Clustering of Translocation Breakpoints
Author(s): Mark R. Segal and Joseph L. Wiemels
Source: *Journal of the American Statistical Association*, Vol. 97, No. 457 (Mar., 2002), pp. 66-76
Published by: American Statistical Association
Stable URL: http://www.jstor.org/stable/3085759
Accessed: 27/04/2010 13:06

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/page/info/about/policies/terms.jsp. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/action/showPublisher?publisherCode=astata.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

American Statistical Association is collaborating with JSTOR to digitize, preserve and extend access to *Journal of the American Statistical Association.*

http://www.jstor.org

Clustering of Translocation Breakpoints

Mark R. SEGAL and Joseph L. WIEMELS

Translocation, a physical movement of genetic material from one chromosome to another, can result in the aberrant linkage of two cellular genes. This type of fusion may disrupt cellular function by producing novel, biologically active fused genes, or by triggering the activation of normally quiescent growth-associated genes. Either of these mechanisms provides a putative oncogenic stimulus, and indeed, several gene fusions from translocations have been identified in leukemias, lymphomas, and sarcomas. Although the oncogenic effects of genes involved in translocations have been under intensive study, little is known regarding the formation of translocation fusions themselves. The locations of these fusions are typically independent of the resultant oncogenic protein because they usually arise within certain bounded noncoding regions of the genes. Thus the resultant proteins can be ignored in studying translocations, and we can focus exclusively on the fusions. A patterned (in particular, clustered) distribution of fusion breakpoints will potentially yield relevant information about the fusion process by identifying regions prone to recombination. Accordingly, the statistical analysis of translocation breakpoints has focused on the extent to which they cluster. Somewhat questionable methods have been used in this regard. After highlighting these shortcomings, we introduce a variety of approaches, including scan statistics, bandwidth tests, and gap statistics, that provide a comprehensive means for appraising clustering. We apply this battery to TEL–AML1 translocations, the most common translocation in childhood acute lymphoblastic leukemia. The results obtained indicate generally weaker evidence for clustering than previously reported, and also highlight differences between the statistical approaches.

KEY WORDS: Bandwidth test; Gap statistic; Gene fusion; Scan statistic.

1. INTRODUCTION

Translocation is defined as the physical movement of genetic material between two nonhomologous chromosomes. In the simplest case, formation of a translocation involves double-strand breaks on two chromosomes followed by the aberrant fusion of the DNA free ends to the wrong partner chromosome. The resulting two derivative chromosomes with swapped arms can be viewed on a glass slide preparation of chromosomes, or a karyotype, of a patient's cells. At the level of the DNA sequence, specific genes may be split in two, resulting in the fusion of two genes not normally associated with each other. This resultant juxtaposition of two cellular genes can generate chimeric protein products in which the functional domains of two separate genes are fused together and/or can alter regulation of gene expression (Rabbitts 1994). More than 600 different acquired translocations in the neoplastic diseases have been described. A given translocation between two particular cellular genes is consistently associated with a specific tumor type. This permits the development of diagnostics and therapeutics based on particular gene fusion products.

Translocations in the leukemias, our focus here, usually result in the formation of a chimeric protein in which the proximal end of one protein is fused to the terminal end of another protein. These proteins are usually transcription factors—proteins in the nucleus that control the expression of other genes involved in the growth and development of blood cells. When the normal development of blood stem cells is disrupted by the aberrant fusion transcription factor, leukemia may result. Genes are structured with protein-coding regions, or exons, interspersed with noncoding regions, or introns. Translocations that produce chimeric oncoproteins are constrained to occur within specific introns to preserve the ordering of exons necessary to generate an oncoprotein. However, within susceptible introns there is great latitude as to where the DNA may be broken and refused on either chromosome. This breakage/refusion site is called a "breakpoint" and is unique to each individual patient diagnosed with a particular translocation.

One of the most common translocations in leukemia is the fusion of the TEL gene on chromosome 12 to the AML1 gene on chromosome 21. This translocation occurs in 25% of cases of childhood acute lymphoblastic leukemia, the most common cancer of childhood. We have shown that the TEL–AML1 fusion occurs prenatally in most children who develop this form of leukemia, even up to age 14 (Wiemels et al. 1999a; Wiemels, Ford, Van Wering, Postma, and Greaves 1999b). Despite this knowledge of the temporal origin of the translocation, little is understood about the process of fusion formation. Considered a "master" transcription factor, AML1 is a critical regulator of the development of nearly all blood cells. Blood cells develop from embryonic precursor cells, or stem cells, into functional types, such as red blood cells, T cells, and B cells. The TEL–AML1 protein is thought to result in the aberrant repression of genes that are normally induced by AML1 during the process of differentiation, or development of blood stem cells into functional types (Guidez et al. 2000). With the process of differentiation "frozen," the blood stem cells may gain a form of immortality, one component of the leukemic cell phenotype. The fusion occurs within the 14,000 base pair (bp) intron 5 of TEL and large 160,000 bp intron 1–2 of AML1. Both TEL and AML1 are involved in a various other translocations in other lymphoid and myeloid leukemia subtypes in children and adults (Greaves 1999), making the study of translocations involving these genes applicable to a wide range of the disease.

The elucidation of some common translocation breakpoint sequences in the lymphomas has resulted in a clear causal mechanism. Very tight clustering has been observed, which implicates the involvement of "recombination site sequences"

Mark R. Segal is Professor and Joseph L. Wiemels is Assistant Professor, Department of Epidemiology and Biostatistics, University of California, San Francisco, CA 94143 (E-mail: *mark@biostat.ucsf.edu*). The authors thank Catherine Loader and Michael Minnotte for providing software. Chuck McCulloch, Tom Louis, a referee and associate editor provided many helpful suggestions. This work was supported by National Institutes of Health grant AI40906.

© 2002 American Statistical Association

(RSSs) in the formation of such translocations (Jager et al. 2000; Tsujimoto, Gorham, Cossman, Jaffe, and Croce 1985). These are short 16-bp motifs whose orientation allows them to be recognized by select cellular enzymes. These enzymes normally rearrange genes of the immune system to produce the antibody repertoire. This gene rearrangement process is critical for formation of the estimated 10^7 different antibodies necessary for immune system function. However, the aberrant recognition of RSS in other cellular genes can have the unfortunate consequence of producing translocations. The fact that the cells from which lymphomas originate normally express these same enzymes serves to implicate RSS in the genesis of translocation.

The situation for the leukemias is different in that breakpoint distributions are seemingly more diffuse, resulting in a poor understanding of etiology. Recombination site sequences are not involved in leukemia translocations. This is because the translocations occur at a very early progenitor stage in blood cell development, preceding the expression of the enzymes that manipulate RSS. Only recently have methods been developed to sequence these leukemia fusions (Reichel et al. 1999; Thandla et al. 1999; Wiemels and Greaves 1999). Several hypotheses for explaining leukemia translocations based on these breakpoints have been advanced. These involve either (a) the primary base sequence of DNA (e.g., motifs for enzyme binding sites), (b) secondary structures (paranemic, or base-unpaired DNA structures), or (c) tertiary organization of chromatin. So far, however, little consensual evidence supports a particular mechanism, with some studies implicating primary or secondary structures (e.g., Wyatt, Rudders, Zelenetz, Delellis, Krontiris 1992; Wiemels and Greaves 1999) and others implicating tertiary structures (Khodarev, Sokolova, Vaughan 1999; Strissel, Strick, Rowley, and Zeleznik-Le 1998).

Establishing of breakpoint clusters in particular regions would imply that specific features (i.e., motifs) of the primary DNA sequence near the location of a cluster may play a role in translocation. Attempts to search for several hypothesized breakpoint motifs have been undertaken for TEL–AML1 and other leukemia translocation fusions. These attempts revealed no hint of association between breakpoint location and these features of primary DNA structure. Although the negative findings by no means preclude a role for some as-yet uncharacterized motif, justification of the labor-intensive nature of a comprehensive motif-based analyses would require some earlier evidence of clustering. So appraising clustering is an expedient preliminary step to undertaking motif search. Accordingly, to the extent that the location of translocation breakpoints has been subject to any statistical treatment, the analyses have focused on evaluating and localizing putative clusters.

