Title
Clustering of translocation breakpoints

Permalink
https://escholarship.org/uc/item/5rr9k945

Authors
Segal, Mark R
Wiemels, Joseph L

Publication Date
2001-06-01

Peer reviewed

Clustering of Translocation Breakpoints

Mark R. Segal and Joseph L. Wiemels

Department of Epidemiology and Biostatistics,

University of California, San Francisco, CA 94143-0560

email: mark@biostat.ucsf.edu

Abstract

Translocation, a physical movement of genetic material from one chromosome to another, can aberrantly juxtapose portions of two cellular genes. This type of fusion may disrupt cellular function by producing novel, biologically-active fused genes, or by 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. While the biological activity of the oncogenic effects of genes involved in translocations are 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 as long as they take place within certain bounded regions within the genes. Because of this independence a patterned, in particular clustered, distribution of fusion breakpoints within a given region will potentially yield relevant information about the etiology of the fusion.

The statistical analysis of translocation breakpoints has, accordingly, focussed on the extent to which they cluster. Somewhat questionable methods have been employed in this regard. After highlighting these shortcomings, we introduce a variety of approaches including scan statistics, smoothed bootstrap, 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. Results obtained indicate generally weaker evidence for clustering than previously reported, and also highlight differences between the statistical approaches.

KEY WORDS: Gene fusion; Gap statistic; Scan statistic; Smoothed bootstrap.

1   Introduction

Translocation is defined as the physical movement of genetic material between two non-homologous chromosomes. In the simplest case, the 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 karyotype, of a patient's cells. At the level of the DNA sequence, specific genes may be cut in half, 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 alter regulation of gene expression (Rabbitts, 1994). Dozens of translocations have been described in the leukemias, lymphomas and sarcomas. A given translocation between two cellular genes is consistently associated with a specific tumor type. This permits the development of diagnostics and/or therapeutics based on the particular gene fusion products.

Translocations in the leukemias, which are our focus, usually result in the formation of a chimeric protein, in which the proximal end of one protein is fused to the terminal end of the other. These proteins are usually transcription factors – proteins present in the nucleus that control the expression of other genes involved in growth and development of blood cells. When the normal development program of the blood stem cells is interrupted by the aberrant fusion transcription factor, leukemia may result. Genes are structured in such a manner so as to have protein coding regions, or exons, interspersed with noncoding regions, or introns. Translocations which 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 re-fused on either chromosome. This breakage/re-fusion site is called a "breakpoint" and is unique to each individual patient diagnosed with a particular translocation. The "clustering" of breakpoints in a specific region indicates that the region is fragile and may be susceptible to cleavage by chemicals or DNA-modifying enzymes. The localization of such putative clusters requires

breakpoints from multiple patients and appropriate statistical validation. The identification and characterization of breakpoint clusters will ultimately aid etiologic, epidemiologic, and diagnostic studies of leukemia.

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 which occurs in one-quarter of childhood acute lymphoblastic leukemia (ALL), 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 the age of 14 (Wiemels et al., 1999a,b). Despite this knowledge of the temporal origin of the translocation, little is understood regarding the process of formation of the fusion. The translocation results in a chimeric oncogene with the first five exons of *TEL* fused to nearly the entire *AML1* protein coding region. *AML1* is considered to be a "master" transcription factor, and is a critical conductor 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 14000 base pair (bp) intron 5 of *TEL* and comparatively large 160000 bp intron 1-2 of *AML1*. Both *TEL* and *AML1* are involved in a variety of 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 swathe 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" (RSS) in the formation of such translocations (Jager et al., 2000; Tsujimoto et al., 1985). These are short, sixteen base-pair motifs, whose orientation allows them to be recognized by select cellular enzymes. These enzymes normally rearrange genes of the immune system in order to produce the antibody repertoire. This gene rearrangement process is critical for formation of the estimated 10^7 different anti-

bodies (and corresponding genes) necessary for immune system function in a given individual. 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 with the leukemias is very different in that breakpoint distributions tend to be far more diffuse, resulting in a poor understanding of their 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 which precedes the expression of the enzymes that manipulate RSS. Only recently were methods developed to sequence these leukemia fusions (Reichel et al., 1999; Wiemels and Greaves, 1999), spawning attendant need for applying statistical methods to analyze breakpoint distributions. The existence of clusters in particular regions suggests that features of the intrinsic DNA sequence and/or chromatin are critical to translocation. Accordingly, to the extent that the location of translocation breakpoints has been subject to any statistical treatment, the analyses have focussed on evaluating and localizing putative clusters.

The purpose of the present paper is to identify some shortcomings in the limited approaches to appraising clustering that have been taken to date. These are reviewed in section 2 where a variety of improvements, drawing on recent statistical work, are also described. These methods include scan statistics with attendant distributional approximations, Silverman's (1981) smoothed bootstrap procedure, and gap statistics (Tibshirani et al., 2000). As illustrated, these methods differ according to whether the emphasis is on appraising a specific cluster or determining the number of clusters. Section 3 presents a reanalysis of the particular *TEL-AML1* fusion described above while section 4 describes some possible extensions and offers concluding discussion.

2   Approaches to Appraising Clustering

As mentioned, very little in the way of formal assessment of clustering is pursued in evaluating translocation breakpoint distributions. Indeed, van der Reijden et al., (1999) assert (in the title itself!) that acute myeloid leukemia-associated inv(16)(p13q22) breakpoints are tightly clustered without undertaking any related analysis. The only formal approach to date is that of Wiemels et al., (2000) and it is on both their data and methods that we subsequently focus. A preview is provided by Figures 1 and 2.

The data itself is displayed in Figure 1, with the top panel depicting *TEL* breakpoints and the bottom panel *AML1* breakpoints. The shaded boxes represent exons of the respective genes, with the breakpoints (primarily) occurring in the intervening introns. In both panels 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 determined.

