The Author(s) *BMC Bioinformatics* 2016, **17**(Suppl 11):320
DOI 10.1186/s12859-016-1173-8

BMC Bioinformatics

RESEARCH Open Access

A novel scan statistics approach for clustering identification and comparison in binary genomic data

Danilo Pellin[1,2*] and Clelia Di Serio[1]

From 11th International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics (CIBB 2014) Cambridge, UK. 26-28 June 2014

Abstract

Background: In biomedical research a relevant issue is to identify time intervals or portions of a n-dimensional support where a particular event of interest is more likely to occur than expected. Algorithms that require to specify a-priori number/dimension/length of clusters assumed for the data suffer from a high degree of arbitrariness whenever no precise information are available, and this may strongly affect final estimation on parameters. Within this framework, spatial scan-statistics have been proposed in the literature, representing a valid non-parametric alternative.

Results: We adapt the so called Bernoulli-model scan statistic to the genomic field and we propose a multivariate extension, named Relative Scan Statistics, for the comparison of two series of Bernoulli r.v. defined over a common support, with the final goal of highlighting unshared event rate variations. Using a probabilistic approach based on success probability estimates and comparison (likelihood based), we can exploit an hypothesis testing procedure to identify clusters and relative clusters. Both the univariate and the novel multivariate extension of the scan statistic confirm previously published findings.

Conclusion: The method described in the paper represents a challenging application of scan statistics framework to problem related to genomic data. From a biological perspective, these tools offer the possibility to clinicians and researcher to improve their knowledge on viral vectors integrations process, allowing to focus their attention to restricted over-targeted portion of the genome.

Keywords: Scan statistics, Viral integration sites, Cluster identification, Binary genomic data

Background

In many different research areas it is of interest to identify time intervals or portions of a *n*-dimensional support where a particular event is more likely to occur than expected. These regions, which in biology are commonly called *clusters* or *hotspots*, are presumable characterized by an increased probability of success and their identification may throw light on a better understanding of the underlying events-generating process. Different perspectives can be adopted according to both classical and Bayesian frameworks, and within parametric and non-parametric approaches. Applications include also the fields of epidemiology, public health, astronomy and neuroscience, ranging from one to *n*-dimensional spaces [1–6].

Many algorithms require to specify a-priori the number of clusters assumed for the data and/or their expected dimension and/or length. These settings may strongly affect the final estimation results and requires a high degree of arbitrariness on the parameters whenever no precise informations are available. Spatial scan has been proposed with wide success in the literature [5] becoming one of the main epidemiological statistics tools in disease

*Correspondence: pellin.danilo@hsr.it
[1]University Center of Statistics for the Biomedical Sciences, Vita-Salute San Raffaele University, Via Olgettina 58, 20132 Milan, Italy
[2]Johann Bernoulli Institute, University of Groningen, Nijenborgh 9, 9747 AG Groningen, Netherlands

surveillance to test the null hypothesis that geographical data are randomly distributed against a localized cluster alternative. This method and its natural extensions are of particular interest since no prior information on parameters or clusters characteristics are required. Indeed, the scan statistic is able to address any of the following interrelated purposes: a) to test if event aggregation occurs (overall clustering), b) cluster localization (detection of cluster) c) to test event distribution on a specific region (focused test).

In a multivariate setting, a challenging goal for researchers may be the identification of regions where two spatial processes - defined over a common support - show different behaviours. More in detail, the processes are allowed to share fluctuations in probability (or rate) of success. To address this type of problem, a few alternatives have been currently proposed. Most of them rely on non-parametric estimation of relative risk function by means of kernel method, as proposed in [7, 8] for environmental epidemiology data analysis.

Scan statistics methodologies have been proposed for the analysis of Poisson and Gaussian distributed random variables, categorical and many other type data. In this paper we are interested in modeling spatial distribution of a particular type of genomic data, such as viral IS retrieved by using Next Generation Sequencing (NGS) platforms [9, 10]. From a statistical point of view, the genome is interpreted as a set of $2\times3\times10^9$ independent Bernoulli random variables $B_{chr,postion,strand}$, where 1 means that a viral integration has been observed mapping to that particular genomic coordinates and 0 otherwise.

