BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 11 2014, pages 1493–1500
doi:10.1093/bioinformatics/btu035

Genome analysis

Advance Access publication January 30, 2014

Comparing DNA integration site clusters with scan statistics

Charles C. Berry[1,*], Karen E. Ocwieja[2], Nirav Malani[2] and Frederic D. Bushman[2]

[1]Division of Biostatistics and BioInformatics, Department of Family and Preventive Medicine, University of California at San Diego, La Jolla, CA 92093-0901 and [2]Department of Microbiology, Perelman School of Medicine at the University of Pennsylvania, 425 Johnson Pavilion, Philadelphia, PA 19104-6076, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Gene therapy with retroviral vectors can induce adverse effects when those vectors integrate in sensitive genomic regions. Retroviral vectors are preferred that target sensitive regions less frequently, motivating the search for localized clusters of integration sites and comparison of the clusters formed by integration of different vectors. Scan statistics allow the discovery of spatial differences in clustering and calculation of false discovery rates providing statistical methods for comparing retroviral vectors.

Results: A scan statistic for comparing two vectors using multiple window widths is proposed with software to detect clustering differentials and compute false discovery rates. Application to several sets of experimentally determined HIV integration sites demonstrates the software. Simulated datasets of various sizes and signal strengths are used to determine the power to discover clusters and evaluate a convenient lower bound. This provides a toolkit for planning evaluations of new gene therapy vectors.

Availability and implementation: The `geneRxCluster` R package containing a simple tutorial and usage hints is available from http://www.bioconductor.org.

Contact: ccberry@ucsd.edu

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

Received on September 16, 2013; revised on January 2, 2014; accepted on January 15, 2014

1 INTRODUCTION

Retroviruses such as HIV prefer to target specific regions and locations of the host genome for integration (Schroder *et al.*, 2002; Wu *et al.*, 2003), and these preferences differ between retroviruses (Mitchell *et al.*, 2004). Extensive studies have shown that the integration targeting preferences characteristic of different retroviruses are preserved in retroviral vectors, which are engineered derivatives used for gene transfer during human gene therapy.

Gene therapy with retroviral vectors benefits patients but carries risks associated with integration in sensitive genomic regions (Deichmann *et al.*, 2007; Hacein-Bey-Abina *et al.*, 2003, 2010), such as the promoter of the LMO2 proto-oncogene. This motivates the development of vectors with reduced preference for sensitive genomic regions. To track such outcome in human gene therapy, the US Food and Drug Administration has advised the monitoring of integration site (IS) distribution in cells from gene-corrected subjects to assess the risk of integration into sensitive regions (U.S. Food and Drug Administration, CBER, 2006). Comparing the risks of candidate gene therapy vectors depends in part on assessing their relative preferences for integration in specific genomic locations of concern.

Vectors can be compared with respect to their preference for targeting known sensitive regions. However, limited understanding of the relationship between the genomic location of an integration and its risk constrains the usefulness of inspecting pre-specified regions and of supervised learning or regression methods for risk assessment, focusing interest instead on clustering detected in empirical studies.

Local clusters of ISs favoring one or another vector reflect preferences for local targets. Scan statistics (generated by moving a window over a defined space and computing a summary at each position) have many applications in genomics (for an early review, see Karlin *et al.*, 1991). It has been shown that combining scans using windows with a range of widths improves power over a single fixed-width scan (Zhang, 2008). The false discovery rate (FDR) of such scan statistics is estimated by

$$\widehat{\mathrm{FDR}} = \frac{\lambda}{1+R} \tag{1}$$

where λ is the expected number of false discoveries and R is the number of discovered *clumps*—i.e. intervals formed by combining adjacent or overlapping windows with scan statistics passing a defined threshold. The estimate is unbiased when the number of false discoveries is Poisson with mean λ and independent of the number of true discoveries. The Poisson assumption is reasonable for clumps formed from locally dependent windows (Aldous, 1988). See Siegmund *et al.* (2011) and the references therein for proofs, details and further discussion.

A method for discovering chromosomal intervals (or clumps) favoring one vector over another using these methods is outlined in the next section. Then it is applied to several sets of HIV ISs to illustrate how the method could be used. In the examples, we examine differential clustering in a relatively subtle setting, involving comparison of methods for recovering ISs made by infection with HIV or an HIV-based vector in lymphoid cells. We then discuss the use of this method in the comparison of gene therapy vectors and some of the analytic issues that arise. In Supplementary S2 (Section 6) it is shown that the method generalizes the definition of so-called common integration sites (Abel *et al.*, 2007) (sets of n ISs from one vector in a defined interval) and shows how FDRs can be estimated for them.

*To whom correspondence should be addressed.

We also present a limited comparison with a machine learning technique that recursively splits data from a chromosome to form genomic intervals (Olshen *et al*., 2004).

2 METHODS

2.1 FDRs for vector clumps

Figure 1 provides an example of the ISs, their locations on the chromosome and the windows that overlie them for one clump. Referring to it first may ease reading the details here. Clumps are formally defined in Supplementary S2 (Section 3). The datum representing one IS in a comparative study has three components: (i) the chromosome on which it is located (e.g. 'chr1'), (ii) the *position* on the chromosome as the count of bases from the short end or *pter* (e.g. 147 251 235) and (iii) an indicator of which vector was used (e.g. 0 for vector A, 1 for vector B). Typically, thousands of ISs are collected using each vector (lest the experiment be uninformative), and the basic data form a table with one row for every