Figure 2 is taken from Wiemels et al., (2000) and depicts breakpoint density estimates using (gaussian) kernel density estimation with prescribed bandwidths. The bandwidths used are 1000 bp and 2000 bp for *TEL* and *AML1* respectively. Later, we show that these are much too small. Regions were the kernel density estimate exceeds a 95% confidence envelope obtained via simulation (described in section 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) use 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). We now focus on *kNN* distances and then discuss density estimation approaches in section 2.2.1. Consider a situation where we have $c \Leftrightarrow 1$ 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 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 with regard 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. Conversely, 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)). The motivation for using average *kNN* distances derives from Cuzick and Edwards (1990), however, they were dealing with a different (case-control) context wherein averaging over all case-control distances was appropriate. While it is the case that average *kNN* distances are distributionally more tractable than minimum *kNN* distances, there are a variety of accurate and readily computable approximations for the latter. We next outline two such approximations which are among those employed for a more formal evaluation of *TEL-AML1* clustering 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 a variety of approximations. Instead of working directly with scan or minimum *kNN* statistics, Huffer and Lin (1997) reformulate

5

in terms of *clumps*. In particular, a $k : d$ clump exists if there are k consecutive points in an interval of length d. Let $Y_{k:d} \equiv Y$ be the number of $k : d$ clumps:

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

Since $Y \ge 1$ if and only if $N_d \ge k$ we have

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

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

Huffer and Lin (1997) pursue 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 utilizes only the first two moments of Y. We later use both this and another approximation based on matching moments to a compound Poisson distribution – the two methods yield very similar results. Explicit formulae for the first two moments of Y are obtained using properties of *spacings* which are distances between consecutive order statistics. The resultant formulae involve the sample size n, number of points k, interval width d, and cumulative binomial and trinomial probabilities; see Huffer and Lin (1997, section 3.2). While quite general, these formulae do not hold for $k \le 3$ and $n < 2(k-1)$, a restriction 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 having transition matrix $\mathbf{P}$ and define $\tilde{Y} = \sum_{i=1}^{w} Z_i$. For notational simplicity write $s = 1/(a + b)$ and $\pi = \pi_1$. Then we have

$$\Pr\{\tilde{Y} \ge 1\} = 1 - (1 - \pi)(1 - \pi/s)^{w-1} \tag{4}$$

$$\mathrm{E}\tilde{Y} = w\pi \tag{5}$$

$$\mathrm{Var}(\tilde{Y}) \approx \pi(1 - \pi)(w + 2(s - 1)(w - s)). \tag{6}$$

Matching (5) and (6) to the first two moments of Y yields closed form solutions for π and s, and whence for $\Pr\{\tilde{Y} \geq 1\}$ by (4). The latter then 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 $w \leq 21$ (since $k > 3$), so there is less basis for appealing to Markov chain stationarity. While limited simulations for this sample size again indicate that the Markov chain (and compound gamma) approximations are very accurate, in order not to rely solely on moment-based approaches we next consider alternative large-deviation approximations for $\Pr\{N_d \geq k\}$.

Loader (1991) considers both one and two dimensional scan statistics as well as distinguishing between d known and unknown. Here, we briefly summarize results for the known d case. While 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 as follows.

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

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

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

where ϵ and $b(k; n, d)$ are as above. 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 below where, with d large, corrections are appreciable.

2.2 Multiplicity: Number of Clusters / Modes

The use of average k nearest neighbor distances for $k = 1, \ldots, 5$ provided an overall assessment as to whether there is significant clustering. If so, it does not provide an indication of cluster location or multiplicity. To remedy this, Wiemels et al., (2000) turn to kernel density estimates. The location of significant modes (clusters) is then established by simulation: repeated breakpoint samples of equal size 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 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 since 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 that the resultant densities almost coincide. This density (Figure 3(a)) is clearly unimodal. The bandwidths are more than 8 times larger than the bandwidth of 1000 bp used by Wiemels et al., (2000). However, for *AML1*, we obtain respective bandwidths of 56792 (STE) and 82829 (DPI) with the former supporting 3 modes and the latter only 2. Viewing the number of modes as a function of bandwidth is central to Silverman's smoothed bootstrap approach, which is described in Section 2.2.1. Irrespective of which bandwidth selection rule is adopted, the estimated bandwidth is appreciably greater than the bandwidth of 2000 bp prescribed by Wiemels et al., (2000).

The question of determining how many modes a density possesses has received considerable attention, with Silverman (1981) providing an easy and compelling prescription for answering it. Perhaps more subtle is whether detecting clusters in data coincides with detecting modes in underlying densities with Silverman (1986) asserting that these are "somewhat indistinct

notions with a slight difference in emphasis", while the Panel on Clustering (1989) contends 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., (2000) described in section 2.2.2.

2.2.1 Silverman's Smoothed Bootstrap

We provide a brief overview of Silverman's smoothed bootstrap 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 ($H_0^j : N(f) \leq j$) while 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) shows 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 vs. 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 (*cf* the shoulder in Figure 3(b)). We have that

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

By using bootstrap resampling we can readily evaluate the right hand side of (9) since there is no need to recalculate $h_j^\diamond$ for each bootstrap replicate.

The prescription for effecting bootstrap testing is 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

$$Y_i^* = c_j(X_i^* + h_j^\diamond \epsilon_i) \quad i = 1, 2, \ldots, n \tag{10}$$

9

where ϵ_i iid $\mathcal{N}(0,1)$ and $c_j = 1/\sqrt{1 + (h_j^\diamond/\mathrm{Var}(X))^2}$ is a scale factor employed so 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 $\hat{f}^{*b}$, $b = 1, 2, \ldots B$.

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

Step 2 corresponds to sampling from $\hat{f}_{h_j^\diamond}$, 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 since $\hat{f}_{h_j^\diamond}$ represents a plausible j mode density that is closest to $j + 1$ modal. The procedure is computationally straightforward. As described by Silverman (1983) and Izenman and Somner (1988) it is also conservative. For this reason, and additionally because the smoothed bootstrap procedure does not readily generalize to more than one dimensional data (see section 4), we consider next an alternative approach to determining the number of clusters.

2.2.2 Gap Statistic

Tibshirani, Walther and Hastie (2000) develop the gap statistic as an adjunct to a clustering algorithm in order 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 applications of the heuristic see Segal (1988) and Sugar et al., (1999). As documented by Tibshirani et al., (2000) the merits of the gap statistic are numerous: (i) strong theoretic underpinnings in one dimension (pertinent to translocation breakpoints), (ii) applicable with any clustering algorithm in arbitrary dimensions, (iii) easily implemented, and (iv) 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 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 $W_m = \sum_{r=1}^{m} D_r / 2 n_r$. 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., (2000) is to compare $\log(W_m)$ to its expectation under an appropriate null referent distribution. They show 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 above) and so constitutes an appropriate ("least favorable") null referent distribution. Relatedly, Hartigan and Hartigan (1985) utilize the uniform as the least favorable unimodal distribution for assessing the power of their dip statistic of unimodality. Tibshirani et al., (2000) further describe choices for the more ambiguous higher dimensional setting which we present in section 2.3.

