BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 18 2014, pages 2551–2558
doi:10.1093/bioinformatics/btu351

Genome analysis

Advance Access publication May 23, 2014

Identification of chromosomal translocation hotspots via scan statistics

Israel T. Silva[1,3,*], Rafael A. Rosales[2], Adriano J. Holanda[2], Michel C. Nussenzweig[1] and Mila Jankovic[1]

[1]Laboratory of Molecular Immunology, The Rockefeller University, 1230 York Avenue, New York, NY 10065, USA, [2]Departamento de Computação e Matemática, Universidade de São Paulo. Av. Bandeirantes, 3900, Ribeirão Preto, CEP 14049-901 and [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 Brudno

ABSTRACT

Motivation: The detection of genomic regions unusually rich in a given pattern is an important undertaking in the analysis of next-generation sequencing data. Recent studies of chromosomal translocations in activated B lymphocytes have identified regions that are frequently translocated to c-myc oncogene. A quantitative method for the identification of translocation hotspots was crucial to this study. Here we improve this analysis by using a simple probabilistic model and the framework provided by scan statistics to define the number and location of translocation breakpoint hotspots. A key feature of our method is that it provides a global chromosome-wide nominal control level to clustering, as opposed to previous methods based on local criteria. While being motivated by a specific application, the detection of unusual clusters is a widespread problem in bioinformatics. We expect our method to be useful in the analysis of data from other experimental approaches such as of ChIP-seq and 4C-seq.

Results: The analysis of translocations from B lymphocytes with the method described here reveals the presence of longer hotspots when compared with those defined previously. Further, we show that the hotspot size changes substantially in the absence of DNA repair protein 53BP1. When 53BP1 deficiency is combined with overexpression of activation-induced cytidine deaminase, the hotspot length increases even further. These changes are not detected by previous methods that use local significance criteria for clustering. Our method is also able to identify several exclusive translocation hotspots located in genes of known tumor supressors.

Availability and implementation: The detection of translocation hotspots is done with `hot_scan`, a program implemented in R and Perl. Source code and documentation are freely available for download at https://github.com/itojal/hot_scan.

Contact: isilva@rockefeller.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 4, 2013; revised on May 8, 2014; accepted on May 13, 2014

*To whom correspondence should be addressed.

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; Rabbitts, 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 found by the scan statistic and discuss the potential biological relevance of our findings.

2 METHODS

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. We also describe explicitly a technique previously used by Klein *et al*. (2011) and Jankovic *et al*. (2013), which attributes a local significance level to a hotspot. A major difference between these methods is that the scan statistic provides a global chromosome-wide control level to clustering.

2.1 Scan statistic approximations

We model the occurrence of translocation breakpoints in a chromosome of length N as realization of an independent and identically distributed sequence of 0-1 Bernoulli random variables, $\xi_1, \ldots, \xi_N$ with $\mathbb{P}\{\xi_j = 1\} = p$ and $\mathbb{P}\{\xi_j = 0\} = 1 - p$, $1 \le j \le N$ for $0 < p < 1$. The event $\{\xi_j = 1\}$ occurs if there is a translocation breakpoint at the j-th base. We refer hereafter to this model as the global chromosome null hypothesis, H_0. A study of the adequateness of H_0 as a description of the datasets considered throughout is included as Supplementary Material. Let m be a positive integer and then let $Y_i = \sum_{j=i}^{i+m-1} \xi_j$, $1 \le i \le N - m + 1$, be the running number of successes in a window of width m. The scan statistic S_m is defined as the maximum number of successes within any of the $N - m + 1$ consecutive windows,

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

The significance of a cluster of translocation events in a window of width m can be assessed by the probability of the tail event $\{S_m \ge y\}$. 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 S_m under H_0. Still in this simple case, its form has remained elusive. Several approximations and asymptotics for the distribution of the scan statistic have been derived under the Bernoulli model, particularly when the number of observed events in N trials, a, are known. Following Naus (1974) and Glaz *et al*. (2001), the conditional probability $\mathbb{P}\{S_m \ge y | a\}$ may be approximated by the function

$$\varphi(y) = 2\sum_{i=y}^{a} H(i; a, m, N) + \left(y\frac{N}{m} - a - 1\right) H(y; a, m, N), \qquad (1)$$