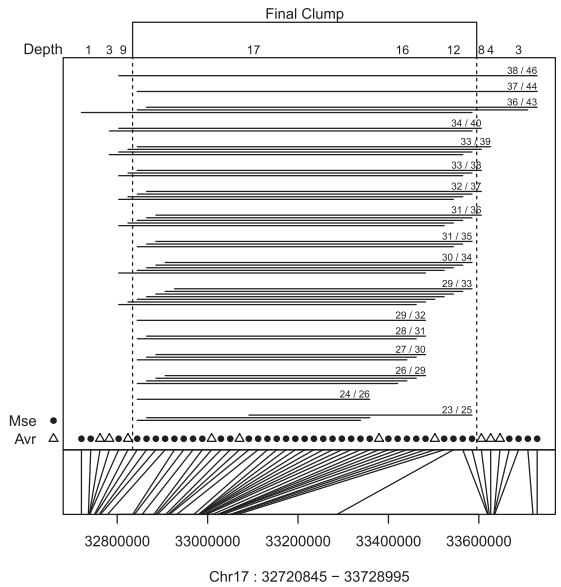

Fig. 1. Clustering of ISs by recovery method. A region of chromosome 17 is shown and marked by locations of HIV vector ISs recovered by each of two methods (see Section 3.3 for details). Sites recovered after cleavage of genomic DNA with Mse are marked with circles; those recovered after cleavage with the Avr cocktail are shown by the triangles. The circles and triangles are shown evenly spaced along the x-axis (lower middle) for ease of visualization, then connected by lines to the chromosome scaffold (bottom) to show their distribution on the chromosome. A clump, or genomic interval enriched for a recovery method, is discovered as follows: Windows containing between 15 and 75 sites were tested for enrichment of Avr or Mse, but only those passing the enrichment test are shown. Each horizontal bar (upper section of figure) represents a window that spans the x-values for a group of sites that show enrichment for Mse. (In this region, no such group shows enrichment for Avr.) The bars are grouped together according to the number of sites spanned and the number of Mse sites required to declare enrichment (see text for details of how the cutpoints were chosen). Those numbers are given just above the right edge of each group as a fraction (required/spanned). The depth, i.e. number of different window size groups over a site, is indicated at the top. Sites with greatest depth are always included in the final clump, but sites at the edges are screened using a likelihood criterion; in this case several sites on each side are excluded. Dashed vertical lines enclose the sites that were ultimately assigned to the clump

IS and three columns for the components just listed. Genomic intervals that host relatively more IS from either vector (chr1:154 268 322–154 272 351, say) are of interest, and such intervals would be subjected to further study using informatic tools, wet bench experiments or monitoring in clinical trials. It is natural to describe the intervals in terms of that sort of molecular address (as long used, e.g. by Abel *et al*., 2007; Karlin *et al*., 1991). However, statistical testing in that coordinate system can be complicated by nuisance parameters for spatial inhomogeneities shared by the two vectors. Treating the spatial order of the integrations as the coordinate reduces the problem to a Bernoulli process (or its generalization when ties are present) that is homogeneous under the null hypothesis, and a scan using a binomial test can identify regions in which one vector is overrepresented. For example, if there are n IS on one chromosome that are ordered by position and the vector indicators collected into windows for the positions indexed by intervals $\{1, \ldots, w\}$, $\{2, \ldots, w+1\}$, $\ldots$, $\{n-w+1, \ldots, n\}$, the sum of the indicators in each interval would be compared with cutpoints for the binomial distribution with w trials and intervals with suitably high or low counts marked. The process is repeated for the other chromosomes, and then all is repeated using another choice of w. Marked intervals that overlap or adjoin are connected to form a single interval or *clump*. The number of clumps discovered is R. The value of λ is the average number of clumps discovered under the null hypothesis. Here it is estimated by the average number of clumps discovered over replications in which the indicators in the third column of the table described above are permuted and the procedure just outlined is applied. The FDR is then estimated by substituting R and that estimate of λ into (1).

Many details have been ignored here. These include what values of w to use, how to choose cutpoints for low and high counts and how to handle contradictions when a window is overlapped by other windows favoring each of the vectors. These are discussed in the following sections and in Supplementary S2 (Sections 3, 4 and 5). Also, the ends of a clump may contain ISs that do not favor either vector, in which case removing them from the clump is sensible. Finally, the introduction of a non-specific filter based on the number of bases covered by a window can improve power (Bourgon *et al*., 2010) and allows control over the scan using distance in bases as the coordinate. The filter is described at the end of Section 2.2.

2.2 Algorithm for finding clumps

Two different sets of ISs generated by integration of two retroviral vectors form a N by three table describing locations (chromosome and position) and vector indicators. There are N sites, indexed by $i = 1, \ldots, N$, whose locations $L = \{(l_{i1}, l_{i2}) : i = 1, \ldots, N\}$ are ordered by chromosome (l_{i1}) and by position (l_{i2}) on each chromosome. The vector that contributes site i is indicated by $m_i = \{0, 1\}$, and local regions will be compared with the genome-wide *background* odds of $\sum_{i=1}^{N} m_i : \sum_{i=1}^{N} (1 - m_i)$, which are determined by experimental design or happenstance but are not of direct interest. Those background odds are used to establish cutpoints (see Section 2.3) for the counts described in the next paragraph. A collection of sliding windows of J different widths, $\{w_j, j = 1, \ldots, J\}$ covers the locations. Typically, every one of a range of widths is used, i.e. $w_j = w_{j-1} + 1$.

