Identification of chromosomal translocation hotspots via scan statistics

Israel T. Silva\textsuperscript{1,3,*}, Rafael A. Rosales\textsuperscript{2}, Adriano J. Holanda\textsuperscript{2}, Michel C. Nussenzweig\textsuperscript{1} and Mila Jankovic\textsuperscript{1}

\textsuperscript{1}Laboratory of Molecular Immunology, The Rockefeller University, 1230 York Avenue, New York, NY 10065, USA, \textsuperscript{2}Departamento de Computação e Matemática, Universidade de São Paulo, Av. Bandeirantes, 3900, Ribeirão Preto, CEP 14049-901 and \textsuperscript{3}National Institute of Science and Technology in Stem Cell and Cell Therapy and Center for Cell Based Therapy. Rua Catão Roxo, 2501, Ribeirão Preto, CEP 14051-140, SP, Brazil

Associate Editor: Michael Brutno

1 INTRODUCTION

The identification of genomic regions that are unusually rich in a given pattern is a recurring problem in bioinformatics, widespread in the analysis of data generated by deep-sequencing. An example of this is the detection of regions with an unlikely high clustering of chromosomal translocation breakpoints (Hasan et al., 2010; Jankovic et al., 2013; Klein et al., 2011). Recurrent chromosomal translocations are associated with hematopoietic malignancies such as leukemia and lymphoma and with some sarcomas and carcinomas (Kumar-Sinha et al., 2008; Kuppers, 2005; Kuppers and Dalla-Favera, 2001; Nussenzweig and Nussenzweig, 2010; Rabbits, 2009). There is growing evidence that translocations are not random. Among basic determinants of these events are the existence of chromosome territories, active transcription and most prominently targeted DNA damage (Chiarle et al., 2011; Hakim et al., 2012; Klein et al., 2011). DNA double-strand breaks (DSBs) are necessary intermediates in chromosome rearrangements, and they occur in the cell during normal metabolic processes, and can be induced by genotoxic agents or during physiological DNA recombination in lymphocytes. The majority of human lymphomas are of mature B cell origin, and many of them carry balanced chromosomal translocations that involve immunoglobulin genes (Kuppers, 2005). This susceptibility is most likely dependent on activation-induced cytidine deaminase (AID), the B lymphocyte-specific enzyme that initiates class switch recombination (CSR) and somatic hypermutation (SHM), two processes necessary for antibody diversification (Ramiro et al., 2006; Revy et al., 2000). AID initiates SHM and CSR by deaminating cytosines in immunoglobulin genes during stalled transcription (Chaudhuri et al., 2004; Pavri et al., 2010; Storb et al., 2007). Several DNA repair pathways process the resulting U-G mismatch to introduce mutations or produce targeted DSB (Di Noia and Neuberger, 2007; Stavnezer et al., 2008). Besides being targeted to immunoglobulin genes, AID targets a large number of non-immunoglobulin genes (Liu et al., 2008; Pavri et al., 2010; Yamane et al., 2011). AID-induced DSBs are recognized by DNA damage response proteins and repaired by non-homologous end joining, a process that can fail and result in chromosomal translocations (Gostissa et al., 2011; Nussenzweig and Nussenzweig, 2010; Zhang et al., 2010). Libraries of AID-dependent translocations from primary B...
cells revealed many discrete sites throughout the genome that are targeted by AID. Some of these targets are known translocation partners identified in mature B cell lymphomas (Chiarle et al., 2011; Klein et al., 2011). Mutations in the components of DNA repair pathways that process AID-induced breaks can lead to defective CSR, and the most severe defect is documented in 53BP1-deficient B cells. 53BP1 is a DNA repair protein that regulates DSB processing and is required for genomic stability. It does so by facilitating distal DSB joining and by protecting DNA ends from resection (Bothmer et al., 2011; Bunting et al., 2010; Difilippantonio et al., 2008). The landscape of AID-induced translocations in 53BP1-deficient B cells is different from the one found in wild-type cells. Deep sequencing of translocation capture libraries from primary 53BP1-deficient B lymphocytes has shown that the profile of translocation hotspots changes most likely owing to increased DNA end resection (Jankovic et al., 2013).

