0^{-1} e^{-\epsilon^2/2} c\}. \quad (\text{B.5})$$

Since there exists a probability measure Q satisfying $(dQ/dP)(\mathbf{Z}) = Y$, it follows from a change of measure to Q that

$$P\{Y \geq \ell_0^{-1} e^{-\epsilon^2/2} c\} \leq \ell_0 e^{\epsilon^2/2} c^{-1},$$

and (B.1) follows from (B.2) and (B.5). \square

Observe that $|\log(\alpha r/m)|^{1/2}$ increases as r decreases from m to 1. If

$$\log(n^* + 1) = o(|\log \alpha|^{1/2}), \quad (\text{B.6})$$

select $r = m$. Otherwise, since $\log(n^* + 1) = o(|\log(\alpha/m)|^{1/2})$, by considering a subsequence if necessary, we can assume that there exists positive integers $r(\leq m) \rightarrow \infty$ as $m \rightarrow \infty$ such that

$$|\log(\alpha r/m)|^{\frac{1-\eta}{2}} \leq \log(n^* + 1) = o(|\log(\alpha r/m)|^{1/2}) \text{ as } m \rightarrow \infty, \quad (\text{B.7})$$

where $\eta > 0$ is given in the statement of Theorem 2. This is possible because the ratio of $|\log(\alpha r/m)|^{1/2}$ and the lower bound in (B.7) tends to infinity as $m \rightarrow \infty$. Theorem 2 then follows from Lemmas 5 and 6 below, under the assumptions of Theorem 2.

LEMMA 5. $\alpha_m = P\{p_{(i)} \leq \alpha i/m \text{ for some } 1 \leq i \leq r\} \sim \alpha \text{ as } m \rightarrow \infty$.

PROOF. Let $\delta > 0$. By Lemma 3 and either (B.6) or (B.7), whichever is appropriate, we can find m large enough such that

$$\frac{(1-\delta)\alpha i}{m} \leq P\left\{p_j \leq \frac{\alpha i}{m}\right\} \leq \frac{(1+\delta)\alpha i}{m} \text{ for all } 1 \leq i \leq r \text{ and } 1 \leq j \leq m. \quad (\text{B.8})$$

By (B.8) and either Seeger (1968) or Simes (1986),

$$\alpha_m \leq (1+\delta)\alpha. \quad (\text{B.9})$$

Let q_j be the actual p-value of $A_j^{(1)}$ and $q_{(1)} \leq \dots \leq q_{(m)}$ the sorted values of q_1, \dots, q_m .

By (B.8), $q_j \geq (1-\delta)p_j$ whenever $p_j \leq \alpha r/m$. Hence

$$\alpha_m \geq \gamma_m = P\{q_{(i)} \leq (1-\delta)\alpha i/m \text{ for some } 1 \leq i \leq r\}. \quad (\text{B.10})$$

Let $t = (1-\delta)\alpha r/m$. Since

$$P\{q_{(i)} \leq it/r \text{ for some } 1 \leq i \leq r | q_{(s)} \leq t < q_{(s+1)}\} = 1 - (s/r) \text{ for } s = 1, \dots, r,$$

see Seeger (1968) p.589,

$$\begin{aligned}
\gamma_m &\geq \sum_{s=1}^r \binom{s}{r} \binom{m}{s} t^s (1-t)^{m-s} + \sum_{s=r+1}^m \binom{m}{s} t^s (1-t)^{m-s} \\
&= (1-\delta)\alpha \sum_{s=1}^m \binom{m-1}{s-1} t^{s-1} (1-t)^{m-s} - \sum_{s=1}^r \binom{s-r}{m} t^{s-1} (1-t)^{m-s} \\
&= (1-\delta)\alpha - m^{-1} E(X-r)^+,
\end{aligned} \tag{B.11}$$

where X is Binomial($m, (1-\delta)\alpha/m$). But $(x-r)^+ \leq [x - (1-\delta)\alpha]^2/r$ for all r large uniformly over $x \geq 0$ (let $x = r + u$ for some $u \geq 0$ and expand the right-hand side). Hence

$$m^{-1} E(X-r)^+ \leq (mr)^{-1} E[X - (1-\delta)\alpha]^2 \leq r^{-1}(1-\delta)\alpha. \tag{B.12}$$

Lemma 5 follows from (B.9)–(B.12) by choosing δ arbitrarily small. \square

LEMMA 6. $\sum_{i=r+1}^m P\{p_{(i)} \leq \alpha i/m\} = o(\alpha)$, (with convention $\sum_{i=m+1}^m = 0$).