Some notation is needed to refer to those windows, the sites they cover, the counts in the window and all the windows covering a given site. For each width, the set of overlapping windows has the element $\mathbf{S}_{ij}$ that is the set of consecutive integers $\{i - w_j + 1, \ldots, i\}$, so $\mathbf{S}_{w_1,1} = \{1, \ldots, w_1\}$ is the first window on the first chromosome for the narrowest width. For notational convenience, sets $\mathbf{S}_{11}$ through $\mathbf{S}_{w_1-1,1}$ are defined as the empty set and so is every other $\mathbf{S}_{ij}$ whose index i appears in the first $w_j - 1$ positions of a chromosome. This convention allows the index i run from 1 to I and for summation over elements of any $\mathbf{S}_{ij}$ (whose sum is 0 for the empty set). As a notational convenience, let $\mathbf{B}_j(L)$ refer to the set of all values of i for which $\mathbf{S}_{ij}$ is a set of w_j consecutive integers. Also, some indexes may be

removed from $\mathbf{B}_j(L)$ and the corresponding $\mathbf{S}_{ij}$ converted to the empty set to filter out collections of sites that sparsely cover a genomic region and to ensure that sites sharing a common location are not split between different windows of the same width. The set of sites covered by a window of width $w_{j,}$ also covering site i is $\mathbf{T}_{ij} = \bigcup_{i'=i}^{i+w_j-1} \mathbf{S}_{i'j}$. Also, $\mathbf{T}_{ij}$ indexes width w_j windows that overlap $\mathbf{S}_{ij}$. The set of windows that cover site i is $\mathbf{T}^*_{ij} = \{i' : i \in \mathbf{S}_{i'j}\}$.

An initial screening marks each window if the count of one vector is in a critical region defined by cutpoints (discussed below). If $i \in \mathbf{B}_j(L)$, the count $m^+_{ij} = \sum_{i' \in \mathbf{S}_{ij}} m_{i'}$ is compared with cutpoints to yield $n_{ij} = \delta(m^+_{ij} \geq \gamma_{j2}) - \delta(m^+_{ij} < \gamma_{j1})$ taking $\delta(\cdot)$ as the indicator function and $\gamma_{j1} \leq \gamma_{j2}$ as lower and upper critical region cutpoints. Otherwise, $n_{ij} = 0$, indicating that the count was neither so high nor so low that the window is seen to depart from the background odds or that $\mathbf{S}_{ij}$ is the empty set. A window is considered marked if it has a non-zero value of n_{ij}.

All marked overlapping or adjacent windows are gathered to form a clump unless there are marks that conflict according to the vector they favor. Such conflicts are resolved site-by-site in favor of the smallest value of j for which there is a mark, thereby making the region of conflict as small as possible.

Taking $c_{i0} = 1, i = 1, \ldots, N$, then

$$c_{ij} = \prod_{j'=1}^{j} \prod_{i' \in \mathbf{T}_{ij'}} \delta(n_{ij} n_{i'j'} \geq 0) c_{i,j'-1}$$

that is, c_{ij} indicates if the windows of width w_j overlapping site i are free of such conflicts. $\tilde{n}_{ij} = n_{ij} c_{ij}$ is corrected for overlap conflicts.

The classification of sites as being marked by windows of width j is given by

$$v_{ij} = \delta\left(0 < \sum_{i' \in \mathbf{T}^*_{ij}} \tilde{n}_{i'j}\right) - \delta\left(0 > \sum_{i' \in \mathbf{T}^*_{ij}} \tilde{n}_{i'j}\right)$$

which yields values of −1, 0 or 1 according to whether vector 1, neither vector, or vector 2 is favored, and the overall classification of each site is

$$u_i = \delta\left(0 < \sum_{j=1}^{J} v_{ij}\right) - \delta\left(0 > \sum_{j=1}^{J} v_{ij}\right)$$

and the covering *depth* is defined as $d_i = \sum_{j=1}^{J} |v_{ij}|$.

The clumps are non-zero runs in u_i, i.e. clumps are identified as $\mathbf{A}_k = \{a_{k1}, \ldots, a_{k2}\}$, where $u_{a_{k1}} u_r = 1,\ a_{k1} \leq r \leq a_{k2},\ (a_{k1}, a_{k2}) = \arg\max_{a_{k1}, a_{k2}} (a_{k2} - a_{k1})$ and $l_{a_{k1},1} = l_{a_{k2},1}$. It is sensible to prune these clumps, as the edges sometimes include sites whose vector proportions do not differ from the background. Each edge of a run is pruned back until highest depth is reached and then added back depth-by-depth using a likelihood criterion to determine the boundary. To do this, the region of greatest depth is identified as $\tilde{a}_{k1} = \min(r)$ and $\tilde{a}_{k2} = \max(r)$, where $a_{k1} \leq (r, s) \leq a_{k2}$ and $d_r = \sup_s(d_s)$. The tail regions are added according to a log likelihood criterion, treating vector identities in the provisional clump as independent Bernoulli events whose probability is the observed relative frequency and the vector identities outside the clump as Bernoulli events whose probability is the background frequency: $\hat{a}_{k1} = \arg\min_{r \in \mathbf{R}_k} (h(a_{k1}, r) + g(r, \tilde{a}_{k2}))$ and $\hat{a}_{k2} = \arg\max_{s \in \mathbf{S}_k} (h(s, a_{k2}) + g(\tilde{a}_{k1}, s))$, where $\mathbf{R}_k = \{r : a_{k1} \leq r \leq \tilde{a}_{k1}, u_r d_r \neq u_{r-1} d_{r-1}\}$ and $\mathbf{S}_k = \{s : \tilde{a}_{k2} \leq s \leq a_{k2}, u_s d_s \neq u_{s+1} d_{s+1}\}$ and $h(i,j) = \log p\left(\sum_{k=i}^{j-1} m_k; j - i, \pi_0\right)$ if $j > i$ and 0 otherwise, $g(i,j) = \log p\left(\sum_{k=i}^{j} m_k; j - i + 1, \sum_{k=i}^{j} m_k/(j - i + 1)\right)$, $p(k; n, \pi) = \pi^k (1 - \pi)^{n-k}$ is the Bernoulli mass function and the background value of its parameter often is taken as $\pi_0 = \sum_{i=1}^{N} m_i / N$.