A quantitative method to determine the clustering of translocations was essential for the analysis of chromosomal rearrangements in Klein et al. (2011) and Jankovic et al. (2013). Translocation hotspots were determined by using a technique similar to that used to define the coordinates of enriched protein-binding regions in ChiP-seq experiments. A translocation cluster is defined by concatenating closely spaced adjacent breakpoints, and its significance is then determined by using a test based on the negative binomial distribution. This method assumes that the observed breakpoints are realization of a Bernoulli process. By taking advantage of this model, here we consider a different approach for the detection of hotspots based on the use of scan statistics (Balakrishnan and Koutras, 2001; Glaz et al., 2001). The scan statistic is well suited for this task because it provides a genome-wide level to breakpoint clustering. Using our method, we are able to show that translocation hotspots induced by AID in activated B lymphocytes are longer than those previously identified by the local method. Furthermore, our method shows that long hotspots are more frequent in the absence of 53BP1. The frequency of the long hotspots is further increased if AID is overexpressed in 53BP1-deficient B cells. We also discover a set of hotspots exclusively depleted in 53BP1-deficient B cells. 53BP1-deficient CSR, and the most severe defect is documented in 53BP1-deficient B cells. The significance of a cluster of translocation events in a window of width $m$ can be assessed by the probability of the tail event $S_m \geq y$.

Several methods have been developed to detect clustering of events when the observations arise from a spatial or temporal point process. This section describes the application of the scan statistic for the detection of genomic regions with a particularly high density of translocations. This approximation is ensured by weak convergence of the hypergeometric distribution toward the binomial law for large populations and becomes accurate in the current application where $N > 1 \times 10^4$ and $m \geq 500$. Furthermore, for large values of $a$, namely, for $1000 \leq a \leq 10000$, as is the case for most chromosomes in the datasets considered here, the summation in (2) may be approximated by

$$
\hat{p} = \frac{1}{\sqrt{2\pi\hat{m}}} e^{-y^2/2\hat{m}}
$$

where $\hat{m}$ is the maximum number of successes within any of the $N - m + 1$ consecutive windows,

$$
S_m = \max_{1 \leq i \leq N-m+1} \{ Y_i \}.
$$

The significance of a cluster of translocation events in a window of width $m$ can be assessed by the probability of the tail event $S_m \geq y$. Small probabilities for this event indicate departures for the Bernoulli model consistent with $H_0$ and could therefore be used to detect hotspots. A considerable effort has been made to derive the distribution of the number of observed events in $N$ trials, $a$, are known. Following Naus (1974) and Glaz et al. (2001), the conditional probability $P(S_m \geq y \mid a)$ may be approximated by the function

$$
\phi(y) = \frac{1}{2} \sum_{i=0}^{a} \binom{y}{i} \binom{N - y}{a - i} \binom{N - m - a + 1}{m - i} H(y; a, m, N),
$$

with $H(y; a, m, N)$ as the hypergeometric distribution,

$$
H(y; a, m, N) = \binom{m}{y} \binom{N - m}{a - y} \binom{N}{a}.
$$

Although the expression in (1) already provides a method to quantify the significance of a cluster, the following asymptotic version allows for an efficient implementation. For sufficiently large $m$ and $N$, the function in (1) may be approximated by

$$
\phi(y) \approx \frac{1}{2} \sum_{i=0}^{a} b(i; a, \theta) = \binom{m}{y} \binom{N - m - a + 1}{y} b(y; a, \theta),
$$