This article was partially motivated by shortcomings identified in the limited approaches to appraising the clustering taken to date. Section 2 reviews these approaches and describes various improvements, drawing on recent statistical work. These methods include scan statistics with attendant distributional approximations, Silverman's (1981) bandwidth test procedure, and gap statistics (Tibshirani, Walther, and Hastie 2001). As illustrated, these methods differ according to whether the emphasis is on verifying a specific cluster or on determining the number of clusters. Section 3 presents a reanalysis of the particular TEL–AML1 fusion data described earlier, and Section 4 describes some possible extensions and offers a concluding discussion.

2. APPROACHES TO APPRAISING CLUSTERING

To date, very little formal assessment of clustering has been undertaken in evaluating translocation breakpoint distributions. Indeed, van der Reijden et al. (1999) asserted (in the title of their article itself!) that acute myeloid leukemia–associated inv(16)(p13q22) breakpoints are tightly clustered without reporting any supportive analysis. The only formal approach to date is that of Wiemels et al. (2000), and we focus on their data and methods. A preview is provided by Figures 1 and 2.

A total of 24 patients were studied, deriving from collaborations spanning a population-based childhood epidemiology study in the United Kingdom (18 patients) and clinics in Buffalo, New York (4 patients) and Valhalla, New York (2 patients). Subject to DNA availability, all potential patients were included. The DNA availability was specific to the hospital of diagnosis and treating physician and is judged to be independent of the pathologic or biologic characteristics of the leukemia. Consequently, the selection process should affect neither the breakpoint distribution nor the likelihood of detecting clustering.

All of the patients were identified to have the TEL–AML1 translocation using reverse-transcriptase polymerase chain reaction (RT-PCR), which specifically targets the TEL exon 5–AML1 exon 2 mRNA. No other mRNA structures have been reported in the leukemias. However, the RT-PCR assay does not determine the patient-specific DNA breakpoint. Only patients with sufficient high molecular weight diagnostic DNA available, as well as additional archived material, can be used for further analysis. The genomic DNA fusions (which provide the patient-specific breakpoints) were identified using molecular biology methods that are precise (error free) in locating the breakpoint. For 20 subjects, breakpoints were sequenced using long-distance inverse polymerase chain reaction methods (Wiemels and Greaves 1999); those for the remaining four patients (from Buffalo) were sequenced using traditional techniques with cloning vectors (Thandla et al. 1999).

The breakpoint data are displayed in Figure 1. The shaded boxes represent exons of the respective genes, with the breakpoints occurring primarily in the intervening introns. The scale is in base pairs; note the much greater range for AML1 than for TEL. Each numeral above the arrow showing breakpoint location is a patient identifier; for each of the 24 patients, the location of breakpoints for both derived chromosomes being displayed. Note that breakpoint data are in fact paired; each patient has breakpoints within both the TEL and AML1 intronic regions. Corresponding bivariate clustering approaches are addressed in Section 2.3.

Figure 2, from Wiemels et al. (2000), depicts breakpoint density estimates using (Gaussian) kernel density estimation with prescribed bandwidths. The bandwidths used are 1000 bp for TEL and 2000 bp for AML1. Later, we show that these

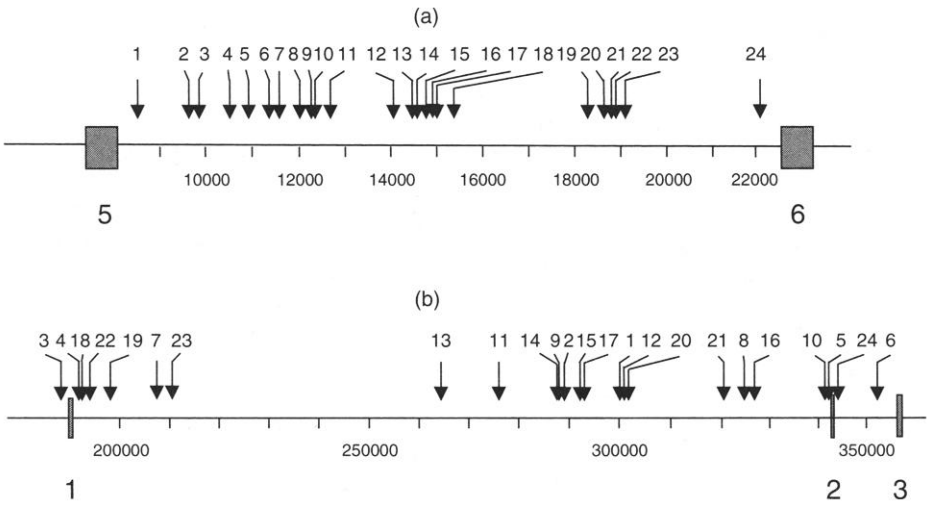

Figure 1. Breakpoint Locations Within the (a) TEL and (b) AML1 Genes. The shaded boxes represent exons of the respective genes. In both panels the scale is in base pairs. The data are paired with the numerals above each arrow showing breakpoint location as a patient identifier.

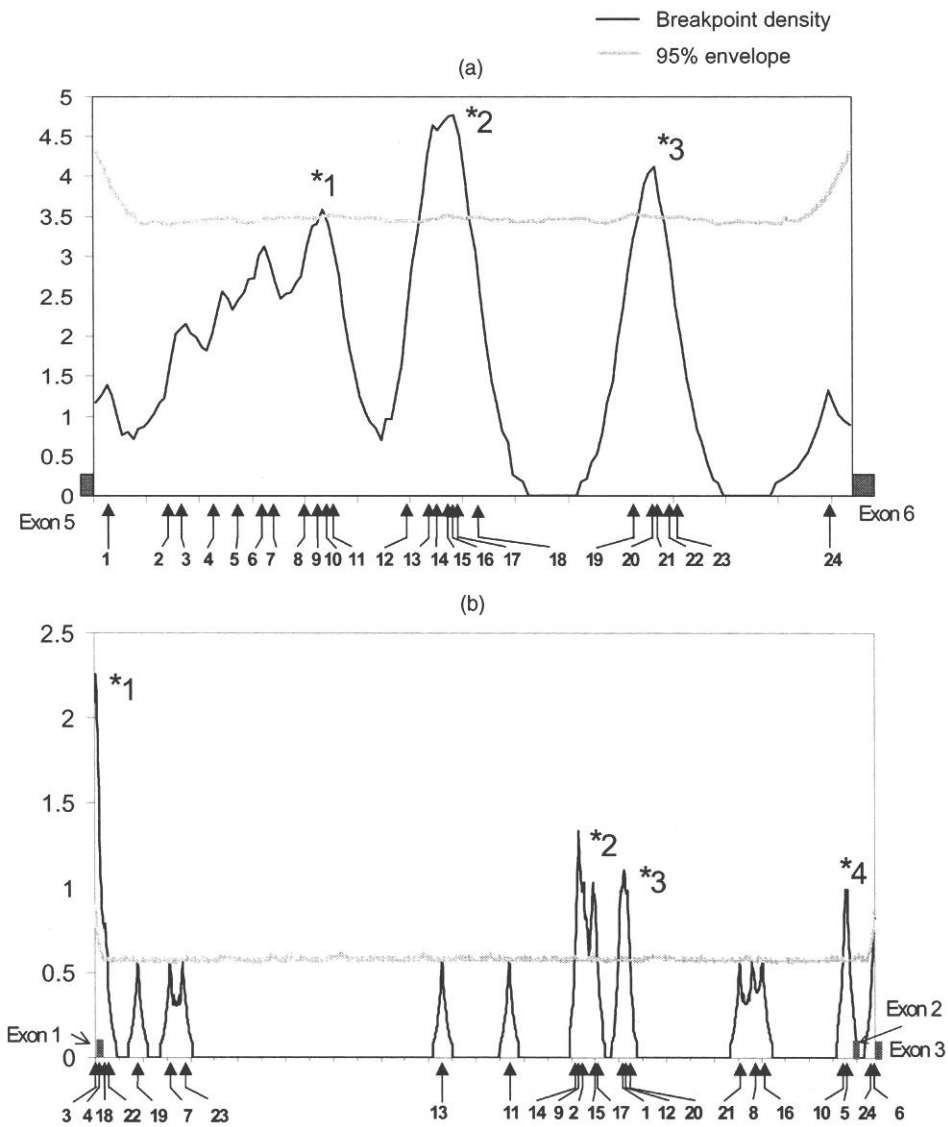

Figure 2. Breakpoint Locations With Corresponding Gaussian Kernel Density Estimates (——) and 95% Pointwise Confidence Envelopes (——) for (a) TEL and (b) AML1. Starred numerals designate the putative clusters reevaluated in Table 1.

are much too small. Regions were the kernel density estimate exceeds a 95% confidence envelope obtained via simulation (described in Sec. 2.2) are designated as clusters, this process yielding the three (four) numbered clusters for TEL (AML1) that we reevaluate via scan statistic approximations, as described next.