PROOF. Assume without loss of generality $r < m$ and consider $r < i \leq m$. Let $p^* = \alpha i/m$ and select c to satisfy

$$\frac{1}{2c\sqrt{\pi \log c}} = p^*.$$

Then for all large m ,

$$\log(n^* + 1) \geq |\log(\alpha r/m)|^{\frac{1-\eta}{2}} [\geq |\log p^*|^{\frac{1-\eta}{2}} \sim (\log c)^{\frac{1-\eta}{2}}], \tag{B.13}$$

see (B.7). Since $p_j = (2A_j)^{-1}(\pi \log A_j)^{-1/2}$, by Lemma 4 with $\epsilon = 1$,

$$P\{p_j \leq p^*\} = P\{A_j^{(1)} \geq c\} \leq p^*[1 + (2\pi e \log c)^{1/2}(\log c + \log n^*)^{1/2} + 4(\pi e \log c)^{1/2}]. \tag{B.14}$$

By (B.13) and the assumptions in Theorem 2,

$$(\log c)^{1/2} \leq (\log c)^{1/2}(\log c + \log n^*)^{1/2} = O((\log(n^* + 1))^{\frac{2}{1-\eta}}) = o(\alpha^{-1+\kappa}) \text{ for some } \kappa > 0. \tag{B.15}$$

By (B.14), (B.15), recalling that $p^* = \alpha i/m$,

$$P\{p_{(i)} \leq p^*\} \leq P\{X \geq i\} \text{ where } X \sim \text{Binomial}(m, p_X), \tag{B.16}$$

and $p_X = p^* \alpha^{-1+\kappa} = i\alpha^\kappa/m$. It follows from a change of measure argument that (with the convention $0^0 = 1$),

$$P\{X \geq i\} \leq \frac{p_X^i (1 - p_X)^{m-i}}{(i/m)^i (1 - i/m)^{m-i}} \leq \alpha^{i\kappa} \left(1 + \frac{i}{m-i}\right)^{m-i} \leq (\alpha^\kappa e)^i, \quad (\text{B.17})$$

and Lemma 6 follows from (B.16) and (B.17) since $r \rightarrow \infty$ as $m \rightarrow \infty$. \square

To show Theorem 2 for $k = 2$, we note the inequality

$$A_i^{(2)} \leq A_i^{(1)} + \tilde{A}_i^{(1)}, \text{ where } \tilde{A}_i^{(1)} = n_i^{-1} \sum_{j=1}^{n_i} e^{Z_{ij}^2/2}.$$

Then $P\{A_i^{(2)} \geq c\} \leq P\{A_i^{(1)} \geq c/2\} + P\{\tilde{A}_i^{(1)} \geq c/2\}$ and Lemmas 5, 6 and Theorem 2 follows from applying Lemma 4 on both $A_i^{(1)}$ and $\tilde{A}_i^{(1)}$.

References

- [1] Bartek, J., and Lukas, J. (2001). Pathways governing G1/S transition and their response to DNA damage. *FEBS Lett.*, 490, 117-122.
- [2] Begun, Hall, Huang and Wellner (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.*, 11, 432-452.
- [3] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a powerful and practical approach to multiple testing. *J. Roy. Statist. Ser B*, 57, 289-300.
- [4] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, 29, 1165-1188.
- [5] Chan, H.P. (2009). Detection of spatial clustering with average likelihood ratio test statistics, accepted for publication by *Ann. Statist.*