where $b(y; a, \theta)$ denotes the Binomial distribution for $a$ trials and success probability $0 < \theta = m/N < 1$. This approximation is ensured by weak convergence of the hypergeometric distribution toward the binomial law for large populations and becomes accurate in the current application where $N > 1 \times 10^4$ and $m \geq 500$. Furthermore, for large values of $a$, namely, for $1000 \leq a \leq 10000$, as is the case for most chromosomes in the datasets considered here, the summation in (2) may be evaluated as

$$
\sum_{i=0}^{a} b(i; a, \theta) = b(y; a, \theta) \sum_{i=0}^{\infty} \binom{m}{y} \binom{N - m - y + 1}{\theta} \frac{\theta^{-i}}{\theta^{-i}},
$$

with $\sum_{i=0}^{\infty} \frac{\theta^{-i}}{\theta^{-i}} = \sum_{i=0}^{\infty} \frac{\theta^{-i}}{\theta^{-i}} = 1$, and $(n)_i$ as the $i$th Pochhammer symbol of $n \in \mathbb{N}$, i.e., $(n)_i = n(n+1) \cdots (n+i-1)$. Note that the second argument of $\sum_{i=0}^{\infty}$ is always negative or zero because $y \leq a$. The series defining $\sum_{i=0}^{\infty}$ is thus finite. With this simplification, the desired P-value

$$
\hat{p} = P[S_m \geq y \mid a]
$$

is approximately

$$
\hat{p} \approx b(y; a, \theta) \left\{ 2 \sum_{i=0}^{\infty} \binom{m}{y} \binom{N - y + 1}{a - i} \right\}.
$$

We observe that (2) is the approximation for the probability of $[S_m \geq y \mid a]$ described by Wallenstein and Neff (1987) in the
well-known continuous case, namely, when \(N\) points are drawn uniformly from \([0, 1]\), and \(S_m\) is the largest number of points to be found in any subinterval of \([0, 1]\) of length \(m\). Despite the existence of several other approximations for the probability of \(S_m \geq y|a\), Glaz (1989) observes that this approximation is precise when the right side of (2) \(\leq 0.01\) and recommends its use in this regime.

The detection of chromosomal translocation breakpoints via scan statistics has also been considered by several authors in the analysis of leukemia (Berger et al., 2013; Busch et al., 2007; Hasan et al., 2010; Reiter et al., 2003; Wielmels et al., 2002). These analyses are based on the method described by Segal and Wielmels (2002), by following a large deviation approximation for the probability of \(S_m \geq y\) described in Loader (1991). Although being derived by using rather different arguments, Naas and Wallenstein (2004) observe that this approximation and the one in (1) produce similar results.

### 2.2 Hotspots

The procedure outlined in Section 2.1 provides a significance test for the existence of hotspots, still their actual number and location have to be determined. Here we describe a method to infer the coordinates of these events.

A chromosome-wide scan with a window of width \(m\) leads to the consideration of the following sequence of local null hypotheses. For \(i=1, \ldots, N-m+1\),

\[
H_{0i}: \xi_i, \ldots, \xi_{i+m}
\]

are i.i.d. Bernoulli(\(p\)) random variables with \(\xi_i = p\).

Let \(y_i\) be the observed number of translocation events in the \(i\)th window, \(w_i\). The hypothesis \(H_{0i}\) is rejected at a prescribed level \(a_H\) if \(p_i = P[S_m \geq y_i|a] \leq a_H\), with \(p_i\) computed according to (3). This significance criterion conditionally supports the one-sided alternative

\[
H_{ai}: \xi_i, \ldots, \xi_{i+m}
\]

are i.i.d. Bernoulli(\(p\)) random variables with \(\xi_i > p\)

against the null, \(H_{0i}\).