2.1 Existence: Nearest-Neighbor and Scan Statistics

Wiemels et al. (2000) used k nearest-neighbor *(kNN) distances averaged over all breakpoints* to establish the existence of clustering and, subsequently, kernel density estimation to localize the clusters (regarded as equivalent to modes). Here k designates the number of nearest neighbors considered. We first focus on *kNN* distances, then discuss density estimation approaches in Section 2.2.1. Cuzick and Edwards (1990) provided a motivation for average *kNN* distances in a case-control context; however, the rationale does not extend to the current setting, as the following example shows. Consider a situation where we have $c-1$ tightly clustered points and one outlying point well separated from the cluster. Now consider an alternate configuration with c points equispaced on an interval of a length equal to the distance between the cluster and the outlier. These two arrangements will have essentially the same average first nearest-neighbor distance despite being diametrically opposite in terms of the extent of clustering. The salient feature of this example is that the use of average (global) nearest-neighbor distances can be insensitive to the presence of clustering because of the influence of (a few) isolated points. In contrast, the use of *minimum kNN* distances is not so affected. Indeed, the use of the *scan statistics*, which is equivalent to the minimum *kNN* distance, is well established for assessing clustering and has been applied in many settings (see, e.g., Wallenstein and Neff 1987; Karlin and Macken 1991). Although minimum *kNN* distances are distributionally less tractable than average *kNN* distances, accurate, computationally feasible approximations are available. We next outline two such approximations, which we illustrate in Section 3.

Without loss of generality, for the purposes of clustering, we can rescale the intronic region where breakpoints arise to the unit interval $(0, 1)$. Let $X_1, X_2, \ldots, X_n$ be independent and identically drawn from $\mathcal{U}(0, 1)$, the uniform distribution on the unit interval, with $X_{(i)}$ the corresponding order statistics. Let $N_{x,x+d} = \#\{X_i : X_i \in (x, x+d)\}$ be the number of points contained in the interval $(x, x+d)$. Then the scan statistic for prescribed interval length d is defined as $N_d = \sup_x N_{x,x+d}$, the maximum number of points in such an interval. If we also define L_k to be the length of smallest subinterval of $(0, 1)$ containing k points, then L_k is the minimum *kNN* statistic, and we have

$$\Pr\{N_d \geq k\} = \Pr\{L_k \leq d\}, \tag{1}$$

so that tests based on the scan and minimum *kNN* statistics are equivalent.

The exact distribution corresponding to (1) is exceedingly complex (see Huntington and Naus 1975) and computationally impractical. This had led to various approximations. Instead of working directly with scan or minimum *kNN* statistics, Huffer and Lin (1997) reformulated in terms of *clumps*. In particular, a $k : d$ clump exists if there are k consecutive points (here translocation breakpoints) in an interval of length d. Let $Y_{k:d} \equiv Y$ be the number of $k : d$ clumps,

$$Y = \sum_{i=1}^{n-k+1} I\{X_{(i+k-1)} - X_{(i)} \leq d\}. \tag{2}$$

Because $Y \geq 1$ if and only if $N_d \geq k$, we have

$$\Pr\{N_d \geq k\} = \Pr\{Y \geq 1\}, \tag{3}$$

so we can effect approximation to the distribution of the scan statistic by approximating $\Pr\{Y \geq 1\}$.

Huffer and Lin (1997) pursued this by finding (in different ways) discrete distributions that match the moments of Y. Here we expand on just one of the simplest approaches, based on Markov chain approximations, which uses only the first two moments of Y. Later we use both this and another approximation based on matching moments to a compound Poisson distribution; the two methods yield very similar results. Explicit formulas for the first two moments of Y are obtained using properties of *spacings*, which are distances between consecutive order statistics. The resultant formulas involve the sample size n, number of points k, interval width d, and cumulative binomial and trinomial probabilities (see Huffer and Lin 1997, sec. 3.2). Although quite general, these formulas do not hold for $k \leq 3$ and $n < 2(k-1)$, a restriction that we address in Section 3.

From (2), we see that Y is defined as sum of $w = n - k + 1$ indicators. The Markov chain approximation is based on the hope that this sequence of indicators behaves like a two-state ($\{0, 1\}$) Markov chain. Consider a two-state Markov chain whose transition matrix $\mathbf{P}$ has off-diagonal entries $p_{01} = a$ and $p_{10} = b$. The stationary distribution for this chain is $\pi_0 = b/(a+b)$ and $\pi_1 = a/(a+b)$. Let $Z_1, Z_2, \ldots$ be a Markov chain started from this stationary distribution with transition matrix $\mathbf{P}$ and define $\widetilde{Y} = \sum_{i=1}^{w} Z_i$. Then we have

$$\Pr\{\widetilde{Y} \geq 1\} = 1 - (1-\pi_1)(1-a)^{w-1}, \tag{4}$$

$$\mathrm{E}\widetilde{Y} = w\pi_1, \tag{5}$$

and

$$\mathrm{Var}(\widetilde{Y}) \approx \pi_1(1-\pi_1) \times \big(w + 2\big(1/(a+b) - 1\big)\big(w - 1/(a+b)\big)\big). \tag{6}$$

Equating (5) and (6) to the first two (central) moments of Y yields closed-form solutions for a and b; recall that $\pi_1 = a/(a+b)$. Substituting these in (4) gives an estimate for $\Pr\{\widetilde{Y} \geq 1\}$. This constitutes the Markov chain approximation for the scan statistic p value in accord with (3). As demonstrated by Huffer and Lin (1997), this approximation is remarkably accurate considering its crudeness. However, their demonstration (by way of simulation) was limited to appreciably larger sample sizes ($n = 100,\ 1000$) than are typically encountered with translocation breakpoint studies. In the present circumstance for TEL–AML1, we have $n = 24$ and $w \leq 21$ (because the approximation is restricted to $k > 3$ breakpoints), so there is less basis for appealing to Markov chain stationarity. Although limited simulations for this sample size

again indicate that the Markov chain (and compound Poisson) approximations are very accurate, to avoid relying solely on moment-based approaches we next consider alternative large-deviation approximations for $\Pr\{N_d \geq k\}$.

Loader (1991) considered both one- and two-dimensional scan statistics and also distinguished between d known and unknown. Here we briefly summarize results for the known d case. Although details of the more complicated unknown d case are deferred to Loader (1991), we do apply the corresponding approximations in Section 3.

The first large-deviation approximation, which is computationally easy and accurate in the upper tail for a range of sample sizes, n, and interval lengths, d, is

$$\Pr\{N_d \geq k\} = n\varepsilon b(k; n, d)\big(1 + o(1)\big), \tag{7}$$

where $\varepsilon = (k - nd)/nd$ and $b(k; n, d)$ is the binomial probability mass function. We require $\varepsilon > 0$ and so need $k > nd$, the expected number of breakpoints in an interval of length d under uniformity. In evaluating TEL and AML1 breakpoint clustering we use an endpoint corrected version of (7). The resultant approximation (Loader 1991, eq. 11) is

$$\Pr\{N_d \geq k\} \approx n\varepsilon b(k; n, d) + \sum_{j=k}^{n} b(j; n, d) + \sum_{j=0}^{k-1} \left(\frac{1 - d - \varepsilon d}{1 + \varepsilon - d - \varepsilon d}\right)^{2(k-j)} b(j; n, d), \tag{8}$$

where ε and $b(k; n, d)$ are as before. In our one-dimensional applications where d is small, the correction afforded by (8) is slight. This contrasts with the example considered by Loader (1991) and the two-dimensional examples that follow where, with d large, corrections are appreciable.