2.3 Choosing cutpoints and setting filters

There are a number of seemingly natural ways to choose the cutpoints for discovery that would reflect departure of vector odds in a window from genome-wide background odds. A small fixed α level could be used for all window widths and cutpoints given by the binomial distribution with the proportion given by the overall frequencies of the two vectors. Alternatively, a target for expected false discoveries might be set and the largest α level satisfying it for each window size determined. There are many other possibilities, and some are discussed in Supplementary S2 (Section 5). Likewise, the fraction of windows to be pre-filtered must depend to some degree on the experimental content. Section 3.2 illustrates these considerations using two datasets.

2.4 Software implementation

The `geneRxCluster` R package implements the algorithm for finding clumps and estimating FDRs. The principal function returns a `GRanges` object (Lawrence *et al.*, 2013) representing the genomic locations of the clumps discovered with metadata annotations indicating the number of sites from each vector and the smallest target for false discoveries—taken as the smallest α level times number of windows ($|B_j|$) attained by any window in each clump. This object can interface to the BioConductor software suite (Gentleman *et al.*, 2004) containing tools for genomic data analysis, browsing and data display.

The evaluation of filtering rules, counting vectors in windows and application of cutpoints to obtain the window marks, n_{ij}, is straightforward. The computation of conflict indicators, c_{ij}, is implemented as a finite-state automaton that traverses the sets of windows covering each site, $\mathbf{T}_{ij}$, with i moving fastest and j in order $1, \ldots, J$. By updating and downdating counts of $n_{ij} > 0$ and $n_{ij} < 0$ (i.e. windows satisfying the critical region for each vector), counts of favored vector for each $\mathbf{S}_{i'j}$ window in $\mathbf{T}_{ij}$ and counts of conflicts based on lesser values of j, only one reference to each n_{ij} is needed. The order of the computation of all the c_{ij}, v_{ij} (window class) and u_i (site class) is NJ. Finally, the order of computation in pruning the clumps depends on the number of points at which pruning cuts can be made.

2.5 Illustrative datasets

The data used to illustrate the method come from two experiments. In one (here called 'Jurkat', see Wang *et al.* (2007) for more details), the ISs were generated by HIV infection of 50 independent cultures of Jurkat cells using an HIV-based vector, the DNA of each divided into two aliquots, one cleaved by the restriction enzyme MseI (here called Mse, recognition site: TTAA) and one by a cocktail of three enzymes (here called Avr, recognition sites: ACTAGT, CCTAGG and GCTAGC). DNA linkers were then ligated onto the free DNA ends, and DNAs were PCR-amplified using primers complementary to the linker and the vector DNA long terminal repeat (LTR). DNA libraries were subsequently pooled, sequenced and mapped to the hg18 freeze of the human genome (Lander *et al.*, 2001; Meyer *et al.*, 2013). It is known that the recovery of ISs after cleavage with a restriction enzyme depends on their juxtaposition with restriction sites (Alonso *et al.*, 2003; Gabriel *et al.*, 2009; Wang *et al.*, 2008), so the use of two different sets of enzymes fueled our expectation that the sites recovered might differ. The other dataset (named 'CD4+') is newly described here and comprises one of the largest datasets determined for an HIV primary isolate (designated HIV89.6) infecting primary human T-cells. An analysis of its association with genomic features is presented in Supplementary S1 where it is described in detail. Briefly, the dataset was generated using three replicate infections (referred to as Infection I, Infection II and Infection III) of CD4+ T cells infected with HIV89.6. The DNA from each infection was cleaved with NLAIII (recognition site: CATG), sequenced and mapped. It was expected that these replicates would not differ from each other but would differ from those of the Jurkat experiment.

These datasets usefully illustrate the kinds of variations that might be anticipated in an actual experiment comparing two gene therapy vectors. It is important to know if replicates show extra-Bernoulli variation

because this would invalidate the permutation estimate of λ. The Avr versus Mse comparison is expected to illustrate subtle differences that might mirror a challenging comparison of vectors. The datasets are large enough (with more than 185 000 distinct ISs) to characterize the genome-wide variation in integration targeting accurately.

3 RESULTS

3.1 Spatial association versus data source

The 147 294 CD4+ and 40 974 Jurkat ISs were ordered by the genomic locations of the sites of integration (Craigie and Bushman, 2012) of the viral DNA. Table 1 gives the identity of the members of each successive pair of integrations under that ordering. Under the null hypothesis of equal target preferences and equal recovery of integrations regardless of restriction enzyme, all rows would be equal. The three 'Infection' rows are equal or nearly so (and $\chi^2 = 6.06$, $4\,df$, $P = 0.1948$). Jurkat (Avr and Mse rows and columns combined) differs from CD4+ (all Infections combined) ($\chi^2 = 4894.12$, $1\,df$, $P < 0.0001$). Avr and Mse differ from one another ($\chi^2 = 183.58$, $1\,df$, $P < 0.0001$) to a lesser degree; that they differ is unsurprising given the bias in the recovery of an IS that depends on its distance from the relevant restriction site. The apparent homogeneity of the three infections is expected, given the use of identical materials and procedures in each replicate. The use of permutation methods to estimate the FDR would be invalidated by inhomogeneous replicates. So, this finding provides support for using permutation methods to compare ISs from experiments in which all samples of a kind are prepared from the same starting materials and processed identically. The difference between the CD4+ data and the Jurkat data might be expected considering that the sources of host cells, the HIV vector or virus and the restriction enzymes used all differ.

3.2 Tuning the clump discovery parameters

The algorithm requires a collection of window widths and upper and lower critical regions for each width. Optionally, the analyst may filter out some windows to improve power or to avoid regions of low integration density that will usually be of low interest for assessing risk.