This procedure partitions the chromosome into two regions:

\[
\Psi = \{w_i : p_i \leq a_H, i \in [1, N-m+1]\} \text{ and } \Psi^v.
\]

The connected components of \(\Psi\) are prospective hotspot candidates formed by one or more scanning windows of width \(m\). To actually account for the bias involved while considering the simultaneous rejection of the multiple hypotheses involved in a given component of \(\Psi\), we adjusted the corresponding \(P\)-values by using the Benjamini–Yekutieli correction (Benjamini and Yekutieli, 2001). This correction accounts for the possibility of having a positive dependence structure among the considered set of hypotheses from overlapping scan windows. A similar control of the false discovery rate associated with a large number of tests produced by the scan statistic has previously been considered while scanning for clusters in random fields by Perone Pacifico et al. (2007). Denote by \(p_i^*\) the adjusted \(P\)-value for the \(i\)th window, so that for the control level \(a_H\), the corrected \(P\)-values define set

\[
\Psi^v = \{w_i : p_i^* \leq a_H, i \in [1, N-m+1]\} \subseteq \Psi.
\]

The inclusion follows because \(p_i^* \leq p_i\). Depending on the value of \(m\), each element of \(\Psi^v\) may end on a translocation event or not. In the latter case, the extra bases starting after the last translocation event are deleted. Let \(\Psi^v\) be the remaining connected regions in \(\Psi^v\) after trimming, consisting of segments that start and end with a translocation breakpoint event. Let \(\theta\) be the length of a generic element of \(\Psi^v\), which may be thought as being obtained while scanning with the scan statistic \(S_{\theta}\). Following Section 2.1, the \(P\)-value of this segment is then computed by using (3) and taking \(\theta = L_i\).

The method just described proceeds in two steps. The first one detects possible hotspot regions while scanning with \(S_m\) for \(m\) and \(a_H\) fixed. Most of the resulting events are then shortened in such a way to end at a breakpoint at a second stage. The resulting segment of final length \(\ell\) may be thought of as being elicited with the scan statistic \(S_{\ell}\). This segment is finally classified as a hotspot if its \(P\)-value is \(\leq a_H\). Hotspots were defined by taking \(a_H = 0.05\) and by using several initial window widths.

We denote hereafter by \(SS_h\) the procedure based on the scan statistic with window \(m\) (in base pairs).

### 2.3 A local approach to hotspot detection

The probabilistic model for the occurrence of translocations described in Section 2.1 is implicit in previous work made by Klein et al. (2011) and Jankovic et al. (2013) while analyzing hotspots. The data consisting of the genome translocation breakpoints are represented as a Bernoulli process with success probability \(p\), estimated as \(a/N\) with \(N\) as the genome length and \(a\) as the total number of observed translocation events.

Suppose \((x_i), i=1, \ldots, a\), are the coordinates of the translocations and let \(L_i\) for \(i=1, \ldots, a-1\) be the number of bases between \(x_i+1\) and \(x_{i+1}\) inclusive. The random variable \(L_i\) records therefore the length until the next translocation starting at \(x_i+1\), namely, \(L_i = x_{i+1} - x_i - 1\). The independence of the underlying Bernoulli process implies that \(L_i\) is a geometric random variable with parameter \(P\), i.e. \(P(L_i = \ell) = (1 - p)\theta^{\ell-1}p\), \(\ell \geq 1\). Small values for

\[
P(L_i \leq \ell_i) = 1 - (1 - p)^{\ell_i}\tag{4}
\]

may thus be used to detect unusual short distances between successive translocation events. In this sense, a hotspot can be defined by concatenating adjacent segments for which \(P(L_i \leq \ell_i) \leq \alpha,\) where \(\alpha\) is a given significance level specified in advance. Suppose that a given sequence of adjacent segments of widths \(\ell_1, \ell_1 + 1, \ldots, \ell_1 + L\) is identified as a hotspot. Let

\[
L'_i = \sum_{j=0}^{i} L_{i+j} \text{ and } \ell'_i = \sum_{j=0}^{i} \ell_{i+j},
\]

so that the significance of this hotspot may be quantified by the \(P\)-value

\[
P = P(L'_i \leq \ell'_i).
\]