2.2 Multiplicity: Number of Clusters/Modes

Wiemels et al. (2000) used average kNN distances ($k = 1, \ldots, 5$) to assess overall cluster significance. But if clustering is deemed significant, then this approach does not provide estimates of cluster location or multiplicity. To remedy this, they turned to kernel density estimation of the frequency distribution of clusters across the intronic region. The location of significant modes (clusters) is established by simulation. Repeated breakpoint samples of size equal to the original are independently drawn from a uniform distribution over the intronic breakpoint region, kernel density estimates are computed for each sample, and a pointwise 95% envelope is obtained from the 95th percentile of the density estimates at each base pair (position) within the region. The results of this procedure are reproduced in Figure 2. The approach uses a priori fixed bandwidths. This is a serious shortcoming, because the arbitrarily prescribed bandwidths will have a profound effect on the identification of significant modes, as evident from considering the implications of very large or very small bandwidth selections.

By way of contrast, Figure 3 displays kernel density estimates for TEL and AML1 breakpoints using so-called "second generation" (Venables and Ripley 1999) bandwidth selection rules due to Sheather and Jones (1991). For TEL, bandwidths from either their "solve-the-equation" (STE) (8099 bp) or "direct plug-in" (DPI) (8080 bp) rules are sufficiently close so that the resultant densities almost coincide. This density [Fig. 3(a)] is clearly unimodal. The bandwidths are more than eight times larger than the bandwidth of 1000 bp used by Wiemels et al. (2000). However, for AML1, we obtain respective bandwidths of 56,792 (STE) and 82,829 (DPI), with the former supporting three modes and the latter supporting only two modes. Viewing the number of modes as a function of bandwidth is central to Silverman's bandwidth test approach, which is described in Section 2.2.1. Whichever selection rule is adopted, the estimated bandwidth is appreciably greater than that of 2000 bp used by Wiemels et al. (2000).

The question of determining the number of modes in a density has received considerable attention, with Silverman (1981) providing an easy prescription for answering it. A perhaps more subtle question is whether detecting clusters in data coincides with detecting modes in underlying densities. Silverman (1986) asserted that these are "somewhat indistinct notions with a slight difference in emphasis," whereas the Panel on Discriminant Analysis, Classification, and Clustering (1989) contended that we can "test for the presence of clustering by testing for multimodality." This latter equivalence is implicit in some of the theoretic results of Tibshirani et al. (2001) described in Section 2.2.2.

2.2.1 Silverman's Bandwidth Test. Here we provide a brief overview of Silverman's bandwidth test procedure for determining the number of modes (see Izenman and Somner 1988 and Efron and Tibshirani 1993 for additional description and applications). Let $N(f)$ be the number of modes of a density f. Consider a series of hypotheses such that the jth null hypothesis, H_0^j, is that f has at most j modes ($H_0^j : N(f) \leq j$), whereas the jth alternative, H_1^j, is that f has more than j modes ($H_1^j : N(f) > j$). Let $\hat{f}_h$ be a kernel density estimate with bandwidth h. Define $h_j^\diamond = \inf\{h : N(\hat{f}_h) \leq j\}$. Silverman (1981) showed that for Gaussian kernels, $N(\hat{f}_h)$ is a right-continuous, decreasing function of h so that $N(\hat{f}_h) > j \Leftrightarrow h < h_j^\diamond$. Thus $h_j^\diamond$ is a natural test statistic for testing H_0^j versus H_1^j. To determine $h_j^\diamond$, we count the modes in density estimates $\hat{f}_h$ for varying h. When $h = h_j^\diamond$, $\hat{f}_{h_j^\diamond}$ will have j modes plus a noticeable shoulder [i.e., the shoulder in Fig. 3(b)]. We have that

$$\Pr_f\{h_j^\diamond > h\} = \Pr\{N(\hat{f}_h) > j | X_1, \ldots, X_n \sim f\}. \tag{9}$$

Using bootstrap resampling, we can readily evaluate the right side of (9), because there is no need to recalculate $h_j^\diamond$ for each bootstrap replicate. The bandwidth test is implemented as follows:

1. Draw a bootstrap sample $X_1^*, X_2^*, \ldots, X_n^*$ from the breakpoint data $X_1, X_2, \ldots, X_n$.
2. Obtain a smooth bootstrap sample $Y_1^*, Y_2^*, \ldots, Y_n^*$ by $Y_i^* = c_j(X_i^* + h_j^\diamond \varepsilon_i)$, $i = 1, 2, \ldots, n$, where ε_i iid $\mathcal{N}(0, 1)$ and $c_j = (1 + (h_j^\diamond/\mathrm{var}(X))^2)^{-1/2}$ is a scale factor such that $\mathrm{var}(Y^*) = \mathrm{var}(X)$.
3. From $Y_1^*, Y_2^*, \ldots, Y_n^*$, compute a kernel density estimate $\hat{f}^*$ using bandwidth $h_j^\diamond$.
4. Repeat steps 1–3 B times, yielding $\hat{f}^{*b}$, $b = 1, 2, \ldots B$.

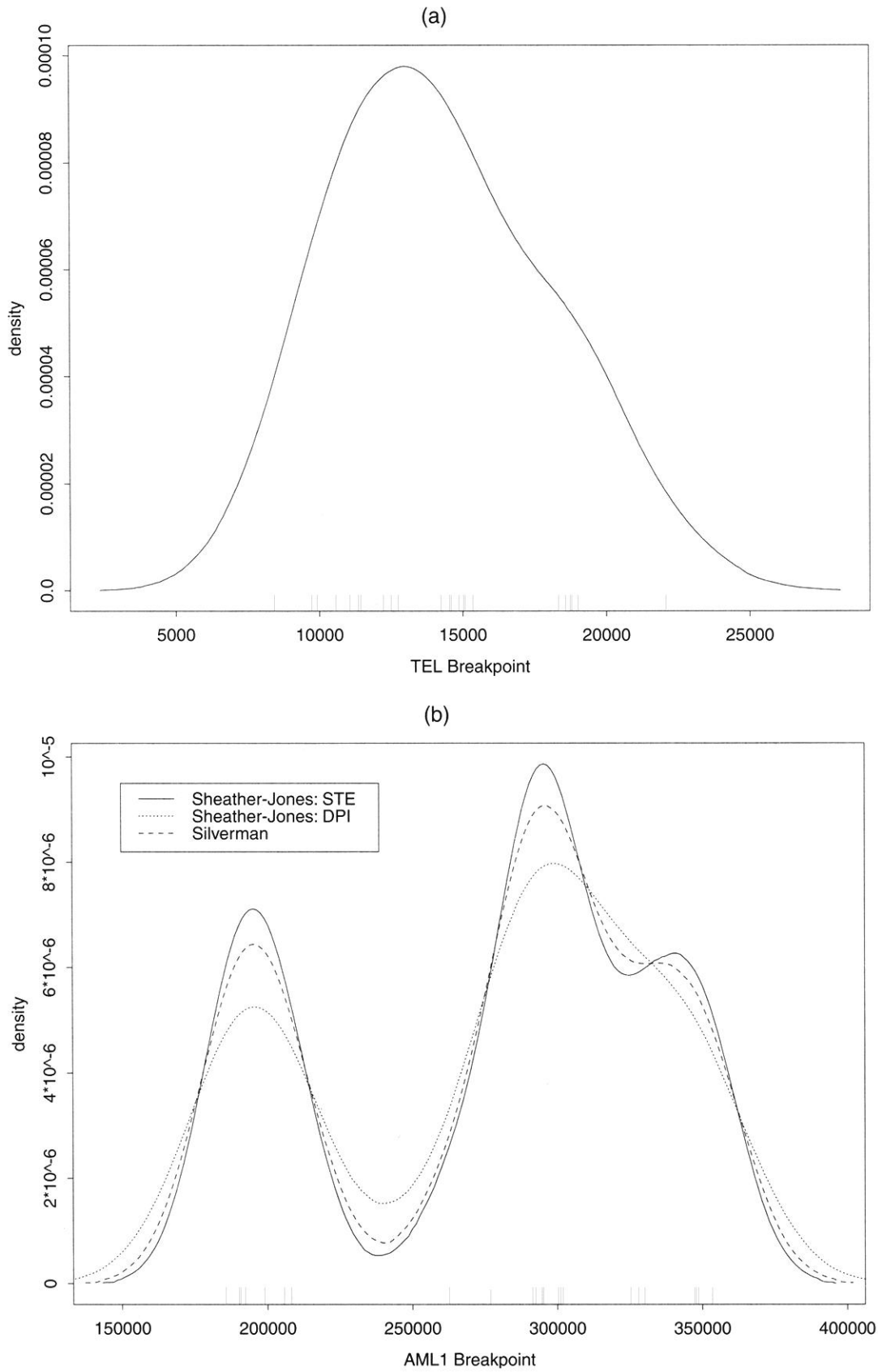

Figure 3. Breakpoint Density Estimates Using Sheather–Jones Bandwidths. (a) TEL breakpoints: The densities using either the DPI rule or the STE rule coincide. (b) AML1 breakpoints: In addition to the DPI and STE estimates, the density corresponding to $h_2^\circ = 64{,}752$ is displayed.