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

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

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

$$\varphi(y) \approx 2\sum_{i=y}^{a} b(i; a, \theta) + \left(y\frac{N}{m} - a - 1\right) b(y; a, \theta), \qquad (2)$$

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 $N > 1 \times 10^6$ and $m \ge 500$. Furthermore, for large values of a, namely, for $1000 \le a \le 10000$, as is the case for most chromosomes in the datasets considered here, the summation in (2) may be evaluated as

$$\sum_{i=y}^{a} b(i; a, \theta) = b(y; a, \theta)\ {}_2F_1\left(1, y - a; 1 + y; \frac{\theta}{\theta - 1}\right),$$

with ${}_2F_1$ as Gauss hypergeometric function, that is, for $|z| < 1$ and n_1 n_2, $n_3 \in \mathbb{Z}$,

$${}_2F_1(n_1, n_2; n_3; z) = \sum_{i=0}^{\infty} \frac{(n_1)_i (n_2)_i}{(n_3)_i} \frac{z^i}{i!}$$

and $(n)_i$ as the ith Pochhammer symbol of $n \in \mathbb{N}$, i. e. $(n)_i = n(n+1)\cdots(n+i-1)$. Note that the second argument of ${}_2F_1$ is always negative or zero because $y \le a$. The series defining ${}_2F_1$ is thus finite. With this simplification, the desired P-value

$$\hat{p} = \mathbb{P}\{S_m \ge y \mid a\}$$

is approximately

$$\hat{p} \approx b(y; a, \theta)\left\{2\ {}_2F_1\left(1, y - a; 1 + y; \frac{\theta}{\theta - 1}\right) + \left(y\frac{N}{m} - a - 1\right)\right\}. \qquad (3)$$

We observe that (2) is the approximation for the probability of $\{S_m \ge y \mid a\}$ 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) ≤ 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 Wiemels (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, Naus 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 $i = 1, \ldots, N - m + 1$,

$$H_{0,i}: \ \xi_i, \ldots, \xi_{i+m} \quad \begin{array}{l}\text{are i.i.d. Bernoulli}(\varepsilon_i) \\ \text{random variables with } \varepsilon_i = p.\end{array}$$

Let y_i be the observed number of translocation events in the ith window, w_i. The hypothesis $H_{0,i}$ is rejected at a prescribed level α_H if $\hat{p}_i = \mathbb{P}\{S_m \geq y_i \mid a\} \leq \alpha_H$, with $\hat{p}_i$ computed according to (3). This significance criterion conditionally supports the one-sided alternative

$$H_{a,i}: \ \xi_i, \ldots, \xi_{i+m} \quad \begin{array}{l}\text{are i.i.d. Bernoulli } (\varepsilon_i) \\ \text{random variables with } \varepsilon_i > p\end{array}$$

against the null, $H_{0,i}$.

This procedure partitions the chromosome into two regions:

$$\mathfrak{B} = \{w_i : \hat{p}_i \leq \alpha_H, i \in [1, N - m + 1]\} \quad \text{and} \quad \mathfrak{B}^c.$$

The connected components of $\mathfrak{B}$ 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 $\mathfrak{B}$, we adjusted the corresponding P-values by using the Benjamini–Yekutieli correction (Benjamini and Yakutieli, 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 ith window, so that for the control level α_H, the corrected P-values define set

$$\mathfrak{B}^* = \{w_i : p_i^* \leq \alpha_H, i \in [1, N - m + 1]\} \subseteq \mathfrak{B}.$$

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

The method just described proceeds in two steps. The first one detects possible hotspot regions while scanning with S_m for m and α_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 ℓ may be thought of as being elicited with the scan statistic S_ℓ. This segment is finally classified as a hotspot if its P-value is $\leq \alpha_H$. Hotspots were defined by taking $\alpha_H = 0.05$ and by using several initial window widths. We denote hereafter by SS_m 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, $\ell_i = x_{i+1} - x_i - 1$. The independence of the underlying Bernoulli process implies that L_i is a geometric random variable with parameter P, i.e. $\mathbb{P}\{L_i = \ell_i\} = (1 - p)^{\ell_i - 1} p$, $\ell_i \geq 1$. Small values for