This probability is directly available because \(L'_i\) is a negative binomial random variable with parameters \(r+1\) and \(a/N\), i.e.

\[
p = \sum_{k=r+1}^{\infty} \frac{k-1}{r} \frac{(a/N)^r (1 - a/N)^{k-r}}{\Gamma(k)},
\]

This method was used by Klein et al. (2011) and Jankovic et al. (2013) to define a set of potential hotspots by taking \(a_H = 0.01\). Any candidate of this set is then identified as a hotspot if

(i) it has more than three translocation breakpoints,

(ii) it has at least one read from each of the two sides of the bait,

(iii) at least 10% of the translocations come from each side of the bait,

(iv) \(p \leq 1 \times 10^{-8}\).

Hereafter, we refer to this procedure as the local method and denote it by \(N_{\Psi}\). We describe results obtained with \(N_{\Psi,0.01}\) and \(N_{\Psi,0.05}\).

### 2.4 TC-Seq and ChiP libraries

The TC-Seq datasets analyzed here are those described by Klein et al. (2011) and Jankovic et al. (2013). These are deposited at Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra) under accession numbers SRA061477 and SRA039959. These datasets are from four different translocation libraries: (i) a library from activated B cells infected with AIDS-expressing retrovirus (denoted hereafter as AID\(Y\)), (ii) a library from AID-deficient B cells (denoted as AID\(D^{-}\)), (iii) a library from...
3 RESULTS

3.1 Scan statistic and local method: simulated data

A fair amount of hotspots simulated as described in Section 2.5 was partially identified by hot_scan and the local method. Differences are due to single breakpoint events that are classified as being part of a hotspot or a non-hotspot region. A finer description about the performance of each method may thus be obtained by studying the ability of correctly classifying each breakpoint. To this end, we simulated 100 chromosomes, each having on average ~20 hotspots according to both conditions in Section 2.5. The results found with the local method and hot_scan at various window widths are summarized in Figure 1. Figures 1A and B present the Receiver operating characteristic (ROC) curves for the classification of breakpoints into hotspot and non-hotspot regions by varying the nominal control level from 0.0001 to 0.5. Both methods are specific, as reflected by the small false-positive rates. This is owing to the fact that both methods correctly classify almost all non-hotspot breakpoints as such. As a consequence, the observed ROC curves raise rapidly. An analysis of the true-positive rate shows that the performance of hot_scan may be superior, but this depends on the scanning window width. For the easiest datasets (Fig. 1A), both procedures show similar true-positive rates. This is revealed by the open circles that correspond to the ROC values at the 0.05 level for NB, SS1000, SS2500 and SS5000. At these window widths, hot_scan presents a higher rate of true positives for the hardest datasets (Fig. 1B). On the other hand, windows as small as 500 bp present a much smaller true-positive rate than the local method for both data conditions. The graphs in Figures 1C and D show that both techniques are able to control the type I error rate at the nominal level. Interestingly, hot_scan method presents lower error rates than the local procedure for all window widths considered, except for the 5000 window at small control levels (Fig. 1D).