5. The achieved significance level for testing H_0^j versus H_1^j is $(1/B)\sum_{b=1}^{B} I\{N(\hat{f}^{*b}) > j\}$.

Step 2 corresponds to sampling from $\hat{f}_{h_j^\circ}$, the (scaled) convolution of the empiric distribution function and a standard normal distribution function. This is appropriate for testing H_0^j versus H_1^j, because $\hat{f}_{h_j^\circ}$ represents a plausible j mode density that is closest to $j+1$ modal. The procedure is computationally straightforward. As described by Silverman (1983) and Cheng and Hall (1998), it is also conservative. For this reason, and also because the bandwidth test procedure does not readily generalize to more than one-dimensional data (see Sec. 4), we next consider an alternative approach to determining the number of clusters. In Section 4 we briefly comment on a refinement of the bandwidth test that allows for local (varying) bandwidths (Minnotte 1997).

2.2.2 Gap Statistic. Tibshirani et al. (2001) developed the gap statistic as an adjunct to a clustering algorithm to formalize the "elbow" heuristic; in graphs plotting a (pooled) within-cluster error measure versus the number of clusters, there is (often) a characteristic kink or elbow, the location of which represents the appropriate number of clusters. (For an application of this heuristic, see Segal 1988.) As documented by Tibshirani et al. (2001), the many merits of the gap statistic include (a) strong theoretic underpinnings in one dimension (pertinent to translocation breakpoints), (b) applicability to any clustering algorithm in arbitrary dimensions, (c) easy implementation, and (d) excellent performance in extensive simulations.

Let $d_{ii'}$ be the distance between observations i and i'. In both our one- and two-dimensional applications, we use just the (Euclidean) distance between the breakpoints. Suppose that our clustering algorithm has generated m clusters, $C_1, C_2, \ldots, C_m$, with C_r denoting the indices of the observations in cluster r and $n_r = |C_r|$ the cluster size. Let $D_r = \sum_{i,i' \in C_r} d_{ii'}$ and $W_m = \sum_{r=1}^{m} D_r/2n_r$. If d is squared Euclidean distance, then W_m is the pooled within-cluster sum of squares around cluster means. The central idea of Tibshirani et al. (2001) is to compare $\log(W_m)$ to its expectation under an appropriate null referent distribution. They showed that in one dimension, $\mathcal{U}(0, 1)$ is most likely to produce spurious clusters (operationalized as single-component log-concave densities, which is analogous to equating clusters with modes as earlier) and so constitutes an appropriate null referent distribution. We apply choices for the more ambiguous higher-dimensional setting in Section 2.3.

The gap statistic is then defined as

$$\text{gap}_n(m) = E_n^*(\log(W_m)) - \log(W_m), \tag{10}$$

where E_n^* denotes expectation under a sample size of n from the null referent distribution; the sample size must be prescribed in view of the adaptive nature of many clustering algorithms. Motivation for (10) was provided by Tibshirani et al. (2001). The optimal number of clusters $\hat{m}$ is determined by maximizing $\text{gap}_n(m)$ after accounting for sampling variation by using a "one standard error rule" akin to that in CART (Breiman, Friedman, Olshen, and Stone 1984). The statistic itself is computed as follows:

1. Using the chosen clustering algorithm, cluster the observed data varying the total number of clusters ($m = 1, 2, \ldots, M$) giving within-dispersion measures W_m.
2. Generate B reference datasets using the uniform distribution. Repeat step 1 on each, giving within-dispersion measures W_{mb}^*, $m = 1, 2, \ldots, M$, $b = 1, 2, \ldots, B$.
3. For each m, compute the estimated gap statistic $\text{gap}(m) = (1/B)\sum_b \log(W_{mb}^*) - \log(W_m)$.

2.3 Two-Dimensional Clustering

As depicted in Figure 1, breakpoint data are paired, with each patient having breakpoints within both the TEL and AML1 intronic regions. Wiemels et al. (2000) examined whether there is corresponding two-dimensional clustering by extending their averaged nearest-neighbor methods. They also were concerned with independence of TEL and AML1 breakpoints. This was pursued by discretizing fifth nearest-neighbor distances and using contingency table methods, a seemingly oblique and inefficient approach. We directly evaluate breakpoint correlation with attendant nonparametric BC_a 95% bootstrap confidence intervals (Efron and Tibshirani 1993).

With regard to clustering, some of the aforementioned approaches generalize to two dimensions, whereas others do not. The gap statistic readily handles arbitrary dimensions. However, as demonstrated by theorem 2 of Tibshirani et al. (2001), unlike in the one-dimensional case, here there is no longer a generally applicable, least favorable referent distribution. This reflects the need to accommodate the "shape" (i.e., covariance structure) of the data at hand. As an ad hoc means of achieving this, for step 2 of the procedure given in Section 2.2.2, they proposed generating independent uniform margins over the principal components of the data. This is effected using the singular value decomposition. In our setting of n patients contributing paired breakpoint data, this works as follows. Designate the $n \times 2$ matrix of breakpoints X. Sweep out the column means and compute the singular value decomposition $X = UDV^T$. Then transform via $X^\star = XV$ and draw independent uniform margins $Z^\star$ over the column ranges of $X^\star$. Finally, create reference data by backtransformation, $Z = Z^\star V^T$. By way of contrast, we also investigate ignoring shape information and obtaining reference data by simply generating independent uniform margins for each dimension.

Extending Silverman's bandwidth test procedure is problematic because the absence of order in R_+^2 precludes relating $N(\hat{f}_h)$ to bivariate kernels with bandwidth $h = (h_1, h_2)$. In the related setting of testing unimodality, Hartigan and Hartigan (1985) proposed using minimal spanning trees to impose order in two or more dimensions. It is unclear whether such an approach is practicable for the bandwidth test.

The scan statistic itself is readily generalized to two dimensions, albeit with the constraint that the cluster regions evaluated are rectangles. Let $X_i = (X_{i1}, X_{i2})$, $x = (x_1, x_2)$ and $d = (d_1, d_2)$ and define $N_{x,x+d}$ as the number of X_i in the region $(x_1, x_1 + d_1) \times (x_2, x_2 + d_2)$. Then the scan statistic is

$$N_{d_1,d_2} = \sup_{x_1, x_2} N_{x,\, x+d}. \tag{11}$$

By defining a convenient ordering, Loader (1991) obtained two-dimensional distributional approximations. The main result is

$$\Pr\{N_{d_1,d_2} \geq k\} = \frac{n^2 d_1 d_2 (1-d_1)(1-d_2)\varepsilon^3}{(1-d_1 d_2)^3(1+\varepsilon)} \times b(k; n, d_1 d_2)(1+o(1)), \quad (12)$$

where now $\varepsilon = (k - nd_1d_2)/nd_1d_2$ and b is the binomial probability mass function as earlier. Again, requiring $\varepsilon > 0$ restricts to $k > nd_1d_2$, the expectation under uniformity. Loader (1991) also provided edge corrections that improve accuracy, at least for select d_1 and d_2, and generalization to the unknown d_1 and d_2 case. In the next section we apply both of these extensions in evaluating clustering of paired TEL–AML1 breakpoints.

3. RESULTS

3.1 One-Dimensional Clustering: Univariate Breakpoints

Considering TEL and AML1 breakpoints separately, and using average *kNN* statistics for $k = 1, \ldots, 5$, Wiemels et al. (2000) obtained (via Monte Carlo simulation) significant indications of clustering for $k = 3, 4$ (TEL) and $k = 2, 4, 5$ (AML1) (see their table 1). In both cases, combining over k and correcting for multiple comparisons was used to declare the presence of significant overall clustering. The locations of the clusters, along with accompanying claims of significance, were then determined via kernel density estimation in accordance with Figure 2.