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

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 $\mathbb{P}\{L_i \leq \ell_i\} \leq \alpha_c$, where α_c is a given significance level specified in advance. Suppose that a given sequence of adjacent segments of widths $\ell_i, \ell_{i+1}, \ldots, \ell_{i+r}$ is identified as a hotspot. Let

$$L_i^r = \sum_{j=0}^{r} L_{i+j} \text{ and } \ell_i^r = \sum_{j=0}^{r} \ell_{i+j},$$

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

$$\overline{p} = \mathbb{P}\{L_i^r \leq \ell_i^r\}.$$

This probability is directly available because L_i^r is a negative binomial random variable with parameters $r + 1$ and a/N, i.e.

$$\overline{p} = \sum_{k=r+1}^{\ell_i^r} \mathbb{P}\left\{L_i^r = k\right\} = \sum_{k=r}^{\ell_i^r} \binom{k}{r} \left(\frac{a}{N}\right)^{r+1} \left(1 - \frac{a}{N}\right)^{k-r}.$$

This method was used by Klein *et al.* (2011) and Jankovic *et al.* (2013) to define a set of potential hotspots by taking $\alpha_c = 0.01$. 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) $\overline{p} \leq 1 \times 10^{-9}$.

Hereafter, we refer to this procedure as the local method and denote it by NB_{α_c}. We describe results obtained with $\mathrm{NB}_{0.01}$ and $\mathrm{NB}_{0.05}$.

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 AID-expressing retrovirus (denoted hereafter as AID^{rv}), (ii) a library from AID-deficient B cells (denoted as $\mathrm{AID}^{-/-}$), (iii) a library from

53BP1-deficient B cells infected with AID retrovirus (denoted as $53BP1^{-/-}AID^{rv}$) and (iv) a library from 53BP1-deficient AID-deficient B cells (denoted as $53BP1^{-/-}AID^{-/-}$). The list with the translocation breakpoints passed on to `hot_scan` in BED format was generated by mapping onto the reference genome as described in Klein *et al.* (2011).

The association between translocation hotspots and RNA polymerase II (*PolII*) accumulation was examined by using ChIP-seq experiments deposited at the Gene Expression Omnibus database http://www.ncbi.nlm.nih.gov/geo under the accession number GSE24178.

2.5 Simulation study

Non-hotspot segments are rather long and characterized by the occasional occurrence of sparsely distributed breakpoint events. Hotspot segments are much shorter and present a higher breakpoint rate. To simulate data with these features, we considered a simple two state hidden Markov model with Bernoulli emission probabilities. Let s_1 and s_2 be the hidden states that specify the instances in non-hotspot and hotspot regions, respectively. Conditionally on being at s_1 or s_2, the observations are consistent with H_0 as stated in Section 2.1. The dynamics between s_1 and s_2 is governed by the transition probability matrix

$$P=\begin{bmatrix} \eta & 1-\eta \\ 1-\nu & \nu \end{bmatrix}.$$

Denote by T_1 and T_2 the waiting time at s_1 and s_2, respectively. It is immediate from the form of the transition matrix that $\mathbb{P}\{T_1 > t\} = \eta^t$, for $t \geq 1$. This implies that $\mathbb{P}\{T_1 = t\} = \eta^{t-1} - \eta^t = \eta^{t-1}(1-\eta)$; hence, T_1 is a geometric random variable with parameter $1-\eta$, and its expected value is therefore $\mathbb{E}[T_1] = 1/(1-\eta)$. Arguing in the same fashion gives $\mathbb{E}[T_2] = 1/(1-\nu)$. Note that $\mathbb{E}[T_1]$ and $\mathbb{E}[T_2]$ correspond to the mean length of a non-hotspot and a hotspot segment, respectively. Let β_1 and β_2 be the probability of observing a breakpoint at any given base, conditionally on being at s_1 or s_2. The parameters η, ν, β_1 and β_2 can be set to reproduce the breakpoint characteristics observed in our data. Let τ_1 and τ_2 be the observed mean length of a non-hotspot and a hotspot segment, respectively. Taking $\tau_1 = \mathbb{E}[T_1]$ and $\tau_2 = \mathbb{E}[T_2]$ leads to

$$\eta = \frac{\tau_1 - 1}{\tau_1} \quad \text{and} \quad \nu = \frac{\tau_2 - 1}{\tau_2}.$$