3.2 Scan statistic and local method: real data

The methods described in Sections 2.2 and 2.3 are compared by plotting the distribution of the hotspot lengths identified in real data samples. Because the observed hotspot lengths vary across several orders, we considered the logarithm of their actual length. The results obtained by analyzing the four datasets described in Section 2.4 are presented in Figure 2. The analysis done with SS5000 shows that the hotspot length distribution is roughly described by two components, one with a mean length of $167 = [e^{0.12}]$ base pairs and the other with mean length equal to $4154 = [e^{4.32}]$. Supplementary Table S1 presents the means.
variances and the weights of these components. Let $0 < \gamma < 1$ and $1 - \gamma$ be the weight of the short and long hotspots components, respectively. While the mean position of these components remains almost the same, the relative weight of long to short hotspots, $\rho = \gamma/(1 - \gamma)$, does show significant changes. A comparison of the AID$^\gamma$ dataset (Fig. 2D) and the 53BP1$^-/^-$AID$^\gamma$ dataset (Fig. 2B) reveals that the relative frequency of long hotspots is higher in the absence of 53BP1. Indeed, the value for 53BP1$^-/^-$AID$^\gamma$ sample, it is four times smaller, namely, $\rho = 0.51$. We conclude that 53BP1 decreases the proportion of long hotspots. A similar effect of 53BP1 deficiency is observed in the absence of AID. The 53BP1$^-/^-$AID$^-/^-$ sample (Fig. 2A) is characterized by $\rho = 1.08$, while the AID$^-/^-$ sample (Fig. 2C) by $\rho = 0.3$. We conclude that in the absence of 53BP1, longer hotspots are more frequent regardless of AID expression. This effect can be attributed to the role of 53BP1 in DNA end protection. In the absence of this protein DNA, end resection is increased, resulting in longer hotspots as suggested previously (Jankovic et al., 2013). A comparison of the plots that present the AID$^-/^-$ and AID$^\gamma$ samples (Figs 2C and D, respectively) shows no substantial changes in the proportion of short to long hotspots, with $\rho = 0.3$ and $\rho = 0.5$. However, in the absence of 53BP1, the frequency of longer hotspots increases significantly when AID is overexpressed (Figs 2A and B). Thus, proper DNA repair that is dependent on 53BP1 ensures the predominance of short hotspots, even when AID is overexpressed. Figure 2 shows that all these features are observed while scanning with windows of width 2000, 2500 and 5000, where hot_scan is shown to be more powerful (Fig. 1).

Most of these results are not observed when analyzing the same data by the local method described in Section 2.3. This becomes clear by inspection of the dashed lines in Figure 2, which correspond to the length distributions for hotspots detected by NB$^0.01$ and NB$^0.05$. Even at $\alpha_C = 0.05$, for which one would expect longer hotspots, the local method is unable to detect the changes in the frequency of the long hotspots to the extent brought by hot_scan. It is important to note that by following (4), the local method would classify two consecutive breakpoints as being part of a hotspot if their distance is smaller than $\ln(1 - \alpha_C)/\ln(1 - \rho)$. Using larger values for $\alpha_C$ allows thus for larger gaps. Values $>0.05$ would, however, correspond to tests with type I errors higher to what is commonly acceptable.

The results in Figure 2 present the differences in the hotspot lengths defined by the scan statistic and the local method. However, they do not provide any information about the relative positions of the hotspots detected by either technique. To address this aspect, we analyzed the relative hotspot positions for all four datasets described in Section 2.4. As an example, Figure 3 presents the hotspots for chromosome 9 estimated via NB$^0.01$, SS$^{2500}$, SS$^{5000}$ and SS$^{10000}$, where NB$^0.01$, SS$^{2500}$ and SS$^{5000}$ are shown as continuous lines, whereas those for NB$^0.05$ and NB$^0.01$ as dashed lines. Each graph corresponds to a (Gaussian) kernel density estimate of the underlying distribution.

![Fig. 1. A and B: breakpoint classification ROC curves for synthetic data resembling the AID$^\gamma$ and 53BP1$^-/^-$AID$^\gamma$ conditions, respectively. Results for the local method are identified as NB, and those for hot_scan according to window width. C and D: type I error control for the same datasets in A and B. The same color schemes used in A and B are also used for C and D. Open circles in A and B correspond to the ROC values found at the 0.05 control level (these are only included for NB and SS$^{2000}$, SS$^{5000}$ and SS$^{10000}$).](image1)