For the reasons presented in Section 2.1, we reevaluate these clusters using scan or minimum *kNN* statistics. The identified clusters furnish the quantities d and k, permitting approximate p value determination using large deviations (8) or the moment-matching schemes in conjunction with (3) as described in Section 2.1. The results are presented in Table 1. The cluster index (first column) for TEL and AML1 corresponds to the respective clusters identified and labeled in Figure 2. We see that only the second AML1 cluster emerges as significant, with marginal results for the second TEL cluster and third AML1 cluster. For the moment approximations, evaluation of the third and fourth AML1 clusters made recourse to simulation based on the minimum *kNN* formulation, because, as previously mentioned, the approximations for such small clusters are not available. Similarly, the large-deviation approximation breaks down for the fourth cluster. The agreement among the approximations is good, especially for small tail probabilities. This is consistent with the simulation results of both Huffer and Lin (1997) and Loader (1991).

When applying the scan statistic in this fashion, it is important to note that the parameter d has been specified to correspond exactly to the clusters identified by Wiemels et al. (2000). Treating d as unknown and optimizing using the likelihood ratio test approach of Loader (1991) gives the following results. For TEL breakpoints, the most significant cluster consists of the five breakpoints labeled 13–17 in the top panel of Figure 1(a) and Figure 2(a), with a large-deviation p value of .12. That this exceeds the p value for the overlapping second TEL cluster in Table 1 is because of accommodation of the adaptation involved in finding the optimal d. For AML1, the optimal cluster consists of the eight breakpoints labeled 14, 9, 2, 15, 17, 1, 12, and 20 in Figure 2(b), with a p value of .0095. By combining clusters 2 and 3 from Table 1, a more significant result is obtained, even when allowing for optimization.

Results obtained from applying Silverman's bandwidth test for determining the number of modes are presented in Table 2. For TEL, the critical bandwidth for testing H_0^1 (at most one mode) versus H_1^1 (two or more modes) is $h_1^\diamond = 6401$, with a corresponding p value of .4, so we terminate the series of hypothesis tests and conclude that the data are unimodal, consistent with Figure 3(a). For AML1, however, we reject H_0^1 in favor of H_1^1—the critical bandwidth $h_1^\diamond = 151{,}383$ being comparable to the range of the AML1 breakpoints (167,611)—and proceed to evaluate H_0^2 (at most two modes) versus H_1^2 (three or more modes). Here we obtain a marginal result ($p = .073$) and so, in accordance with the recommendations of Izenman and Somner (1988), continue testing. Note that the critical bandwidth $h_2^\diamond = 64{,}752$ interpolates the Sheather–Jones bandwidths (56,792 for STE; and 82,829 for DPI), as is apparent from the densities in Figure 3(b). The density corresponding to $h_2^\diamond$ has a shoulder, which on further bandwidth decrease would give rise to a (third) mode as exemplified by the STE density.

Gap statistics results for $m = 1, \ldots, 5$ are presented in Figure 4. The $\hat{m}$ values obtained for TEL and AML1 are $\hat{m} = 1$ and $\hat{m} = 3$. Thus the gap statistic suggests that a single cluster/mode is indicated for TEL breakpoints, whereas three clusters are indicated for AML1 breakpoints.

Table 1. Scan Statistic p Values

Cluster	Approximation method: Markov chain	Compound poisson	Large deviation
	TEL Breakpoints		
1	.585	.588	.716
2	.097	.097	.097
3	.325	.327	.347
	AML1 Breakpoints		
1	.423	.424	.480
2	.021	.021	.021
3	.126*	.126*	.195
4	.526*	.526*	.526*

*Obtained via simulation (see text).

Table 2. Bandwidth Test Results

Number of modes	Critical bandwidth	p value
	TEL Breakpoints	
1	6401	.405
	AML1 Breakpoints	
1	151383	.001
2	64752	.073
3	35207	.395

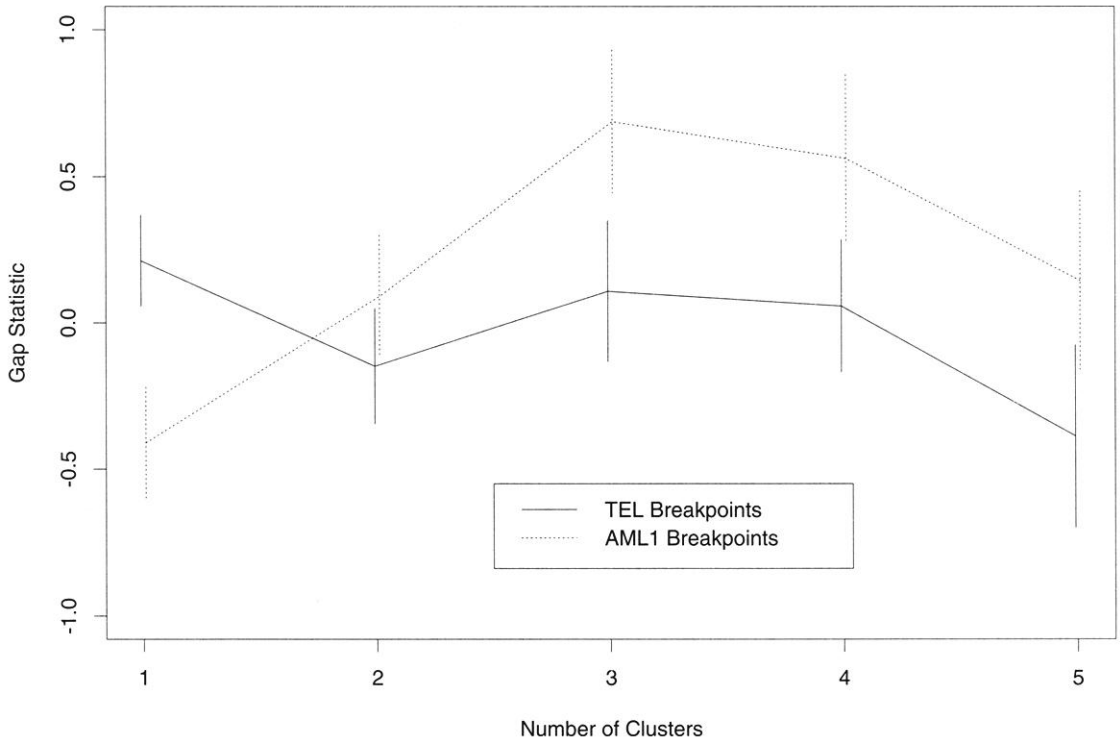

Figure 4. Gap Statistic Estimates and Standard Errors for TEL Breakpoints and AML1 Breakpoints. These have been jittered for clarity.

So, synthesizing results from the various approaches to appraising one-dimensional clustering, we see consistency with regard to TEL breakpoints; a single cluster/mode is all that is supported. The situation is less clear with regard to AML1 breakpoints, with the scan statistic confirming only one of the four clusters identified by Wiemels et al. (2000), Silverman's bandwidth test suggesting two (possibly three) clusters, and the gap statistic indicating three clusters. The latter disparity is perhaps attributable to the cited conservatism of the bandwidth test procedure. We had thought that further reconciliation of these results could be obtained by reevaluating the scan statistic for the clusters identified by the other approaches. This is because most of the clusters identified by the Wiemels et al. (2000) kernel density estimation procedures were small due to the small prescribed bandwidths and hence potentially specious. However, this reevaluation did not change the picture (irrespective of the scan statistic approximation method used); only the eight breakpoints previously itemized as yielding the best cluster when optimizing over d emerged as a significant cluster. We discuss these discrepancies between approaches in more general terms in Section 4.

3.2 Two-Dimensional Clustering: Bivariate Breakpoints

Interestingly, TEL and AML1 breakpoints are not correlated: $\rho = -.036$, 95% nonparametric bootstrap BC_a interval $(-.72, .31)$. However, this obviously does not imply an absence of bivariate clustering. We commence evaluation of two-dimensional clustering by applying the gap statistic. Whether we use referent data based on uniform margins with or without transforming according to the singular value decomposition, we obtain the same result as to the optimal number of clusters: $\hat{m} = 3$. This equivalence is not surprising in view of the aforementioned lack of dependence. Furthermore, the resultant three clusters (as determined using various algorithms with Euclidean distances) coincide with clusters based on AML1 alone; see Figure 5 and note the extensive range of within-cluster TEL breakpoints.