- [6] Chan, H.P., Tu, I. and Zhang, N.R. (2009). Boundary crossing probability computations in the analysis of scan statistics, In *Scan Statistics: Methods and Applications*, (Ed. Glaz, J., Pozdnyakov, V. and Wallenstein, S.), 87–111, Springer, New York.
- [7] Dahl, F., Stenberg, J., Fredriksson, S., Welch, K., Zhang, M., Nilsson, M., Bicknell, D., Bodmer, W.F., Davis, R.W., and Ji, H. (2007). Multigene amplication and massively parallel sequencing for cancer mutation discovery. *Proc. Natl. Acad. Sci.*, 104, 9387–9392.
- [8] Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, 32, 962–994.
- [9] Efron, B. and Tibishirani, R. (2007). On testing the significance of sets of genes, *Ann. Appl. Statist.*, 1, 107–129.
- [10] Eklund, G. (1963). Massignifikansproblemet. Unpublished seminar papers, Uppsala Univ. Institute of Statistics.
- [11] Fisher, R.A. (1932). *Statistical Methods for Research Workers*, 4th edition. Edinburgh: Oliver and Boyd.
- [12] Gangnon, R. and Clayton, M. (2001). A weighted average likelihood ratio test for spatial clustering of disease. *Statist. in Med.*, 20, 2977–2987.
- [13] Hajek, J. (1972). Local asymptotic minimax and admissibility in estimation. *Proc. Sixth Berkeley Symp. Math. Statist. Probab.*, 1, 175–194. Univ. of California Press, Berkeley.
- [14] Lai, T.L. and Siegmund, D. (1977). A non-linear renewal theory with applications to sequential analysis I. *Ann. Statist.*, 5, 946–954.
- [15] LeCam, L. (1972). Limits of experiments. *Proc. Sixth Berkeley Symp. Math. Statist. Probab.*, 1, 245–261. Univ. of California Press, Berkeley.

- [16] Momand, J., et al. (2001). MDM2—master regulator of the p53 tumor suppressor protein. *Gene*, 242, 15-29.
- [17] Murphy, S. and van der Vaart, A.W. (2000). On profile likelihood. *J. Amer. Statist. Assoc.*, 95, 449–465.
- [18] Needleman, H., Gunnoe, C., Leviton, A., Reed, R. Presie, H., Maher, C. and Barrett, P. (1979). Deficits in psychologic and classroom performance of children with elevated denture lead levels. *New England J. Medicine*, 300, 689–695.
- [19] Newton, M., Quintana, F., Den Boon, J., Sengupta, S. and Ahlquist, P. (2006). Random set methods identify distinct aspects of the enrichment signal in gene-set analysis. *Technical report, Dept. Statistics, Univ. of Wisconsin, Madison.*
- [20] Nobel, A.B. and Write, F.A. (2005). Significance analysis of functional categories in gene expression studies: A structured permutation approach. *Bioinformatics* 19:1433-1440.
- [21] Pavlidis, P., Lewis, D. and Noble, W. (2002). Exploring gene expression data with class scores. *Pac. Symp. Biocomputing* 474-485.
- [22] Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.
- [23] Pop, M. and Salzberg, S.L. Bioinformatics challenges of new sequencing technology. *Trends in Genetics*, 24, 142–149.
- [24] Rahnenfhrer, J., Domingues, F. S., Maydt, J. and Lengauer, T. (2004). Calculating the statistical significance of changes in pathway activity from gene expression data. *Stat. Appl. Genet. Mol. Biol.* 3, 131.
- [25] Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, 273, 1516-7.

- [26] Seeger, P. (1968). A note on a method for the analysis of significances en masse. *Technometrics*, 10, 586–593.
- [27] Shendure, J., Mitra, R.D., Varma, C., and Church, G.M. (2004). Advanced sequencing technologies: methods and goals. *Nature Reviews Genetics*, 5, 335–344.
- [28] Shorack, G.R. and Wellner, J.A. (1986). *Empirical Processes with Applications to Statistics*, Wiley, New York.
- [29] Siegmund, D. (2001). Is peak height sufficient? *Genetic Epidemiology*, 20, 403–408.
- [30] Siegmund, D. and Yakir, B. (2003). Significance level in interval mapping. In: Zhang, H. Huang, J. (ed) *Development of modern statistics and related topics in celebration of Yaoting Zhang's 70th birthday*. World Scientific, Singapore.
- [31] Siegmund, D. and Yakir, B. (2007). *The Statistics of Gene Mapping*, Springer, New York.
- [32] Simes (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73, 751–754.
- [33] Storey, J. (2002). A direct approach to false discovery rates. *J. Roy. Statist. Ser. B*, 64, 479–498.
- [34] Storey, J. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann. Statist.*, 31, 2013–2035.
- [35] Storey, J., Taylor, J. and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. Roy. Statist. Ser. B*, 66, 187–205.
- [36] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovitch, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. and Mesirov, J. P. (2005).

Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* 102, 15545-15550.

[37] Westfall, P.H. and Young, S.S. (1993). *Resampling Based Multiple Testing*, Wiley, New York.

[38] Woodroffe, M. (1976). A renewal theorem for curved boundaries and moments of first passage times. *Ann. Probab.*, 4, 67–80.

[39] Zhang, N.R., *et al.* (2009). Statistics for point mutation detection using high throughput sequencing of matched tumor and normal pairs. *In preparation.*