The gap statistic is then defined as

$$\text{Gap}_n(m) = E_n^*(\log(W_m)) \Leftrightarrow \log(W_m) \tag{11}$$

where E_n^* denotes expectation under a sample size of n from the null referent distribution; it is necessary to prescribe the sample size in view of the adaptive nature of many clustering algorithms. Motivation for the definition (11) is provided by Tibshirani et al., (2000). 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 employed in CART (Breiman et al., 1984). The computational procedure is 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 prescription. 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

$$\text{Gap}(m) = (1/B) \sum_b \log(W_{mb}^*) \Leftrightarrow \log(W_m). \tag{12}$$

11

4. Let $\bar{l} = (1/B)\sum_b \log(W^*_{mb})$. Compute the standard deviation of $\log(W^*_m)$'s as $\mathrm{sd}_m = [(1/B)\sum_b(\log(W^*_{mb}) \Leftrightarrow \bar{l})^2]^{1/2}$, and define $s_m = \mathrm{sd}_m\sqrt{1 + 1/B}$.

5. Choose the number of clusters via

$$\hat{m} = \min_m\{\mathrm{Gap}(m) \geq \mathrm{Gap}(m+1) \Leftrightarrow s_{m+1}\}. \tag{13}$$

2.3 Two-Dimensional Clustering

Breakpoint data are, in fact, paired – here each patient has breakpoints within both the *TEL* and *AML1* intronic regions. Wiemels et al., (2000) examine whether there is corresponding two-dimensional clustering by extending their averaged nearest neighbor methods. They also are concerned with independence of *TEL* and *AML1* breakpoints. This is pursued by discretizing fifth nearest neighbor distances and using contingency table methods which is seemingly both oblique and inefficient. We directly evaluate breakpoint correlation with attendant non-parametric BC_a 95% bootstrap confidence intervals (Efron and Tibshirani, 1993).

With regard clustering, some of the above approaches generalize to two dimensions whereas others do not. The gap statistic readily handles arbitrary dimensions, although there are issues surrounding choice of an appropriate referent distribution. As demonstrated by Theorem 2 of Tibshirani et al., (2000), unlike the one-dimensional case, there is no longer a generally applicable, least favorable referent distribution. This reflects the need to accommodate the "shape" (covariance structure) of the data at hand. As an ad hoc means of achieving this they propose, for step 2 of the procedure given in section 2.2.2, 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 smooth bootstrap procedure is problematic since 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) propose using minimal spanning trees to impose order in two or more dimensions. It is unclear whether such an approach is practicable for the smoothed bootstrap.

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

$$N_{d_1,d_2} = \sup_{x_1,x_2} N_{x,x+d} \tag{14}$$

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

$$\Pr\{N_{d_1,d_2} \geq k\} = \frac{n^2 \, d_1 \, d_2 \, (1 \Leftrightarrow d_1)(1 \Leftrightarrow d_2)\epsilon^3}{(1 \Leftrightarrow d_1 d_2)^3 (1 + \epsilon)} \, b\,(k; n, d_1 d_2)(1 + o(1)) \tag{15}$$

where now $\epsilon = (k \Leftrightarrow n \, d_1 \, d_2)/n \, d_1 \, d_2$ and b is the binomial probability mass function as previously. Again, requiring $\epsilon > 0$ restricts to $k > n \, d_1 \, d_2$, the expectation under uniformity. Loader (1991) also provides (i) edge corrections that improve accuracy, at least for select d_1, d_2 and (ii) generalization to the unknown d_1, d_2 case. Both these extensions are applied in evaluating clustering of paired *TEL-AML1* breakpoints in the next section.

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) obtain (via Monte Carlo simulation) significant indications

of clustering for $k = 3, 4$ (*TEL*) and $k = 2, 4, 5$ (*AML1*) (see their Table 1). The fact that the most significant results obtained with $k = 3$ (*TEL*) and $k = 2$ (*AML1*) is used to infer that multiple clusters exist. 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 as per 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 Huffer and Lin (1997) 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, since, as previously mentioned, the approximations are not available for such small clusters. 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).

In applying the scan statistic in this fashion it is important to note that the parameter d has been specified so as to correspond exactly to the respective clusters as identified by Wiemels et al., (2000). If instead we treat d as unknown and optimize using the likelihood ratio test prescription of Loader (1991) (Theorem 2.2) we obtain the following results. For *TEL* breakpoints, the most significant cluster consists of the five breakpoints labeled 13 through 17 in the top panel of Figure 1 and Figure 2A, with large-deviation p-value of 0.12. That this exceeds the p-value for the overlapping 2nd *TEL* cluster in Table 1(a) is due to accommodating 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, 20 in Figure 2B, with p-value 0.0095. By combining clusters 2 and 3 from Table 1(b) a much more significant result is obtained, despite

allowing for the optimization.

Results from applying Silverman's smoothed bootstrap method 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 0.4, so we terminate the series of hypothesis tests and conclude that the data is unimodal, consistent with Figure 3(a). For *AML1* however, we reject H_0^1 in favor of H_1^1 – the critical bandwidth $h_1^\diamond = 151383$ being comparable to the range of the *AML1* breakpoints (167611) – and proceed to evaluating H_0^2 (at most two modes) versus H_1^2 (three or more modes). Here we obtain a marginal result (p = 0.073) and so, in accord with the recommendations of Izenman and Somner (1988), continue testing. Note that the critical bandwidth $h_2^\diamond = 64752$ interpolates the Sheather-Jones bandwidths (56792 - STE; 82829 - DPI) as is apparent from the densities in Figure 3(b): the density corresponding to $h_2^\diamond$ has a shoulder which on further decrease in bandwidth 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 $\hat{m} = 3$ respectively. Thus, the gap statistic suggests that a single cluster/mode is indicated for *TEL* breakpoints, while 3 clusters are indicated for *AML1* breakpoints.