Table 1. Successive pairs of ISs

IS source	Infection I	Infection II	Infection III	Avr	Mse	Total
Infection I	0.31	0.27	0.24	0.08	0.10	55 835
Infection II	0.32	0.27	0.23	0.08	0.10	47 553
Infection III	0.32	0.27	0.24	0.09	0.10	42 051
Avr	0.24	0.21	0.18	0.21	0.16	19 424
Mse	0.26	0.22	0.20	0.15	0.18	21 202

Note: Rows identify the source of the site at the lesser chromosomal position and columns that of the site at the higher position. The cell values are the proportions of the total given in the last column rounded to two digits (each row adds to 1.0). Locations occupied by multiple integrations are omitted, as filtering to remove PCR duplications undercounts independent integrations at the same site in the same replicate.

What values of these parameters should be chosen? A sensitive region that is much more attractive to a candidate retroviral vector may precipitate an adverse event. Current understanding suggests that sensitive regions are of limited size (e.g. the promoter region and first intron of the LMO2 proto-oncogene) and usually do not cover many megabases. Discovering that a broad region is modestly more attractive to ISs is not likely to be useful because it may include subregions of varying sensitivity, and a modest increment in integrations only modestly increases the chance that an integration would trigger an adverse event. However, a narrow region with a high rate of integration can be inspected (e.g. using a genome browser) for hints that integrations there would heighten risk. So, finding relatively small regions with high rates of integration is most useful. Further, having a few false-positive discoveries will not seriously impede investigating the risk potential of all discoveries or monitoring patients for adverse events, such as expansion of clones hosting sites in discovered regions.

A window covering many more bases than usual for a fixed number of ISs would not represent a region of high integration even if only one of the vectors accounted for all sites (unless one of the two vectors contributed almost all of the ISs). When rates of integration vary widely across the genome (as typical of many vector–host combinations), filtering out such windows can improve power (Bourgon *et al.*, 2010) and avoid focusing on regions of low interest. Supplementary S2 and Figures 1 and 2 show that most sites are in regions with lower integration rates, and the distribution is highly skewed. So, filtering out many of the sites would seem in order. For now, windows spanning more than the median number of bases for windows with the same number of sites are filtered out. For the Jurkat data, this means that windows of 15 sites will span no more than 486 735 bases, and windows of 75 sites will span no more than 4 407 247 bases. For the combined Jurkat and CD4+ data, the corresponding limits are 43 519 and 443 665 bases.

Critical regions ought to be chosen to balance true discoveries with false discoveries. With a given dataset, filtering rule and set of critical regions, the expected number of false discovered clumps can be estimated [e.g. by using permutation or subsampling (Bickel *et al.*, 2010) methods]. However, insights useful in setting critical regions can be had by studying the behavior of windows of fixed width. Figure 2 shows the critical regions for two comparisons at selected window widths when the filtering removes windows exceeding the median number of bases for each width. The critical regions are chosen so that the expected number of false discoveries in each tail of each window width is at most five. Figure 3 shows the relationship between the expected number for each window width (called target false discoveries) and the expected false discoveries based on 200 permutations of the data. It seems that choosing the critical regions to satisfy $\alpha_j \leq \frac{r/2}{\left|\bigcup_{i=1}^{N} \mathbf{S}_{ij}\right|}$ for each tail (i.e. to have α_j smaller than a target, $r/2$, for the number of falsely discovered clumps divided by the number of windows after filtering) leads to an expected number of false discoveries on the order of r. This is not too surprising: for a single window width, one expects the number of windows falsely identified to be r if the mass in the critical regions sums to exactly $2\alpha_j$, but usually their sum will be somewhat less. The overlap of discovered windows results in the

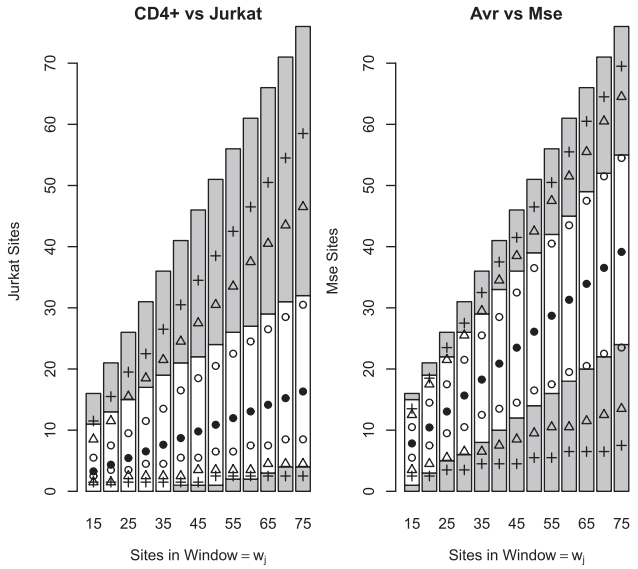

Fig. 2. Critical regions and window-wise power. Windows between 15 and 75 sites are shown (x-axis). The y-axis shows the number of successes, and expected numbers of Jurkat or Mse sites are shown (solid dots) given the background relative frequency. Critical regions (gray) are determined (see text) as are the lower and upper tails each covering 80% of the binomial mass function given window size and an alternative probability whose odds are 3 (circles), 7 (triangles), 15 (+) times (or divided into) the null background odds for the upper tail (or for the lower tail)