The three clusters so identified were used as a basis for prescribing interval lengths (d_1, d_2) for the two-dimensional scan statistic (11), the significance of which was assessed using the edge-corrected refinement of (12). None of the clusters attained significance, with respective p values of .24, .22, and .72. As described in Section 4, this disparity likely reflects the global nature of the gap statistic. It remains possible that optimizing the choice of (d_1, d_2) would detect a significant cluster. Using the result in theorem 3.2 of Loader (1991), we obtain a p value of .005 for optimized (d_1, d_2) corresponding to the four boxed breakpoints in Figure 5. The very small size of this and the closest suboptimal clusters ($k = 3$) makes their biological meaning questionable.

4. DISCUSSION

As delineated in Section 2, the three methods used differ with respect to establishing the existence of a cluster (scan statistic) versus determining the number of clusters (bandwidth test, gap statistic). This is reflected in the extent to which the methods are global (i.e., use all of the data) or local (i.e., effectively condition on individual clusters). The gap statistic is the most global approach, because it is based on an exhaustive and exclusive clustering of *all* breakpoints, implicit in step 1 of the algorithm outlined in Section 2.2.2. Thus the gap statistic estimates $\hat{m} = 3$ AML1 clusters, despite the fact that only one of these is significant by the scan statistic, because this

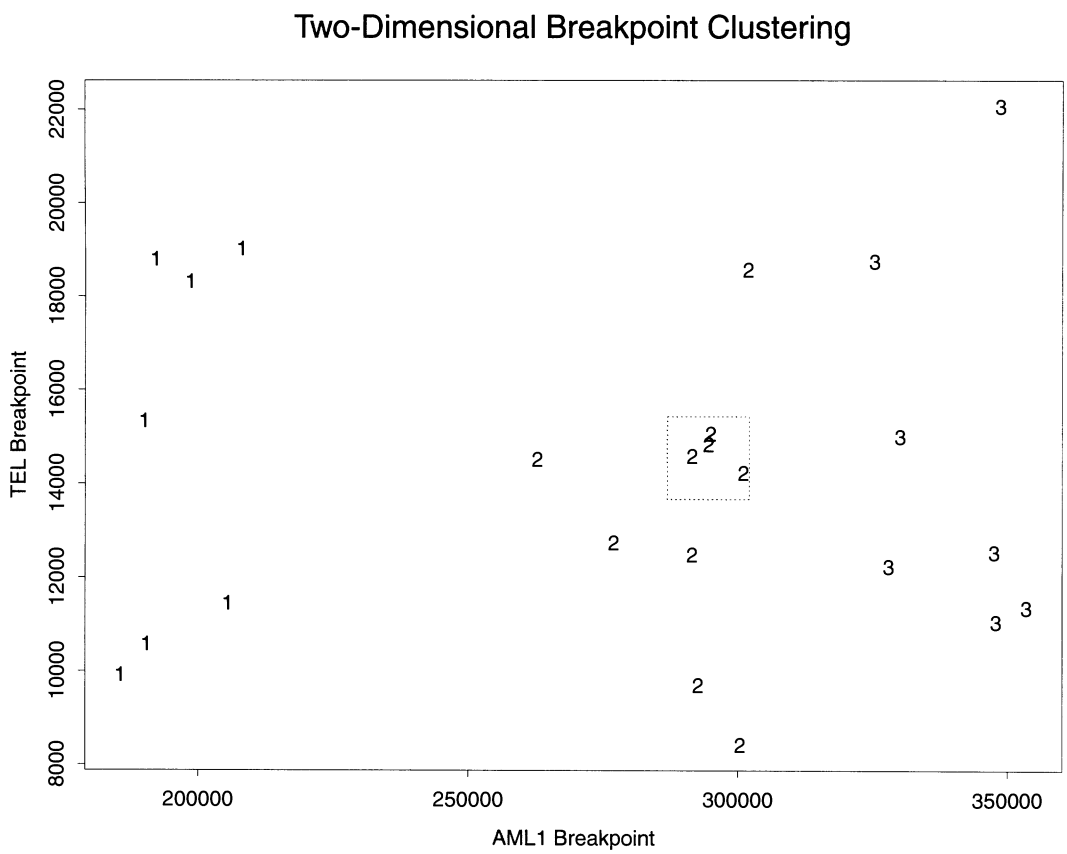

Figure 5. Bivariate Breakpoint Clustering. Breakpoints are plotted as numerals, designating which of the three gap statistic–derived clusters they belong to. The dashed box contains the cluster deemed optimal by using the two-dimensional scan statistic with unknown (d_1,d_2) (see text).

provides the optimal number of groups [according to (10)] for partitioning all of the breakpoints. The gap statistic is not designed to identify individual clusters.

Conversely, the scan statistic that is so designed is the most local approach. Given an optimal cluster (in either the d known or unknown case), it is only the number, not the distribution, of points outside that cluster that affects significance. Silverman's bandwidth test is an intermediate approach. A more local version resulting from use of variable bandwidth smoothing has been developed by Minnotte (1997). Application of his companion software gave results (not shown) comparable to the scan statistic: a single significant mode for each of TEL and AML1.

In light of these distinctions, we view the scan statistic as the frontline method for evaluating clustering of translocation breakpoints. This is because the underlying biologic interest is in identifying (and subsequently validating/testing) local regions susceptible to breakage. The exhaustive clustering of all breakpoints is not an objective in this context. Nonetheless, the gap statistic and bandwidth test provide useful complements. By identifying the collection of modes, the bandwidth test procedure can pinpoint suboptimal clusters (secondary modes) for evaluation via the scan statistic. In two dimensions, where the bandwidth test is unavailable and the scan statistic is limited to appraising rectangular regions (Loader 1991), the gap statistic is useful for initial identification of potential clusters.

As illustrated, the utility of the scan statistic is greatly enhanced by the availability of accurate approximations. It is the case, however, that because of the small sample sizes encountered with translocation breakpoint studies and the fact that data are at most two dimensional, evaluation of significance simply by recourse to simulation is straightforward, thus obviating the need for approximations. This is especially pertinent with respect to the Huffer and Lin (1997) moment-based approximations, the implementation of which are reliant on the symbolic mathematics package MAPLE.

In settings where an exhaustive clustering of all objects is desired, we believe that the gap statistic has merit in view of the properties previously itemized. For example, the analysis of cDNA microarray data has made extensive use of a variety of such clustering algorithms. A number of ad hoc procedures for determining the number of clusters have emerged: (see, e.g., Bittner et al. 2000). The easily implemented gap statistic provides a compelling addition.

Tempering all results here are questions of power for the clustering/multimodality tests used. Such issues, which arise in general, are especially pertinent in the context of breakpoint studies in view of the typically small sample sizes and difficulties in specifying alternatives given the broad competing hypotheses outlined in Section 1. This latter aspect limits even simulation-based assessments. Additionally, the specification of appropriate null distributions is also uncertain, as exemplified by the discussion in Section 2.3 and the work of Cheng and Hall (1998) and Minnotte (1997). The only mitigating

aspect to these concerns is the prospect of larger future studies resulting from more readily implemented assays.

The emergence of potential clusters for both TEL and AML1 invites further exploration via investigation of the associated motifs, as described in Section 1, to assess whether these motifs are elsewhere associated with translocation and gene fusion. A further consequence of cluster validation will be the development of directed PCR assays to more rapidly identify breakpoints for treatment follow-up studies. Our current inverse PCR methods are far too cumbersome for routine clinical use. A more routine translocation sequencing assay that targets clusters would allow the use of these breakpoints as clonotypic markers of the leukemic cell for follow-up assays of "minimal residual disease" in treated patients. Such analysis has proven highly useful in predicting relapse and tailoring therapy. Development of such a DNA-based assay would represent an important advance over the current state-of-the art assays, which are based on mRNA fusion transcripts and are inherently unstable and difficult to work with in clinical laboratories.

[Received November 2000. Revised October 2001.]

REFERENCES

Bittner, M., Meltzer, P., Chen, Y., Jian, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., Sampas, N., Dougherty, E., Wang, E., Marincola, F., Gooden, C., Lueders, J., Glatfelter, A., Pollock, P., Carpten, J., Gillanders, E., Leja, D., Dietrich, K., Beaudry, C., Berens, M., Alberts, D., and Sondak, V. (2000), "Molecular Classification of Cutaneous Malignant Melanoma by Gene Expression Profiling," *Nature*, 406, 536–540.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), *Classification and Regression Trees*, Belmont, CA: Wadsworth.