So, synthesizing results from the various approaches to appraising one-dimensional clustering, we see consistency with regard *TEL* breakpoints: a single cluster/mode is all that is supported. The situation is less clear with regard *AML1* breakpoints with the scan statistic only affirming one of the four clusters identified by Wiemels et al., (2000), Silverman's smoothed bootstrap suggesting 2 (possibly 3) clusters, and the gap statistic indicating three clusters. The latter disparity is perhaps attributable to the cited conservatism of the smoothed bootstrap procedure. We thought further reconciliation of these results could be obtained by re-evaluating the scan statistic for the clusters identified by the other approaches. This is because most of the clusters identified by Wiemels et al., (2000) kernel density estimation were small due to the small prescribed bandwidths and hence potentially specious. However, this re-evaluation 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 further 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: $\rho = -0.036$, 95% nonparametric bootstrap BC_a interval (-0.72, 0.31). 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 above lack of dependence. Furthermore, the resultant three clusters (as determined using a variety of clustering 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 3 clusters so identified were used as a basis for prescribing interval lengths (d_1, d_2) for the two-dimensional scan statistic (14), the significance of which was assessed using the edge corrected refinement of (15). None of the clusters attained significance with respective p-values of 0.24, 0.22 and 0.72. 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 0.005 for optimized (d_1, d_2) corresponding to the 4 boxed breakpoints in Figure 5. The very small size of this and the closest sub-optimal clusters ($k = 3$) makes their biological meaning questionable.

16

4 Discussion

As delineated in section 2, the three methods employed differ with respect to establishing existence of a cluster (scan statistic) versus determining the number of clusters (smoothed bootstrap, gap statistic). This is reflected in the extent to which the methods are global (i.e., utilize all the data) or local (i.e., effectively condition on individual clusters). The gap statistic is the most global approach as it is based on an exhaustive and exclusive clustering *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 only one of these being significant according to the scan statistic, since this provides the optimal number of groups for partitioning all the breakpoints. The gap statistic is not designed to extract individual clusters.

Conversely, the scan statistic which 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, and not the distribution of points, outside that cluster that affects significance. Silverman's smoothed bootstrap testing is an intermediary approach. While a more local version would seemingly result from use of variable bandwidth smoothing, this would complicate the one-to-one relationship between bandwidth and number of modes, upon which the methodology relies. implementing such an approach is prohibitive.

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 smoothed bootstrap provide useful complements. By identifying the collection of modes, the smoothed bootstrap procedure can pinpoint suboptimal clusters (secondary modes) for evaluation via the scan statistic. In two dimensions, where the smoothed bootstrap is unavailable and the scan statistic is limited to appraising rectangular regions (Loader, 1991), the gap statistic is useful for initial extraction 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 typically small sample sizes encountered with translocation breakpoint studies coupled with the fact that data is at most two dimensional, evaluation of significance by recourse to simulation is straightforward. This is especially pertinent with respect to the Huffer and Lin (1997) moment based approximations, which are reliant on the symbolic mathematics package MAPLE.

In settings where an exhaustive clustering of all objects is desired we believe the gap statistic has merit in view of the properties previously itemized. 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.

Extensions to be investigated for studying translocation breakpoints include (a) devising methods for appraising whether there is common breakpoint clustering across differing patient groups, and (b) utilizing sequence database search methods (e.g., Altschul et al., 1997) for assessing whether characteristic breakpoint motifs are elsewhere associated with translocation and gene fusion.

Acknowledgements

References

Altschul SF, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**:3389-3402.

Bittner M, Meltzer P, Chen Y, et al., (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, **406**:536-540.

Breiman L, Friedman JH, Olshen RA, Stone CJ. (1984). *Classification and Regression Trees*. Belmont: Wadsworth.

Cuzick J, Edwards R. (1990). Tests for spatial clustering in heterogeneous populations. *Journal of the Royal Statistical Society* A, **52**:73-104.

Efron B, Tibshirani R. (1993). *An Introduction to the Bootstrap*. London: 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 AM, Lu H, Bennett CA, MacGregor A, Hannemann J, Ito Y, Ghysdael J, Greaves M, Wiedemann LM, 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**:2557-61.

Hartigan JA, Hartigan PM. (1985). The dip test of unimodality. *Annals of Statistics*, **13**:70-84.

Huffer F, 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 RJ, Naus JI. (1975). A simpler expression for the kth nearest neighbor coincidence probabilities. *Annals of Probability* **3**:894-896.

Izenman AJ, Somner SJ. (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, 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, Macken C. (1991). Some statistical problems in the assessment of inhomogeneities of DNA sequence data. *Journal of the American Statistical Association*, **86**:27-35.

Loader CR. (1991). Large deviation approximations to the distribution of scan statistics. *Advances in Applied Probability*, **23**:751-71.

Panel on Discriminant Analysis, Classification, and Clustering. (1989). Discriminant analysis and clustering. *Statistical Science*, **4**:34-69.

Rabbitts TH. (1994). Chromosomal translocations in human cancer. *Nature*, **372**:143-149.

Reichel M, Gillert E, Breitenlohner I, Repp R, Greil J, Beck JD, Fey GH, 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 MR. (1988). Regression trees for censored data. *Biometrics*, **44**:35–47.

Sheather SJ, Jones MC. (1991). A reliable data-based bandwidth selection estimator for kernel density estimation. *Journal of the Royal Statistical Society* B, **53**:683-690.

Silverman BW. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society* B, **43**:97-99.

Silverman BW. (1983). Some properties of a test for multimodality based on kernel density estimates. In *Probability, Statistics and Analysis*, (Kingman and Reuter, eds). Cambridge: Cambridge University Press.

Silverman BW. (1986). *Density Estimation.* London: Chapman and Hall.

Sugar C, Lenert L, Olshen R. (1999). An application of cluster analysis to health services research: Empirically defined health states for depression from the sf-12. Technical Report, Department of Statistics, Stanford University.

Tibshirani RJ, Walther G, Hastie TJ. (2000). Estimating the number of clusters in a dataset via the gap statistic. Technical Report, Department of Statistics, Stanford University.

Tsujimoto Y, Gorham J, Cossman J, Jaffe E, Croce CM. (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 BA, Dauwerse HG, Giles RH, et al., (1999). Genomic acute myeloid leukemia-associated inv(16)(p13q22) breakpoints are tightly clustered. *Oncogene*, **18**:543-550.