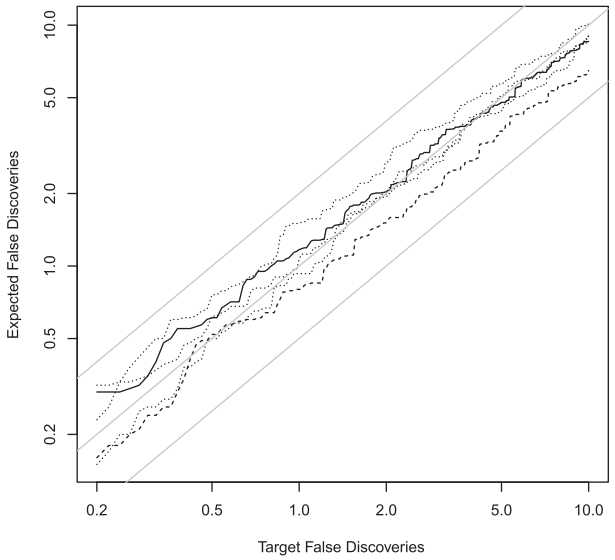

Fig. 3. Target versus expected false discoveries. Expected false discoveries are determined via permutation for target false discoveries up to 10. The plots are Avr versus Mse (solid line), CD4+ versus Jurkat (dashed line) and each replicate infection versus the others (dotted lines). The gray lines show 1/2, 1 and 2 times as many expected as target discoveries

number of clumps falsely discovered being less than the number of windows falsely discovered, so the effect of testing many overlapping windows is muted (by a factor depending on the average width of falsely discovered clumps, see Aldous, 1988).

The 'Avr versus Mse' critical regions in Figure 2 are almost symmetrical, as the background odds (i.e. the ratio of sites contributed by each vector overall) are nearly equal, whereas the asymmetrical 'CD4+ versus Jurkat' critical regions reflect the roughly four to one odds for CD4+ to Jurkat sites. Each panel shows boundaries marking the outer 80% of the tail under alternatives with higher and lower odds than the genomic average. Again, the 'Avr versus Mse' comparison is almost symmetrical—the positions of the boundary relative to the critical region for an increase versus a decrease by the same factor are quite similar. But for the 'CD4+ versus Jurkat' comparison, the boundaries cross into the critical regions at different window sizes. For a 7-fold increase in the odds of Jurkat sites, the boundary is in the critical region when the window width is 25 sites. For a 7-fold decrease, the boundary is still not in the lower critical region when the window width is 75 sites. So, it is expected that better power is obtained for clumps favoring Jurkat sites. One use of plots, like Figure 2, is to guide the choice of limits for the set of w_js; if the boundaries for suitable alternatives are well within the critical region at the highest widths, there is little point in adding wider windows. If the boundary is not in a critical region, augmenting the collection may be worthwhile. Another use is to see the effect of using a more aggressive filter to reduce the number of windows screened, which will tend to expand the critical regions.

The probability of detecting a region of changed odds must be determined by simulation. However, this simple lower bound refines the visual impression obtained from the inspection of power boundaries and is useful in laying out simulations for the ultimate determination of power:

$$1-\beta \geq \sum_{j=1}^{\infty} \exp(-\mu)\frac{\mu^{t(j)}}{t(j)!}\left(1-\sum_{k=\gamma_{s(j)1}}^{\gamma_{s(j)2}-1} p(k; w_{s(j)}, \pi_1)\right)$$

where $t(j) = w_{s(j)} + j - s(j)$, $s(j) = \min(j, J)$, and $p(\cdot)$ is the binomial mass function. $1-\beta$ is the probability of discovery using the cutpoints in γ of a genomic interval in which μ sites are expected and π_1 is the expected proportion of one of the vectors. The density of sites in the flanking regions also figures into β. The term on the right side is the probability of a result in the critical region if the sites in a pre-specified genomic interval are tested using the cutpoints for the observed number of sites in that interval (or J if there are more sites) given Poisson sampling of the sites. Figure 4 gives the probability that a genomic region will be discovered using the cutpoints of Figure 2 under several different scenarios. In each dataset, the background odds (which are an artifact of the data collection methods) are multiplied or divided by 3, 7 or 15 to determine the relative intensity of integration in the region for one vector compared with the other. Because the background odds are not 1:1, the effect in each tail is different. For both vectors, the expected sum of ISs in the region is set to 20, 40 or 60. The region is supposed to be flanked by regions either so sparse in integrations that filtering removes all windows that overlap sites in the flanking regions or so dense that none are removed. The sites in the flanking regions follow the background integration odds. As expected, more extreme odds and wider windows are associated with higher power.

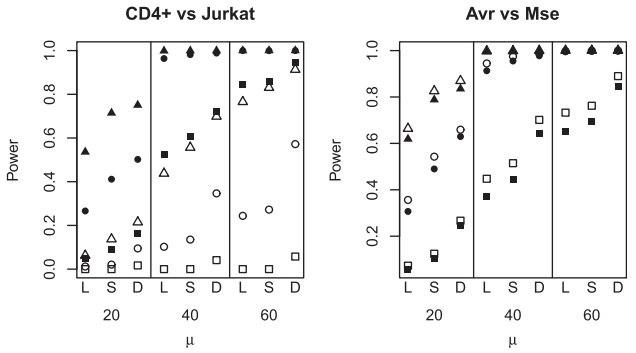

Fig. 4. Power for discovery. Alternatives depend on relative odds (squares = 3-fold, circles = 7-fold, triangles = 15-fold) and on expected number of ISs, μ and whether odds are increased (solid) or decreased (hollow) versus background. A lower bound (L) for power is computed (see text), and power is simulated when the interval is embedded in a sparsely (S) or densely (D) targeted region. Critical regions are as shown in Figure 2

The power to detect a difference in the region must be higher when there are densely populated flanking regions, and the effect is usually to increase power by ~0.10 when the power is between 0.1 and 0.9. The lower bound is closer to the simulated values when μ is higher. The bound achieved some very high values, and when it is <0.60, neither simulated value surpassed 0.80. These results do not directly depend on the number of bases in the region of interest, but it should be noted that typically regions covered by w_j sites will contain many fewer bases in datasets with more ISs. The windows for the Jurkat data alone covered 10 times as many bases at their median widths as the combined CD4+ and Jurkat sites, but the latter set had only 5 times as many sites. So, power comparisons of different datasets ought to select μ for each set according to the number of integrations in each set.