Cheng, M.-Y., and Hall, P. (1998), "Calibrating the Excess Mass and Dip Tests of Modality," *Journal of the Royal Statistical Society*, Ser. B, 60, 579–589.

Cuzick, J., and Edwards, R. (1990), "Tests for Spatial Clustering in Heterogeneous Populations," *Journal of the Royal Statistical Society*, Ser. A, 52, 73–104.

Efron, B., and Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, New York: Chapman and Hall.

Greaves, M. (1999), "Molecular Genetics, Natural History and the Demise of Childhood Leukaemia," *European Journal of Cancer*, 35, 173–185.

Guidez, F., Petrie, K., Ford, A. M., Lu, H., Bennett, C. A., MacGregor, A., Hannemann, J., Ito, Y., Ghysdael, J., Greaves, M., Wiedemann, L. M., and Zelent, A. (2000), "Recruitment of the Nuclear Receptor Corepressor N-CoR by the TEL Moiety of the Childhood Leukemia-Associated TEL–AML1 Oncoprotein," *Blood*, 96, 2557–2561.

Hartigan, J. A., and Hartigan, P. M. (1985), "The Dip Test of Unimodality," *The Annals of Statistics*, 13, 70–84.

Huffer, F., and Lin, C.-T. (1997), "Approximating the Distribution of the Scan Statistic Using Moments of the Number of Clumps," *Journal of the American Statistical Association*, 92, 1466–1475.

Huntington, R. J., and Naus, J. I. (1975), A Simpler Expression for the kth Nearest-Neighbor Coincidence Probabilities," *Annals of Probability*, 3, 894–896.

Izenman, A. J., and Somner, S. J. (1988), "Philatelic Mixtures and Multimodal Densities," *Journal of the American Statistical Association*, 83, 941–953.

Jager, U., Bocskor, S., Le, T., Mitterbauer, G., Bolz, I., Chott, A., Kneba, M., Mannhalter, C., and Nadel, B. (2000), "Follicular Lymphomas' BCL-2/IgH Junctions Contain Templated Nucleotide Insertions: Novel Insights Into the Mechanism of t(14;18) Translocation," *Blood*, 95, 3520–3529.

Karlin, S., and Macken, C. (1991), "Some Statistical Problems in the Assessment of Inhomogeneities of DNA Sequence Data," *Journal of the American Statistical Association*, 86, 27–35.

Khodarev, N. N., Sokolova, I. A., and Vaughan, A. T. (1999), "Abortive Apoptosis as an Initiator of Chromosomal Translocations," *Medical Hypotheses*, 52, 373–376.

Krowczynska, A. M., Rudders, R. A., and Krontiris, T. G. (1990), "The Human Minisatellite Consensus at Breakpoints of Oncogene Translocations," *Nucleic Acids Research*, 18, 1121–1127.

Loader, C. R. (1991), "Large Deviation Approximations to the Distribution of Scan Statistics," *Advances in Applied Probability*, 23, 751–771.

Minnotte, M. C. (1997), "Nonparametric Testing of the Existence of Modes," *The Annals of Statistics*, 25, 1646–1660.

Panel on Discriminant Analysis, Classification, and Clustering. (1989), "Discriminant Analysis and Clustering," *Statistical Science*, 4, 34–69.

Rabbitts, T. H. (1994), "Chromosomal Translocations in Human Cancer," *Nature*, 372, 143–149.

Reichel, M., Gillert, E., Breitenlohner, I., Repp, R., Greil, J., Beck, J. D., Fey, G. H., and Marschalek, R. (1999), "Rapid Isolation of Chromosomal Breakpoints From Patients With t(4;11) Acute Lymphoblastic Leukemia: Implications for Basic and Clinical Research," *Cancer Research*, 59, 3357–3362.

Segal, M. R. (1988), "Regression Trees for Censored Data," *Biometrics*, 44, 35–47.

Sheather, S. J., and Jones, M. C. (1991), "A Reliable Data-Based Bandwidth Selection Estimator for Kernel Density Estimation," *Journal of the Royal Statistical Society*, Ser. B, 53, 683–690.

Silverman, B. W. (1981), "Using Kernel Density Estimates to Investigate Multimodality," *Journal of the Royal Statistical Society*, Ser. B, 43, 97–99.

——— (1983), "Some Properties of a Test for Multimodality Based on Kernel Density Estimates," in *Probability, Statistics and Analysis*, eds. J. F. C. Kingman and G. E. H. Reuter, Cambridge, U.K.: Cambridge University Press.

——— (1986), *Density Estimation*, London: Chapman and Hall.

Strissel, P. L., Strick, R., Rowley, J.D., and Zeleznik-Le, N. J. (1998), "An *in Vivo* Topoisomerase II Cleavage Site and a DNase I Hypersensitive Site Colocalize Near Exon 9 in the MLL Breakpoint Cluster Region," *Blood*, 92, 3793–3803.

Thandla, S. P., Ploski, J. E., Raza-Egilmez, S. Z., Chhalliyil, P. P., Block, A. W., de Jong, P. J., and Aplan, P. D. (1999), "ETV6-AML1 Translocation Breakpoints Cluster Near a Purine/Pyrimidine Repeat Region in the ETV6 Gene," *Blood*, 93, 293–299.

Tibshirani, R. J., Walther, G., and Hastie, T. J. (2001), "Estimating the Number of Clusters in a Dataset via the Gap Statistic," *Journal of the Royal Statistical Society*, Ser. B, 63, 411–423.

Tsujimoto, Y., Gorham, J., Cossman, J., Jaffe, E., and Croce, C. M. (1985), "The t(14;18) Chromosome Translocations Involved in B-cell Neoplasms Result From Mistakes in VDJ Joining," *Science*, 229, 1390–1393.

van der Reijden, B. A., Dauwerse, H. G., Giles, R. H., Jagmohan-Changur, S., Wijmenga, C., Liu, P. P., Smit, B., Wessels, H. W., Beverstock, G. C., Jotterand-Bellomo, M., Martinet, D., Muhlematter, D., Lafage-Pochitaloff, M., Gabert, J., Reiffers, J., Bilhou-Nabera, C., van Ommen, G. J., Hagemeijer, A., and Breuning, M. H. (1999), "Genomic Acute Myeloid Leukemia-Associated inv(16)(p13q22) Breakpoints are Tightly Clustered," *Oncogene*, 18, 543–550.

Venables, W. N., and Ripley, B. D. (1999), *Modern Applied Statistics with S-PLUS*, New York: Springer.

Wallenstein, S., and Neff, N. (1987), "An Approximation for the Distribution of the Scan Statistic," *Statistics in Medicine*, 6, 197–207.

Wiemels, J. L., Alexander, F. E., Cazzaniga, G., Biondi, A., Mayer, S. P., and Greaves, M. (2000), "Microclustering of TEL–AML1 Translocation Breakpoints in Childhood Acute Lymphoblastic Leukemia," *Genes, Chromosomes and Cancer*, 29, 219–228.

Wiemels, J. L., Cazzaniga, G., Daniotti, M., Eden, O. B., Addison, G. M., Masera, G., Saha, V., Biondi, A., and Greaves, M. F. (1999a), "Prenatal Origin of Acute Lymphoblastic Leukaemia in Children," *Lancet*, 354, 1499–1503.

Wiemels, J. L., Ford, A. M., Van Wering, E. R., Postma, A., and Greaves, M. (1999b), "Protracted and Variable Latency of Acute Lymphoblastic Leukemia After TEL–AML1 Gene Fusion *in Utero*," *Blood*, 94, 1057–1062.

Wiemels, J. L., and Greaves, M. (1999), "Structure and Possible Mechanisms of TEL–AML1 Gene Fusions in Childhood Acute Lymphoblastic Leukemia," *Cancer Research*, 59, 4075–4082.

Wyatt, R. T., Rudders, R. A., Zelenetz, A., Delellis, R. A., and Krontiris, T. G. (1992), "Bcl2 Oncogene Translocation is Mediated by a Chi-Like Consensus," *Journal of Experimental Medicine*, 175, 1575–88.