Venables WN, Ripley BD. (1999). *Modern Applied Statistics with S-Plus*. New York: Springer.

Wallenstein S, Neff N. (1987). An approximation for the distribution of the scan statistic. *Statistics in Medicine*, **6**:197-207.

Wiemels JL, Cazzaniga G, Daniotti M, Eden OB, Addison GM, Masera G, Saha V, Biondi A, Greaves MF. (1999a). Prenatal origin of acute lymphoblastic leukaemia in children. *Lancet*, **354**:1499-1503.

Wiemels JL, Ford AM, Van Wering ER, Postma A, Greaves M. (1999b). Protracted and variable latency of acute lymphoblastic leukemia after *TEL-AML1* gene fusion in utero. *Blood*, **94**:1057-1062.

Wiemels JL, Greaves M. (1999). Structure and possible mechanisms of *TEL-AML1* gene fusions in childhood acute lymphoblastic leukemia. *Cancer Research*, **59**:4075-4082.

Wiemels JL, Alexander FE, Cazzaniga G, Biondi A, Mayer SP, Greaves M. (2000). Microclustering of *TEL-AML1* translocation breakpoints in childhood acute lymphoblastic leukemia. *Genes, Chromosomes and Cancer*, **29**:219-228.

Table 1: Scan Statistic P-Values

(a) *TEL Breakpoints*

	Approximation Method		
Cluster	Markov Chain	Compound Poisson	Large Deviation
1	0.585	0.588	0.716
2	0.097	0.097	0.097
3	0.325	0.327	0.347

(b) *AML1 Breakpoints*

	Approximation Method		
Cluster	Markov Chain	Compound Poisson	Large Deviation
1	0.423	0.424	0.480
2	0.021	0.021	0.021
3	$0.126^{\dagger}$	$0.126^{\dagger}$	0.195
4	$0.526^{\dagger}$	$0.526^{\dagger}$	$0.526^{\dagger}$

$\dagger$: obtained via simulation (see text).

Table 2: Smooth Bootstrap Results

TEL Breakpoints

Number of Modes	Critical Bandwidth	P-value
1	6401	0.405

AML1 Breakpoints

Number of Modes	Critical Bandwidth	P-value
1	151383	0.001
2	64752	0.073
3	35207	0.395

Figure Captions

Figure 1: Breakpoint locations within the *TEL* and *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 being a patient identifier.

Figure 2: A: *TEL* breakpoint locations with corresponding gaussian kernel density estimate and 95% pointwise confidence envelope (see text). Starred numerals designate the putative clusters re-evaluated in Table 1. **B**: Same for *AML1*.

Figure 3: Breakpoint density estimates using Sheather-Jones bandwidths: (a) *TEL* breakpoints: the densities using either the direct plug-in (DPI) rule or solve-the-equation (STE) rule coincide; (b) *AML1* breakpoints: in addition to the DPI and STE estimates, the density corresponding to $h_2^\diamond = 64752$ is displayed.

Figure 4: Gap statistic estimates and standard errors: (a) *TEL* breakpoints; (b) *AML1* breakpoints.