Let ζ_1 and ζ_2 be the mean number of observed breakpoints in a non-hotspot and a hotspot segment. Because the number of breakpoints in a given segment is a Binomial random variable, it follows that

$$\beta_1 = \frac{\zeta_1}{\tau_1} \quad \text{and} \quad \beta_2 = \frac{\zeta_2}{\tau_2}.$$

To asses the performance of the methods described here, we simulated data resembling the AID^{rv} and the $53BP1^{-/-}AID^{rv}$ conditions. The former represents a relatively simple task and the later a more challenging one owing to a higher breakpoint frequency in non-hotspot regions. Both datasets gave similar estimates for τ_1 and τ_2, namely, $\tau_1 = 6.5 \times 10^6$ and $\tau_2 = 4154$ leading to $\eta = 0.9999998$ and $\nu = 0.9997592$. For the $53BP1^{-/-}AID^{rv}$ data, we found that $\zeta_1 = 692.061$ and hence $\beta_1 = 1.065 \times 10^{-4}$. For the AID^{rv} data, we found that $\zeta_1 = 176.03055$ and thus $\beta_1 = 2.708 \times 10^{-5}$. For both datasets, we found approximately that $\zeta_2 = 12.462$, which gives $\beta_2 = 0.003$. Harder inferential tasks may be considered with these parameters by taking β_1 closer to β_2.

2.6 Enrichment analysis

The analysis of genes targeted by AID that are discovered by the procedures in Sections 2.2 and 2.3 was made by using *WebGestalt* (Wang *et al.*, 2013). The set of genes targeted by AID was compared with the mouse genome using the hypergeometric test, followed by correction for multiple testing using the Benjamini & Hochberg method at a significance level of Top10. The high-level functional classification was based on *Gene Ontology (GO) Slim* for all three major GO term categories, namely, biological process, cellular component and molecular function.

2.7 Software implementation

The genome-wide scan for hotspots according to the method described in Section 2.2 is implemented by a program we call `hot_scan`. `hot_scan` is written in Perl and R, and depends on the Perl modules Parallel::ForkManager and Math::GSL::SF, available via CPAN search (http://search.cpan.org). The former is required for simple parallelization and the latter to evaluate the hypergeometric function in (3).

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, SS_{2000}, SS_{2500} and SS_{5000}. 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 SS_{5000} shows that the hotspot length distribution is roughly described by two components, one with a mean length of $167 = \lfloor e^{5.12} \rfloor$ base pairs and the other with mean length equal to $4154 = \lfloor e^{8.332} \rfloor$. Supplementary Table S1 presents the means,

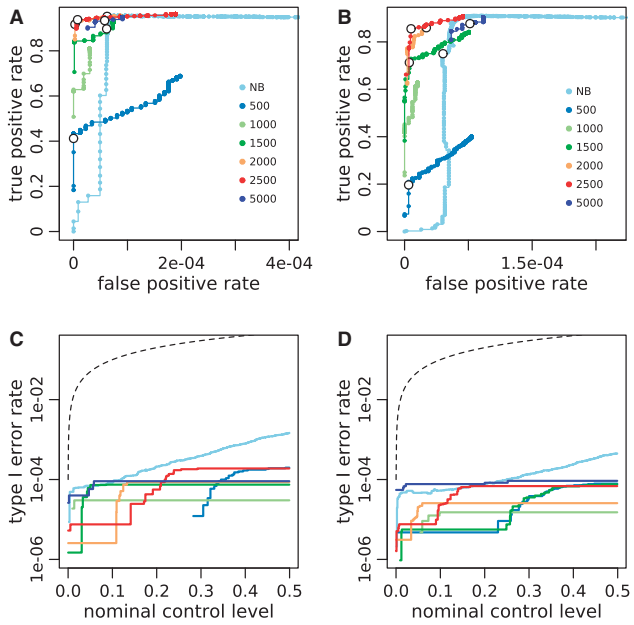

Fig. 1. **A** and **B**: breakpoint classification ROC curves for synthetic data resembling the AID^{rv} and 53BP1$^{-/-}$AID^{rv} conditions, respectively. Results for the local method are identified as NB, and those for `hot_scan` according to window width. **C** and **D**: type I error control for 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_{2500} and SS_{5000})