![Fig. 2. Hotspot length distributions. A presents the distributions for the 53BP1$^-/^-$AID$^-/^-$ sample, B for 53BP1$^-/^-$AID$^\gamma$, C for AID$^-/^-$ and D for AID$^\gamma$. The distributions for the hotspots determined with SS$^{2500}$, SS$^{5000}$ and SS$^{10000}$ are shown as continuous lines, whereas those for NB$^0.05$ and NB$^0.01$ as dashed lines. Each graph corresponds to a (Gaussian) kernel density estimate of the underlying distribution.](image2)
hotspots found by NB$_{0.01}$. The merging of several smaller hotspots into a larger one is justified by the sparsity of the data, which only becomes apparent at larger scales. These features are clearly overseen by the NB method because of its local nature.

Few examples of the scaling effect are shown by the examples in Figures 3E, G and H. These results are consistent with those observed for other chromosomes (Supplementary Figs S1 and S2).
Overexpression of AID in the absence of 53BP1 results not only in the increase in the number of translocation but also defines larger regions where these events cluster. In addition, AID overexpression in 53BP1−/− cells results in elongation of pre-existing hotspot regions. This is apparent when comparing the outermost track and the neighboring one on the circular graph that corresponds to the analysis with SS500 (Fig. 3C) for the 53BP1−/− AID−/− and the 53BP1−/− AID+ data. The analysis of the same data with a larger scanning window, namely, with SS2500 and SS5000, gives a similar result, but the affected regions are much larger (Figs 3D and B). The length of most hotspots in this situation is greatly reduced in 53BP1 sufficient samples. This is apparent for the hotspots from the 53BP1−/− AID+ and the AID+ data. Interestingly, the effect of 53BP1 correlates with the significance of the hotspots.

We conclude that hotspot length is dependent on 53BP1, and that AID overexpression in the absence of 53BP1 results in translocations that cluster over large regions.

### 3.3 Exclusive hotspots

Most of the more prominent hotspots are defined by both the scan statistic and the local method (Figs 3A and C and Supplementary Figs S1A, C and S2A and C). However, both methods reveal exclusive clustering regions (Supplementary Tables S2 and S3). To identify AID-dependent hotspots that are exclusive to each method, we compared the AID+ and the AID− data. We found 36 exclusive hotspots with hot_scan (Supplementary Table S2) and 27 exclusive hotspots with NB0.05 and NB0.01 (Supplementary Table S3). The exclusive hotspots obtained by the scan statistic were defined using different window widths (50, 100, 150, 250, 500, 1000, 2500 and 5000 bp). Regions that are identified as exclusive AID hotspots were also analyzed for several biologically relevant markers. First, we analyzed whether our exclusive hotspots correlate with replication protein A (RPA) binding sites in activated B cells (Hakim et al., 2012; Yamane et al., 2011). The sites of RPA accumulation have been shown to overlap well with AID targets genome-wide, and it was proposed that RPA marks AID-induced DNA-strand breaks. Further, we analyzed the overlap with sites where RNA Polymerase II (PolII) accumulates, as it was shown that transcription is necessary for AID targeting (Di Noia and Neuberger, 2007). We also analyzed the overlap with known fragile regions, namely, by the early replicating fragile sites and common fragile sites (CFS) (Barlow et al., 2013). The results of all these comparisons are summarized in Supplementary Tables S2 and S3. A total of 20 of the 36 (55.5%) exclusive hotspots found by the scan statistic were common to all sites. Notably, all of these sites are associated with the PolII signal (Supplementary Table S2). On the other hand, only 8 of the 27 (29.6%) exclusive hotspots of the local method fall within these sites, and 6 are associated with the PolII signal (Supplementary Table S3). Thus, the hotspots defined by the scan statistic show higher correlation with active transcription, RPA accumulation and CFS than those defined by the local method.