Figure 5: Bivariate breakpoint clustering. Breakpoints are plotted as numerals, designating which of the 3 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).
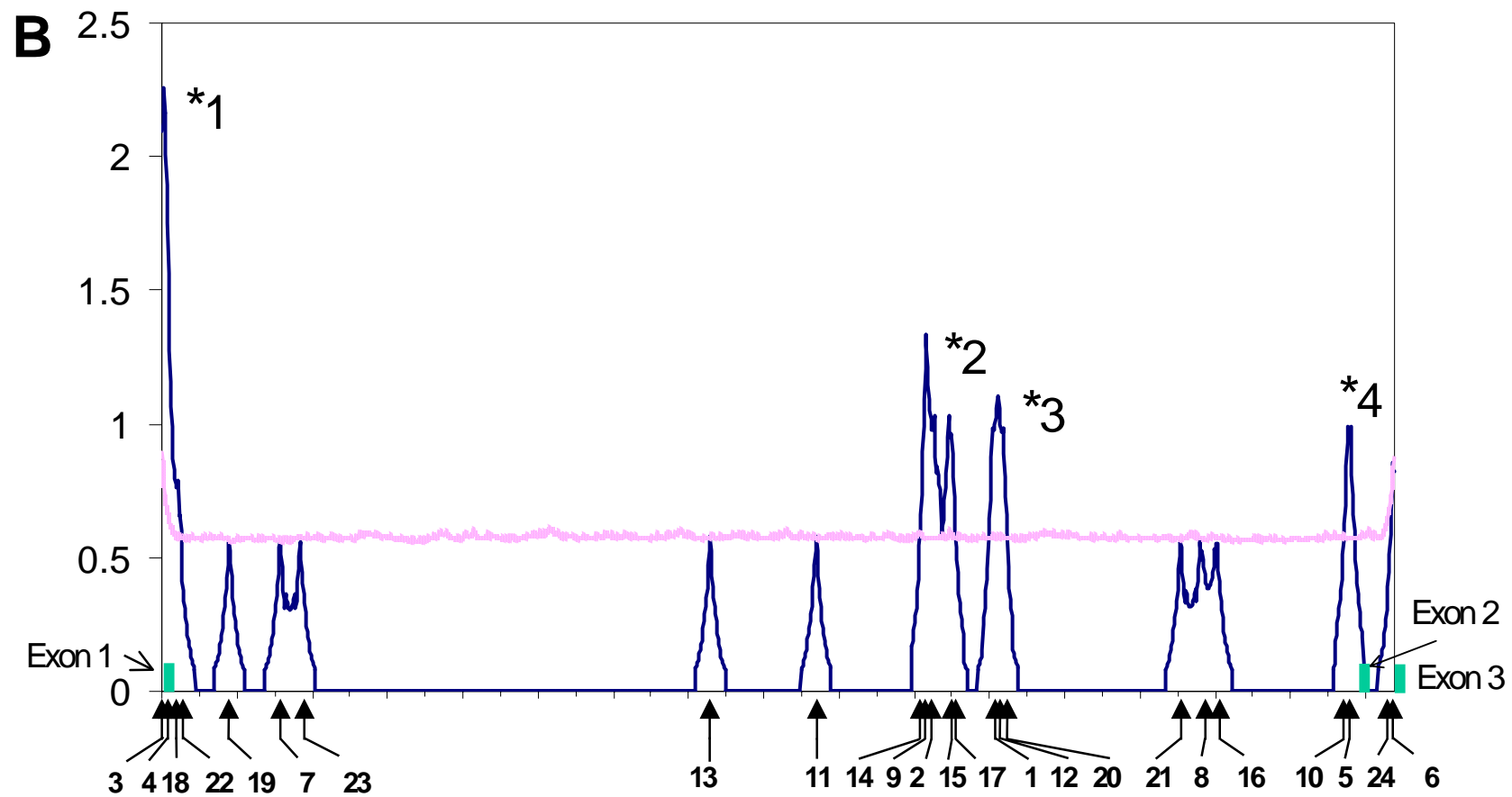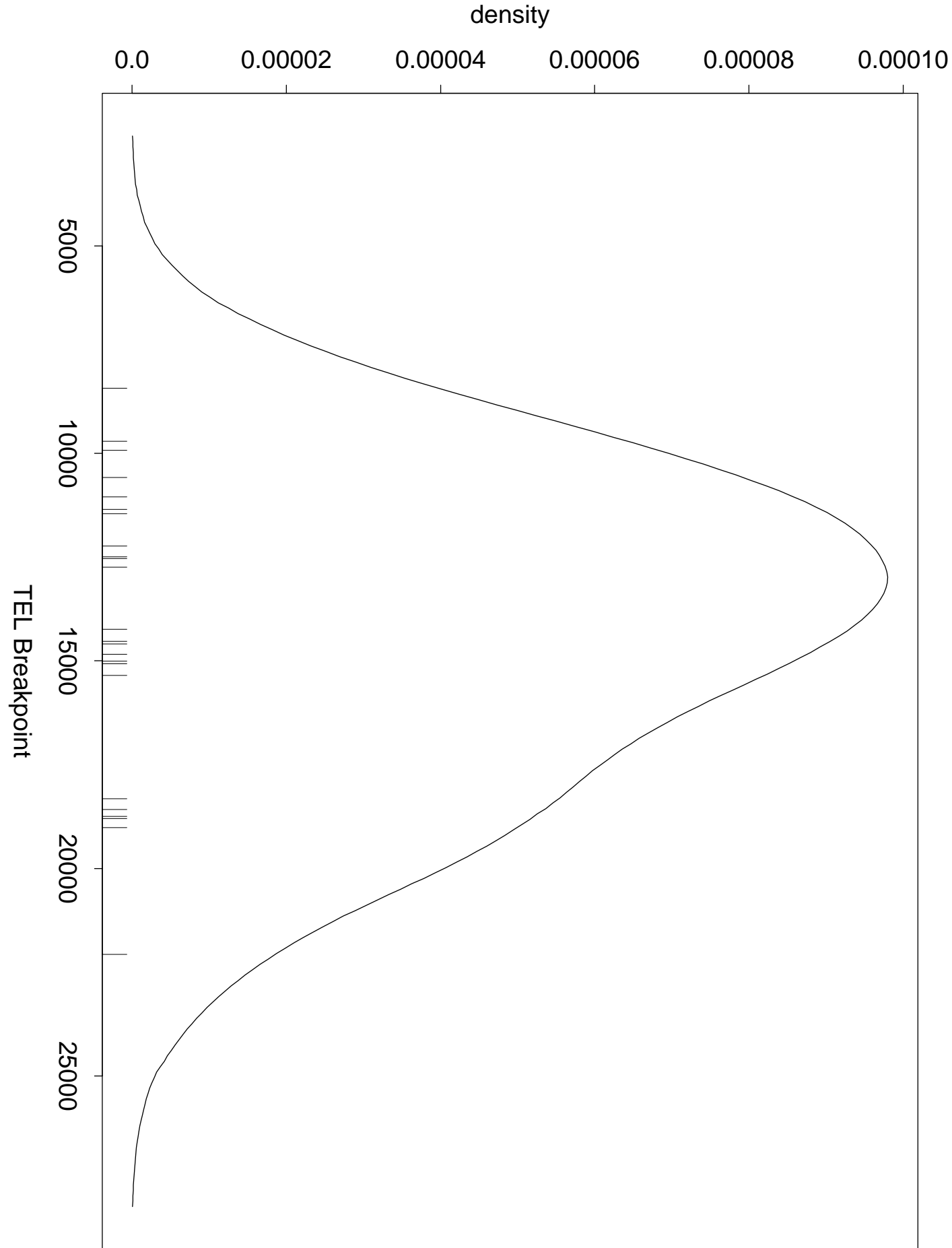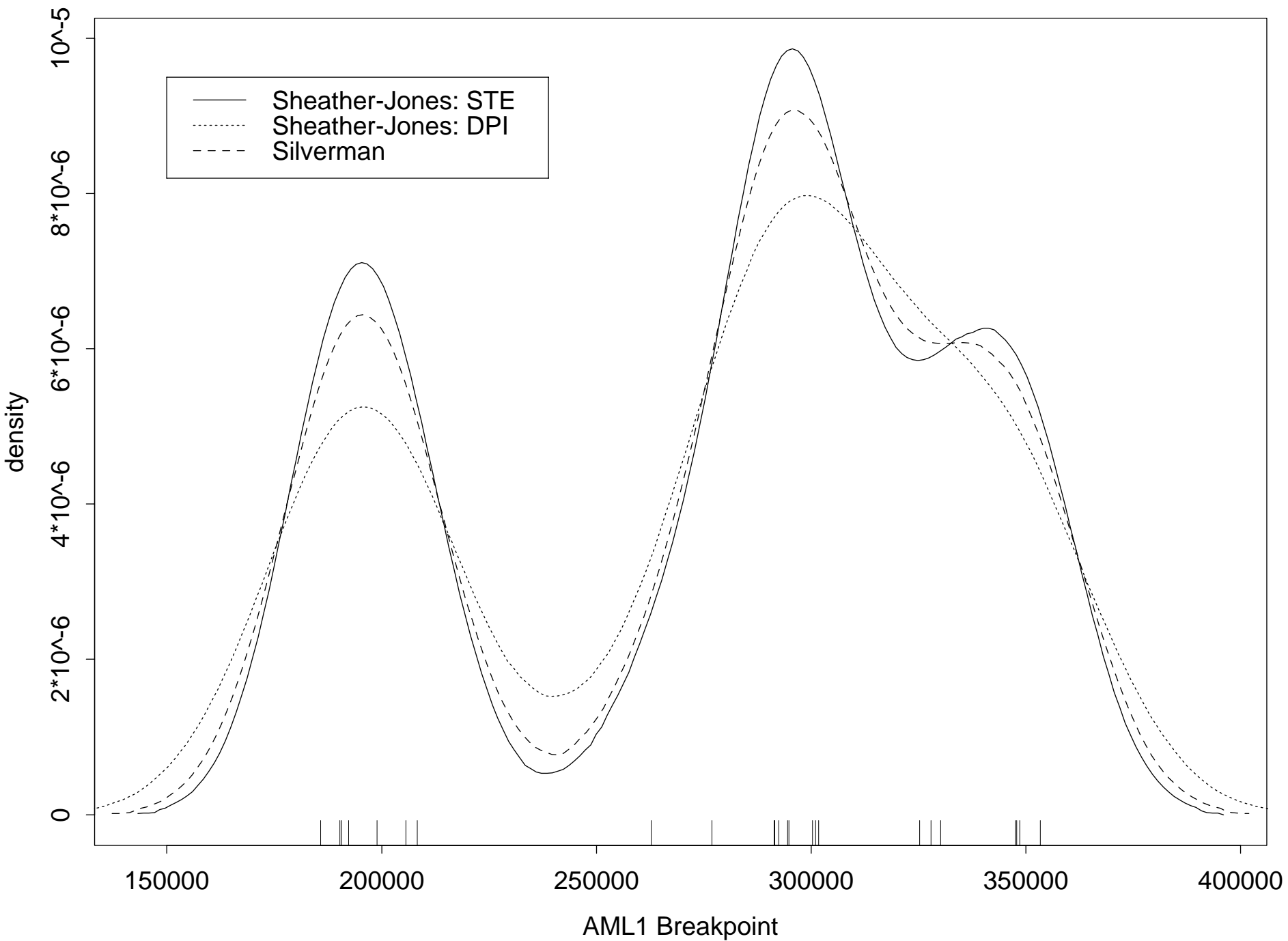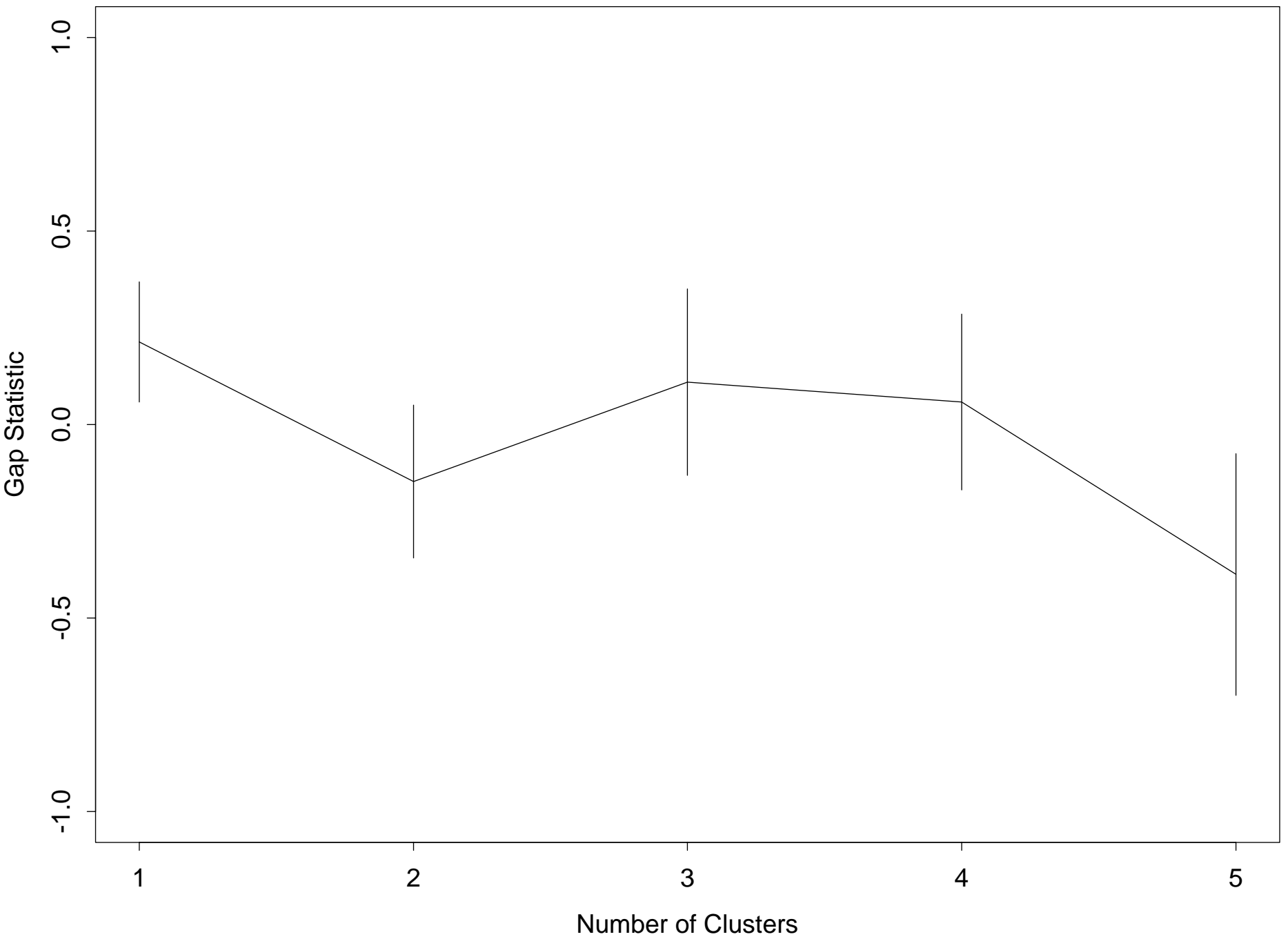