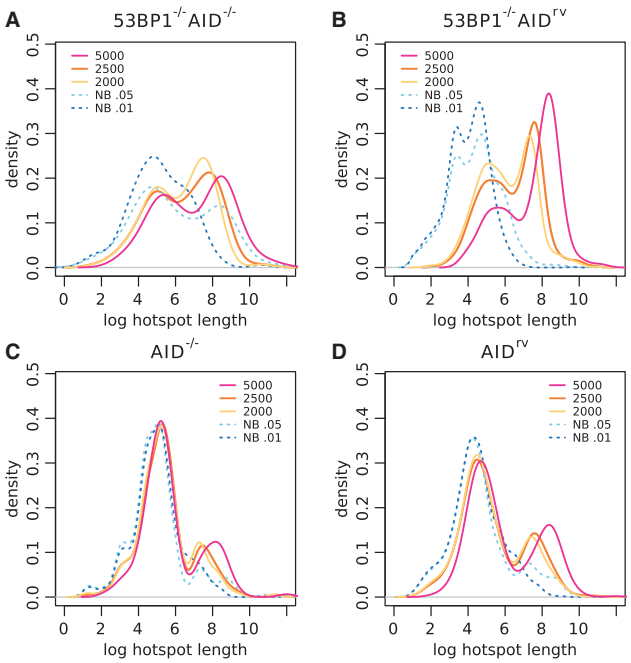

Fig. 2. Hotspot length distributions. **A** presents the distributions for the 53BP1$^{-/-}$AID$^{-/-}$ sample, **B** for 53BP1$^{-/-}$AID^{rv}, **C** for AID$^{-/-}$ and **D** for AID^{rv}. The distributions for the hotspots determined with SS_{5000}, SS_{2500} and SS_{2000} are shown as continuous 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

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^{rv} dataset (Fig. 2D) and the 53BP1$^{-/-}$AID^{rv} dataset (Fig. 2B) reveals that the relative frequency of long hotspots is higher in the absence of 53BP1. Indeed, the value for ρ in the 53BP1$^{-/-}$AID^{rv} sample is 1.95, while for the AID^{rv} 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^{rv} 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 $\ln(1-\alpha_c)/\ln(1-p)$. Using larger values for α_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_{500}, SS_{2500} and SS_{5000}. A comparison of the hotspots found by $NB_{0.01}$ (Fig. 3A) and those by the scan statistic with a relatively small window, namely, by SS_{500} (Fig. 3C), shows that the hotspots defined by either method share the same location. However, the analysis with SS_{2500} and SS_{5000} (Figs 3D and B) reveals the existence of longer hotspots, which include one or few smaller