Supplementary S2 (Section 5) explores other choices and methods of finding cutpoints and criteria for filtering that may be preferred depending on study objectives. One interesting choice is to set the power required for each w_j and an odds ratio; the number of windows passing the filter is adjusted for each w_j to meet the target for false discoveries. For small values of w_j, only a handful of windows can pass the filter, showing how challenging it can be to detect local variations in integration rates in very small regions.

In summary, there are many options for establishing cutpoints and filtering criteria. The toolkit provided here allows exploration of different criteria for filtering and establishing cutpoints and the effects those criteria have on false discovery and power.

3.3 Clump example

Figure 1 shows a clump of ISs comparing Mse with Avr sites in the Jurkat dataset. Values of w_j (of 15, 16, ..., 74, 75) were used. The target for false discoveries was 0.5 for each tail, and again windows spanning more than the median number of base pairs were filtered out. Most of the sites in that clump were recovered with the Mse method compared with about half in the full dataset. Seventeen different values of w_j (between 25 and 46) marked a region containing most of the sites. The likelihood ratio criterion rejected the six sites at the left (3 Avr, 3 Mse), seven on the right (3 Avr, 4 Mse) and retained the remainder.

Table 2. Avr versus Mse FDRs

Target	Clumps discovered		FDR
	Observed	Expected	
0.2	4	0.30	0.06
1.0	7	1.24	0.13
2.0	15	2.10	0.13
10	34	8.93	0.26

Note: Cumulative number of discoveries (Observed) and false discoveries (Expected) and the FDR, according to the target for false discoveries.

3.4 FDRs

Tables of the discoveries were prepared using targets for expected false discoveries of 0.2, 1, 2 and 10. Table 2 shows that 34 clumps were discovered for Avr versus Mse at an FDR of 0.26. In a comparison of actual gene therapy vectors, this would be a small enough number of clumps to inspect them one by one using a genome browser and design wet-bench follow on studies. This FDR would probably be acceptable in that context. The CD4+ versus Jurkat comparison yielded 350 discoveries at an FDR of 0.0176, which if seen in a comparison of candidate gene therapy vectors would rule out clump-by-clump inspection but allow for comparisons via statistical analysis and data mining. For the comparison of each of the CD4+ replicate infections to the others, there were 6, 6 and 8 clumps discovered with expectations of 9.84, 8.58 and 8.27, respectively. The results for the CD4+ replicates are thus consistent with all of the clumps being false discoveries.

The clumps discovered in the Avr versus Mse comparison at target false discovery ≤10.0 are shown in Figure 5. The odds ratio for restriction sites are odds of `Mse` to `Avr` for restriction sites in the region occupied by a clump divided by its genomic average. Log odds ratios for ISs are calculated as $\log\left(0.5+\sum_{i=\hat{a}_{k1}}^{\hat{a}_{k2}} m_i\right)-\log\left(0.5+\sum_{i=\hat{a}_{k1}}^{\hat{a}_{k2}}(1-m_i)\right)-\log(\pi_0/(1-\pi_0))$ for each clump. The Spearman correlation is 0.697 ($P<0.0001$). Thus, the availability of restriction sites strongly affects the vector composition of individual clumps, as was expected.

With the clumps discovered in hand, there are various options for the biomedical scientist to develop an understanding of the mechanisms that caused them or their implications for patient care. There are 350 clumps discovered when comparing Jurkat cells with the CD4+ cells. Figure 6 shows the histogram of the log odds ratios (compared with background) for the two cell types. There are 117 clumps that favor Jurkat and 233 sites that favor CD4+. If one were faced with this many clumps in a comparison of actual gene therapy vectors, it would be possible to explore the differences between them using data mining tools. In addition, browsing the regions occupied by the clumps with extreme log odds ratios might also be productive, as those regions will be targeted more intensely by one of the vectors.

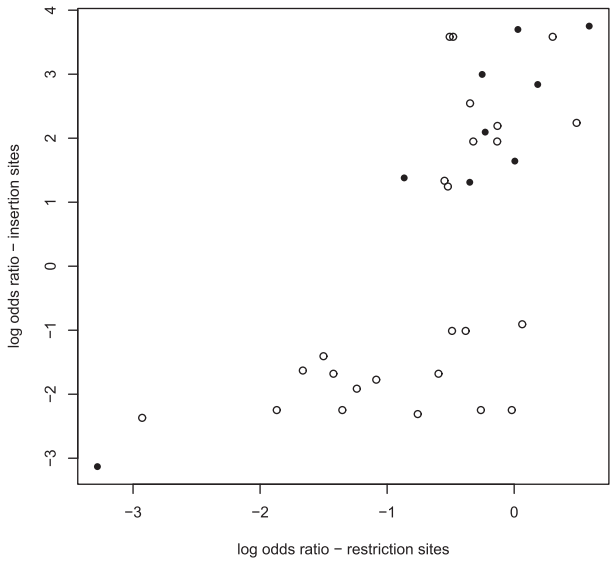

Fig. 5. Integration versus restriction sites. Log odds ratios for Mse versus Avr ISs and log odds ratios for Mse versus Avr restriction sites. See text for details. Positive values on each axis indicate that Mse sites are favored. Solid (hollow) dots indicate target for false discoveries <1.0 (≥ 1.0)