A
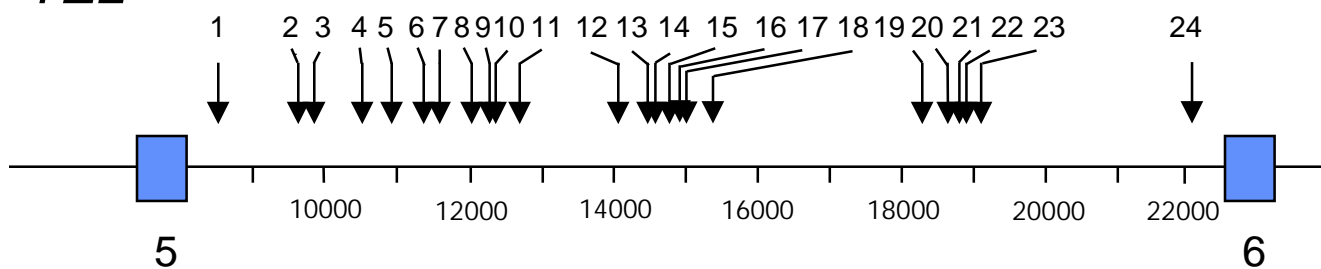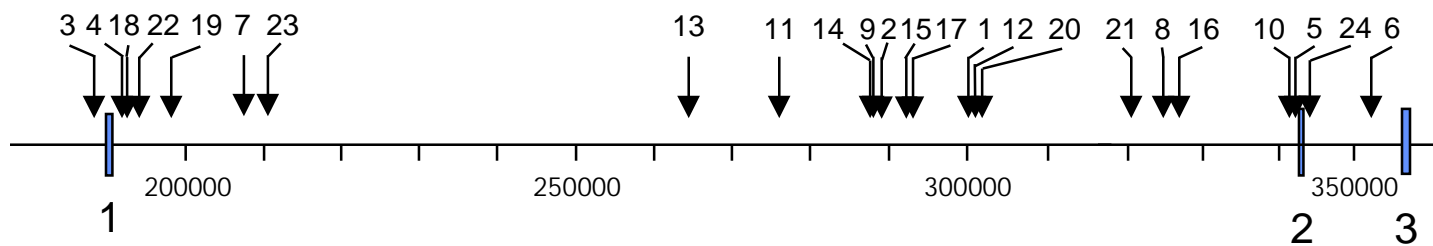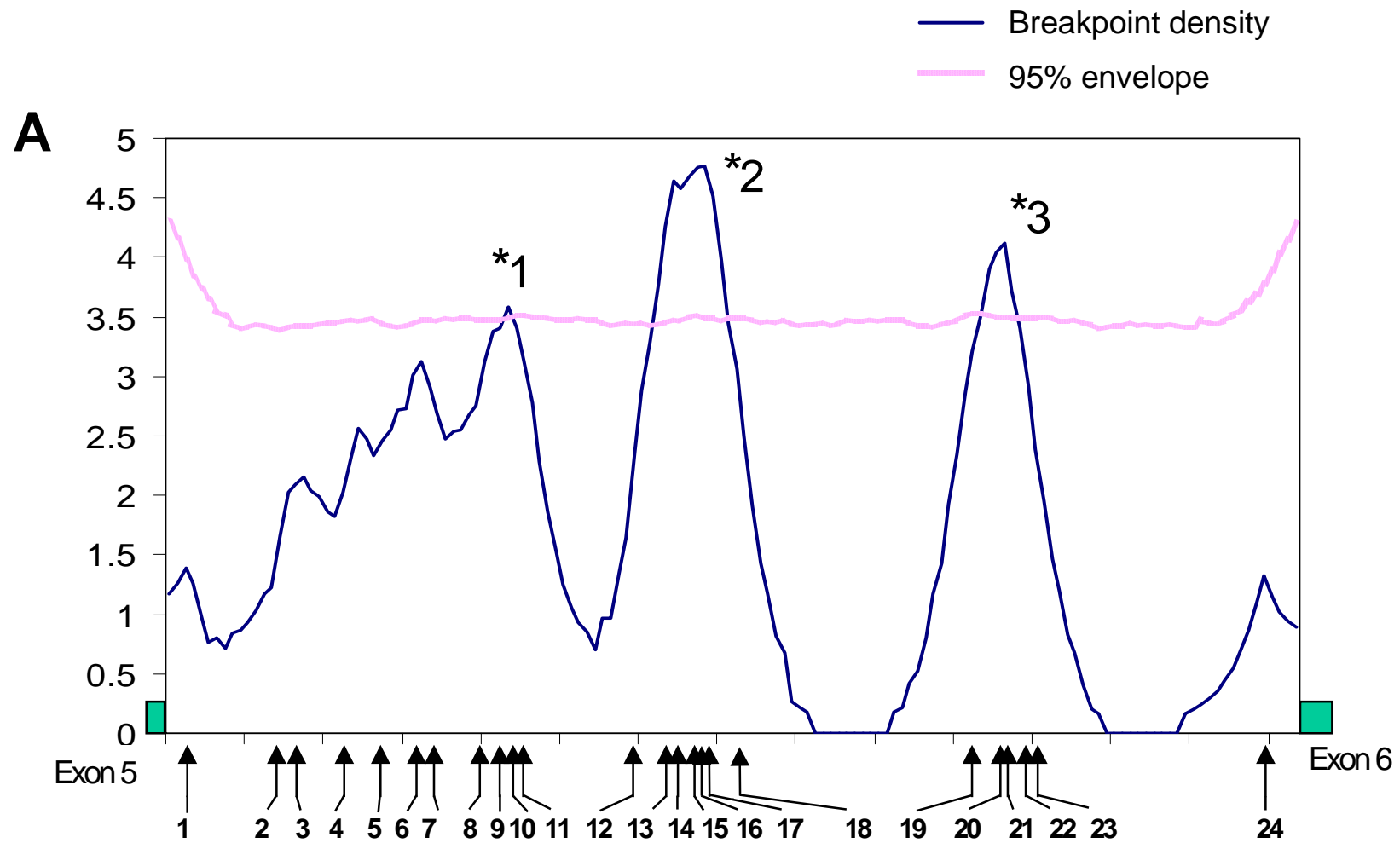
TEL Breakpoints

AML1 Breakpoints