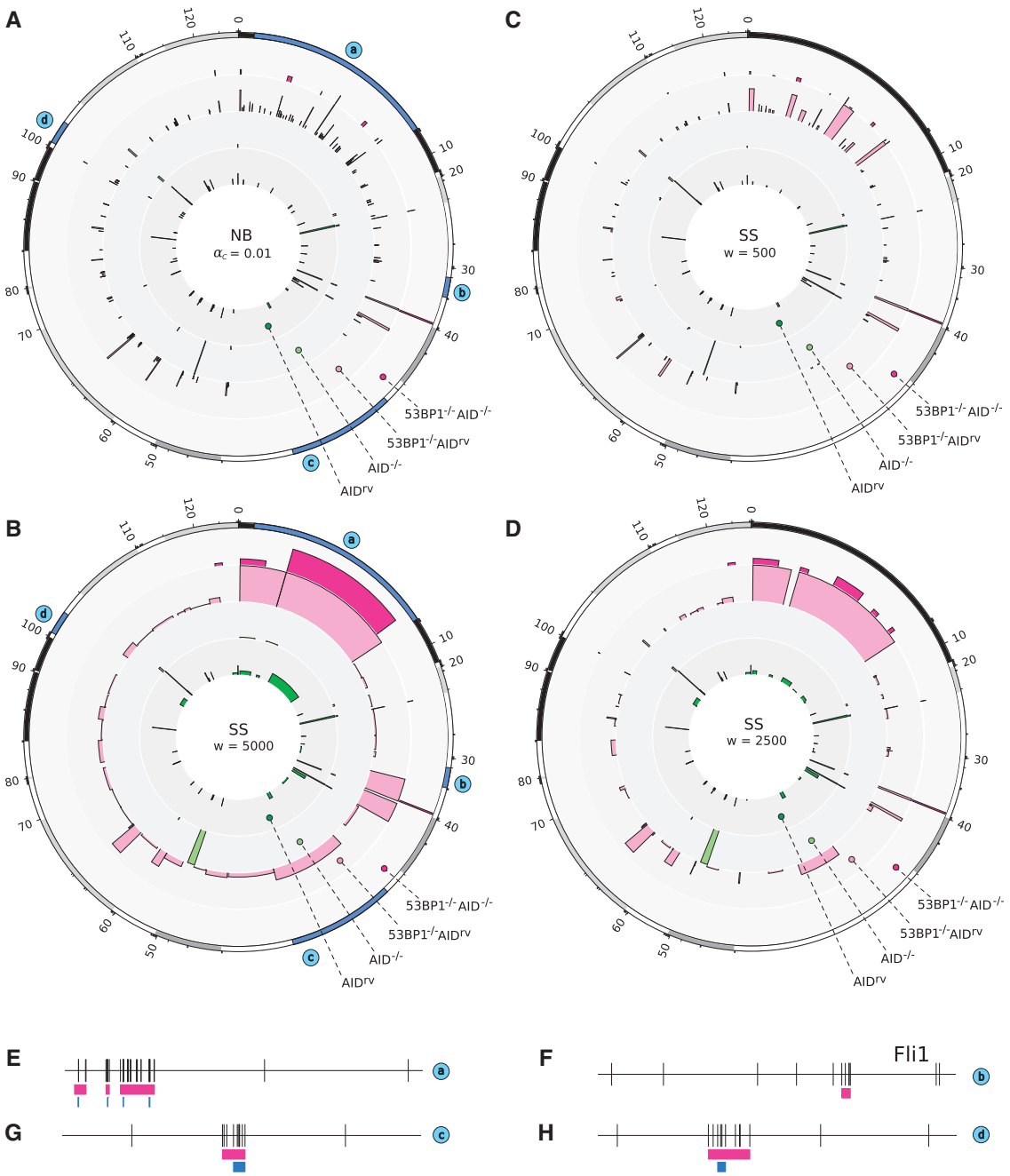

Fig. 3. Relative position for the hotspots in chromosome 9. The circular plots in **A**, **B**, **C** and **D** show, respectively, the hotspots inferred by the local method and the scan statistic with $w = 500$, 5000 and 2500. The innermost track presents the hotspots for the AID^{rv} data, the next track shows those for the AID$^{-/-}$ sample, then those for 53BP1$^{-/-}$AID^{rv} and the outermost track for 53BP1$^{-/-}$AID$^{-/-}$. Traces in **E–H** present few examples of the actual translocation breakpoints and the estimated hotspots by both techniques for the AID^{rv} data. The hotspots by the scan statistic (with $w = 5000$) are shown in pink and those by the local method in blue. The actual locations of these segments within the chromosome are identified with the labels (a)–(d) in A and B. The segment in F corresponds to the full extension of the *Fli1* gene. Circular plots were made with circos (Krzywinski *et al.*, 2009), by plotting the significance of each hotspot along the radial axis as $-\log(\overline{p})$ in A and $-\log(\hat{p})$ in B, C and D. Owing to the sparsity of the translocation hotspots, all the hotspot regions in B were expanded by using a scaling factor equal to 10 000. The same regions defined in B were also expanded in A, C and D to allow for comparisons