AID leads to the accumulation of somatic mutations in a large number of non-immunoglobulin genes (Nussenzweig and Nussenzweig, 2010). An analysis for the presence of SHMs in 1496058 bp from activated B cells (Yamane et al., 2011) revealed a number of non-immunoglobulin genes with AID-dependent mutations: Il4ra, Grap, Hist1h1c, Ly6e, Gadd45g and I4i1. Three of these, namely, Il4ra, Grap and Ly6e, were detected as genes with AID-dependent hotspots by both methods, but a hotspot in Hist1h1c (mutation rate in Igx-AID Ung−/−: 7.97 x 10−5) was only found by hot_scan (Supplementary Table S2 and Supplementary Fig. S3). Three other genes associated with chromosomal translocations were detected exclusively by hot_scan, namely, Fill, Dlx5 and Birc3. The Fill (Friend leukemia integration 1) gene (Fig. 3F) is translocated in 90% of Ewing sarcomas and is important in tumorigenesis (Riggi and Stamenkovic, 2007). Dlx5 (distal-less homeobox 5) is implicated in T-cell lymphomas (Tan et al., 2008). Finally, the Birc3 (baculoviral IAP repeat-containing 3) gene encodes an apoptosis inhibitor that is associated with MALT lymphomas (Dierlamm et al., 1999). A complementary enrichment analysis (Wang et al., 2013) for the genes identified by hot_scan is included in Supplementary Figures S5 and S8 and Supplementary Tables S4 and S5. The functional categories associated with the scan statistic hotspots indicate that the top ranked genes are important in B lymphocytes.

### 4 DISCUSSION

Here we describe a method for the identification of chromosomal translocation hotspots. In contrast to a previous procedure, which we refer here to as the local method, the control level for the detection of a cluster is defined on a chromosome-wide basis by using scan statistics. We show via simulations that scan statistics perform equally well as the local method in more challenging situations, characterized by a higher translocation rate outside hotspot regions. This depends on the width of the scanning window, and its choice requires some calibration. We present a method that is able to accomplish this.

We show that inferences made with scan statistics have important consequences in the analysis of translocation hotspots in primary B cells. The previous study by Jankovic et al. (2013) showed that 53BP1 deficiency results in an increase of rearrangements to intergenic regions and changes the frequency and distribution of translocations in γδ, γ1 immunoglobulin switch regions and other 16 prominent hotspots. Our analysis adds to these findings by showing that the 53BP1 deficiency results in the overall enrichment of longer hotspots. These results support the previous conclusion that 53BP1 prevents the resection of DNA, thus resulting in shorter hotspots (Jankovic et al., 2013). Our analysis here also shows that an increased amount of AID results in a substantial enlargement of pre-existing hotspot regions. These changes can only be observed with wider scanning windows (i.e with w = 2000, 2500 and 5000), and are not detected by previous methods because of their local characterization of clustering. The success of the scan statistic here is brought by its ability to detect events spread across several scales as is shown by the analyses made with several scanning window widths. Our analysis with the scan statistic is able to identify several exclusive hotspots whose authenticity is supported by independent
experimental approaches. Some of these exclusive events are localized in genes that are known to be relevant in tumorigenesis.

The approach presented here may be applied to a variety of questions related to the detection of unusual clustering of a given pattern throughout the genome. Few recent examples of particular interest are the detection of enriched genomic interaction regions such as those defined via ChiP-seq experiments (Ma and Wong, 2011), 4C-seq experiments (Simonis et al., 2006) and DNA-DNA contact sites (de Wit and de Laat, 2012). We expect our method to be especially useful for the analysis of data where a global significance to clustering can be considered.

ACKNOWLEDGEMENTS
I.T.S. wishes to thank T. Oliveira for kindly providing the script for the local method described in Section 2,3, and R.A.R. thanks K.J. Abraham for useful discussions. M.C.N. is a Howard Hughes Medical Institute Investigator. We are also grateful to the anonymous referees for helpful comments and suggestions.

Funding: The work was supported by a NIH grant to M.C.N., number AI037526.

Conflict of Interest: none declared.

REFERENCES