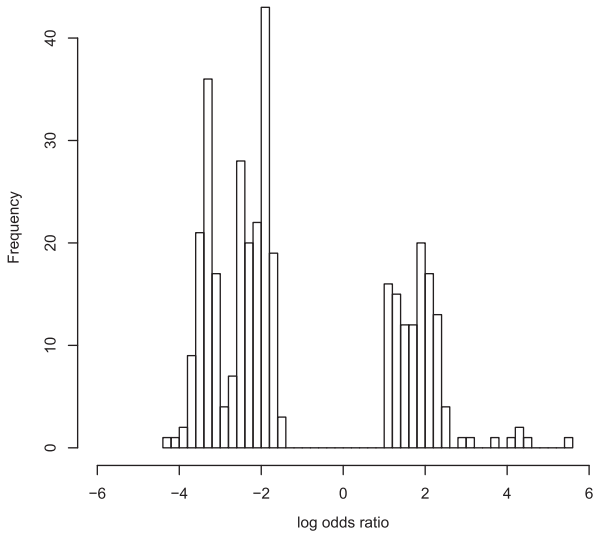

Fig. 6. Odds ratios for Jurkat versus CD4+ sites in each clump. Log odds ratios are calculated as for Figure 5 (see text). Positive values indicate that Jurkat sites are favored in the region occupied by the clump

The Jurkat clumps range from 1567 to 889 570 bases with a median of 90 254 bases, whereas the CD4+ clumps range from 4299 to 1 108 435 bases with a median of 134 158 bases. These regions are small enough to allow productive use of a genome browser to inspect them.

4 DISCUSSION

The clumping method provides a flexible toolkit for exploring local differences in collections of retroviral ISs and for planning experiments. The comparison of the Avr with Mse recovery methods may mirror future studies of gene therapy vectors—only subtle differences exist but they may have profound implications for the risk profile of a newly engineered vector.

In clinical trial reporting, the CONSORT criteria (Schulz *et al*., 2010) require reporting of the power of the trial for the pre-specified endpoint. This report steps in the direction of allowing a deliberative approach to study planning. Supplementary S2 (Section 5) illustrates some of the possibilities of using this toolkit for planning with an eye toward choosing sensible rules for filtering and sensible cutpoints for discovery.

There are various parameters that affect the discovery of clumps and the FDRs associated with them. Ideally, these would be set using prior knowledge and without dependence on statistics that correlate with the ultimate test statistic (Bourgon *et al*., 2010). Optimizing the FDR reported by searching over the parameter space has the potential to introduce resubstitution bias. When such optimization is desired, strategies should be implemented, such as training on one set of data and using an independent test set to validate the clumps discovered. Further work may explore how sensitive resubstitution bias is to different parameters.

One issue that arises in clinical trials in humans is patient-to-patient variability in the genome and other host factors that may yield patient-specific clumps of integrations. On the one hand, this challenges clumping methods that depend on assuming homogeneity among patients, and methods that resample patients or subsample genomic segments (Bickel *et al*., 2010) will be required for estimation of FDRs. Furthermore, planning for human studies will need to take heterogeneity into account.

Funding: National Institutes of Health (2R01 AI052845 and 5R01 AI082020).

Conflict of Interest: none declared.

REFERENCES

Abel,U. *et al.* (2007) Real-time definition of non-randomness in the distribution of genomic events. *PLoS One*, **2**, e570.

Aldous,D. (1988) *Probability Approximations via the Poisson Clumping Heuristic*. Springer, New York.

Alonso,J. *et al.* (2003) Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science*, **301**, 653–657.

Bickel,P.J. *et al.* (2010) Subsampling methods for genomic inference. *Ann. Appl. Stat*., **4**, 1660–1697.

Bourgon,R. *et al.* (2010) Independent filtering increases detection power for high-throughput experiments. In: *Proceedings of the National Academy of Sciences*. 9546.

Craigie,R. and Bushman,F. (2012) HIV DNA integration. *Cold Spring Harb. Perspect. Med*., **2**, a006890.

Deichmann,A. *et al.* (2007) Vector integration is nonrandom and clustered and influences the fate of lymphopoiesis in SCID-X1 gene therapy. *J. Clin. Invest*., **117**, 2225–2232.

Gabriel,R. *et al.* (2009) Comprehensive genomic access to vector integration in clinical gene therapy. *Nat. Med*., **15**, 1431–1436.

Gentleman,R.C. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*., **5**, R80.

Hacein-Bey-Abina,S. *et al.* (2003) A serious adverse event after successful gene therapy for X-linked severe combined immunodeficiency. *N. Engl. J. Med*., **348**, 255–256.

Hacein-Bey-Abina,S. *et al.* (2010) Efficacy of gene therapy for X-linked severe combined immunodeficiency. *N. Engl. J. Med.*, **363**, 355–364.

Karlin,S. *et al.* (1991) Statistical methods and insights for protein and DNA sequences. *Annu. Rev. Biophys. Biophys. Chem.*, **20**, 175–203.

Lander,E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Lawrence,M. *et al.* (2013) Software for computing and annotating genomic ranges. *PLoS Comput. Biol.*, **9**, e1003118.

Meyer,L.R. *et al.* (2013) The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res.*, **41**, D64–D69.

Mitchell,R.S. *et al.* (2004) Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol.*, **2**, e234.

Olshen,A.B. *et al.* (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.

Schroder,A.R.W. *et al.* (2002) HIV-1 integration in the human genome favors active genes and local hotspots. *Cell*, **110**, 521–529.

Schulz,K.F. *et al.* (2010) Consort 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMC Med.*, **8**, 18.

Siegmund,D. *et al.* (2011) False discovery rate for scanning statistics. *Biometrika*, **98**, 979–985.

U.S. Food and Drug Administration, CBER. (2006) *Gene Therapy Clinical Trials Observing Subjects for Delayed Adverse Events*. FDA: Maryland.

Wang,G.P. *et al.* (2007) HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Res.*, **17**, 1186–1194.

Wang,G.P. *et al.* (2008) DNA bar coding and pyrosequencing to analyze adverse events in therapeutic gene transfer. *Nucleic Acids Res.*, **36**, e49.

Wu,X. *et al.* (2003) Transcription start regions in the human genome are favored targets for MLV integration. *Science*, **300**, 1749–1751.

Zhang,Y. (2008) Poisson approximation for significance in genome-wide ChIP-chip tiling arrays. *Bioinformatics*, **24**, 2825–2831.