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 SS_{500} (Fig. 3C) for the $53BP1^{-/-}AID^{-/-}$ and the $53BP1^{-/-}AID^{rv}$ data. The analysis of the same data with a larger scanning window, namely, with SS_{2500} and SS_{5000}, 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 $53BP1^{-/-}AID^{rv}$ and the AID^{rv} 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^{rv} and the $AID^{-/-}$ data. We found 36 exclusive hotspots with `hot_scan` (Supplementary Table S2) and 27 exclusive hotspots with $NB_{0.05}$ and $NB_{<0.01}$ (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 double-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 1 496 058 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 *Il4i1*. 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 *Igκ-AID* $Ung^{-/-}$: 79.7 $x10^{-5}$) 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, *Fli1*, *Dlx5* and *Birc3*. The *Fli1* (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 chomosome-wide basis by using scan statistics. We show via simulations that scan statistics perform equally well as the local method in datasets where the hotspots are relatively obvious. Scan statistics are superior than 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 γ_3, γ_1 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

Balakrishnan,N. and Koutras,M.V. (2001) *Runs and Scans with Applications*. Volume 415 of *Wiley Series in Statistics*. Wiley, New York.

Barlow,J.H. *et al.* (2013) Identification of early replicating fragile sites that contribute to genome instability. *Cell*, **152**, 620–632.

Benjamini,Y. and Yakutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat*., **29**, 1165–1188.

Berger,M. *et al.* (2013) Genomic *ews-fli1* fusion sequences in Ewing sarcoma resemble breakpoint characteristics of immature lymphoid malignancies. *PLoS One*, **8**, e56408.

Bothmer,A. *et al.* (2011) Regulation of DNA end joining, resection, and immunoglobulin class switch recombination by 53BP1. *Mol. Cell*, **42**, 319–329.

Bunting,S.F. *et al.* (2010) 53BP1 inhibits homologous recombination in Brca1-deficient cells by blocking resection of DNA breaks. *Cell*, **141**, 243–254.

Busch,K. *et al.* (2007) Identification of two distinct MYC breakpoint clusters and their association with IGH breakpoint regions in the t(8; 14) translocations in sporadic Burkitt-lymphoma. *Leukemia*, **21**, 1739–1751.

Chaudhuri,J. *et al.* (2004) Replication protein A interacts with AID to promote deamination of somatic hypermutation targets. *Nature*, **430**, 992–998.

Chiarle,R. *et al.* (2011) Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells. *Cell*, **147**, 107–119.

de Wit,E. and de Laat,W. (2012) A decade of 3C technologies: insights into nuclear organization. *Genes Dev*., **26**, 11–24.

Di Noia,J.M. and Neuberger,M.S. (2007) Molecular mechanisms of antibody somatic hypermutation. *Annu. Rev. Biochem*., **76**, 1–22.

Dierlamm,J. *et al.* (1999) The apoptosis inhibitor gene API2 and a novel 18q gene, MLT, are recurrently rearranged in the t(11;18)(q21;q21) associated with mucosa-associated lymphoid tissue lymphomas. *Blood*, **93**, 3601–3609.

Difilippantonio,S. *et al.* (2008) 53BP1 facilitates long-range DNA end-joining during V(D)J recombination. *Nature*, **456**, 529–533.

Glaz,J. (1989) Approximations and bounds for the distribution of the scan statistic. *J. Am. Stat. Assoc*., **84**, 560–566.

Glaz,J. *et al.* (2001) *Scan Statistics. Springer series in Statistics*. Springer Verlag, New York.

Gostissa,M. *et al.* (2011) Mechanisms that promote and suppress chromosomal translocations in lymphocytes. *Annu. Rev. Immunol*., **29**, 319–350.

Hakim,O. *et al.* (2012) DNA damage defines sites of recurrent chromosomal translocations in B lymphocytes. *Nature*, **484**, 69–74.

Hasan,S.K. *et al.* (2010) Analysis of t(15;17) chromosomal breakpoint sequences in therapy-related versus de novo acute promyelocytic leukemia: association of DNA breaks with specific DNA motifs at PML and RARA loci. *Genes Chromosomes Cancer*, **49**, 726–732.

Jankovic,M. *et al.* (2013) 53BP1 alters the landscape of DNA rearrangements and suppresses AID-induced B cell lymphoma. *Mol. Cell*, **49**, 623–631.

Klein,I.A. *et al.* (2011) Translocation-capture sequencing reveals the extent and nature of chromosomal rearrangements in B lymphocytes. *Cell*, **147**, 95–106.

Krzywinski,M.I. *et al.* (2009) Circos: an information aesthetic for comparative genomics. *Genome Res*., **19**, 1639–1645.

Kumar-Sinha,C. *et al.* (2008) Recurrent gene fusions in prostate cancer. *Nat. Rev. Cancer*, **8**, 497–511.

Kuppers,R. (2005) Mechanisms of B-cell lymphoma pathogenesis. *Nat. Rev. Cancer*, **5**, 251–262.

Kuppers,R. and Dalla-Favera,R. (2001) Mechanisms of chromosomal translocations in B cell lymphomas. *Oncogene*, **20**, 5580–5594.

Liu,M. *et al.* (2008) Two levels of protection for the B cell genome during somatic hypermutation. *Nature*, **451**, 841–845.

Loader,C.R. (1991) Large deviation approximations to the distribution of scan statistics. *Adv. Appl. Prob*., **23**, 751–771.

Ma,W. and Wong,W.H. (2011) The analysis of ChIP-Seq data. *Methods Enzymol*., **497**, 51–73.

Naus,J.I. (1974) Probabilities for a generalized birthday problem. *J. Am. Stat. Assoc*., **69**, 810–815.

Naus,J.I. and Wallenstein,S. (2004) Multiple window and cluster size scan procedures. *Methodol. Comput. Appl. Probab*., **6**, 389–400.

Nussenzweig,A. and Nussenzweig,M.C. (2010) Origin of chromosomal translocations in lymphoid cancer. *Cell*, **141**, 27–38.

Pavri,R. *et al.* (2010) Activation-induced cytidine deaminase targets DNA at sites of RNA polymerase II stalling by interaction with Spt5. *Cell*, **143**, 122–133.

Perone Pacifico,M. *et al.* (2007) Scan clustering: a false discovery approach. *J. Multivariate Anal*., **98**, 1141–1469.

Rabbitts,T.H. (2009) Commonality but diversity in cancer gene fusions. *Cell*, **137**, 391–395.

Ramiro,A.R. *et al.* (2006) Role of genomic instability and p53 in AID-induced c-myc-Igh translocations. *Nature*, **440**, 105–109.

Reiter,A. *et al.* (2003) Genomic anatomy of the specific reciprocal translocation t(15;17) in acute promyelocytic leukemia. *Genes Chromosomes Cancer*, **36**, 175–188.

Revy,P. *et al.* (2000) Activation-induced cytidine deaminase (AID) deficiency causes the autosomal recessive form of the Hyper-IgM syndrome (HIGM2). *Cell*, **102**, 565–575.

Riggi,N. and Stamenkovic,I. (2007) The biology of Ewing sarcoma. *Cancer Lett*., **254**, 1–10.

Segal,M.R. and Wiemels,J.L. (2002) Clustering of translocation breakpoints. *J. Am. Stat. Assoc*., **97**, 66–76.

Simonis,M. *et al.* (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet*., **38**, 1348–1354.

Stavnezer,J. *et al.* (2008) Mechanism and regulation of class switch recombination. *Annu. Rev. Immunol*., **26**, 261–292.

Storb,U. *et al.* (2007) Targeting of AID to immunoglobulin genes. *Adv. Exp. Med. Biol*., **596**, 83–91.

Tan,Y. *et al.* (2008) A novel recurrent chromosomal inversion implicates the homeobox gene Dlx5 in T-cell lymphomas from Lck-Akt2 transgenic mice. *Cancer Res*., **68**, 1296–1302.

Wallenstein,S. and Neff,N. (1987) An approximation for the distribution of the scan statistic. *Stat. Med*., **6**, 197–207.

Wang,J. *et al.* (2013) WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Res*., **41**, 77–83.

Wielmels,J.L. *et al.* (2002) Site-specific translocation and evidence of postnatal origin of the t(1;19) e2a-pbx1 fusion in childhood acute lymphoblastic leukemia. *PNAS*, **99**, 15101–15106.

Yamane,A. *et al.* (2011) Deep-sequencing identification of the genomic targets of the cytidine deaminase AID and its cofactor RPA in B lymphocytes. *Nat. Immunol*., **12**, 62–69.

Zhang,Y. *et al.* (2010) The role of mechanistic factors in promoting chromosomal translocations found in lymphoid and other cancers. *Adv. Immunol*., **106**, 93–133.