In genomics a few alternatives have been proposed to identify clusters of ISs, termed Common Integration Sites (CIS) or hotspots. The most popular in the biological literature is a gene integration frequency based method, involving Grubbs test [11] for outlier identification [12]. This approach suffers from an important limitation since ISs located outside genes and their neighborhoods are excluded from the analysis, thus leading to miss possible important intergenic CISs potentially very informative. To overcome this problem, an alternative method based on DBSCAN [13] algorithm has been proposed in [10]. The main drawback of this algorithm is the strong dependence of results on tuning parameters settings, difficult to calibrate for different sized data sets involving viral vectors with different clustering behaviours. To solve this issue, in [10] authors proposed a framework based on resampling in-silico generated ISs to select an optimal distance parameter, by controlling the probability of smaller clusters (3 events) identification. However, the impact of this procedure on bigger clusters investigation is unclear.

Insertional mutagenesis [14] provide a good setting in clinical genomics to understand the importance of comparing two integration patterns. This phenomenon is caused by virus integration trajectory within particular dangerous genomic regions, such as oncogenic regions. Since many studies revealed different patterns in site selection process among available viral vectors, a statistical procedure that allows to identify differently targeted regions represents a fundamental tool in limiting insertional mutagenesis risk. Another framework where tools for detecting genomic clustering might be extremely helpful for biological research is the investigation of active regulatory element involved in differentiation process. This can be performed by exploiting the capability of particular viral vectors, such as the *Murine Leukemia Virus* (MLV) derived vectors, in marking transcription start site of active genes [15, 16].

Some approaches have been proposed in the literature [17] based on kernel methods where two separate non-parametric kernel densities are estimated by means of Gaussian kernels. Comparative clusters of integrations (hotspots) can be selected in those genomic areas where no overlapping among confidence intervals for densities were detected. However, the arbitrary choice of smoothing parameters (bandwidth) strongly affects the detecting procedure.

In this paper we propose to overcome several problematic issues in the existing procedures, by extending the Bernoulli model proposed in [5] to the genomic field. We first study more in depth the preliminary results presented in [18] for clusters identification in univariate setting. We also propose a novel multivariate alternative, that we call Relative Scan Statistics for comparing two integration patterns by the identification of *comparative* or *relative clusters*. Multivariate extensions of scan statistics have already been proposed in the literature [19],to detect disease outbreaks by means of simultaneous analysis of different data sets. To our knowledge, there are no paper focusing on detecting *differences* among data sets using scan statistics. Finally, the proposed methods are compared to the existing ones, like the DBSCAN algorithm and the comparative hotspot [17] procedure.

The paper is organized as it follows. In Section Methods we introduce the Kulldorff scan statistics for Bernoulli data, we illustrate how the method can be used to compare two genomic data sets and the algorithm implementation is presented. In Section Results and discussion real data sets are descibed and results obtained for the univariate and multivariate analysis are discussed. Final consideration and conclusion are provided in Section Conclusions.

Methods

Kulldorff spatial scan statistics for Bernoulli model

The method proposed by [5] can be adopted to face clusters identification as a general problem. In this work, we focus on Bernoulli model, since we consider a particular

type of genomic data – derived by viral vector integration in gene therapy – that reveal presence or absence of a genomic event (namely the integration). A brief description of the underlying idea and the specification of the method for the univariate data analysis previously proposed in [18], is next introduced. Let define the whole study area under investigation as G, $\mathcal{Z}$ the collection of zones $Z \subset G$ obtained by scanning the support by means of a window of variable size.

The spatial scan statistics, S, is defined as the maximum likelihood ratio over all possible zone $Z \in \mathcal{Z}$:

$$S = \frac{\max\{L(Z)\}}{L_0} = \max_Z \left\{ \frac{L(Z)}{L_0} \right\}. \quad (1)$$

S simultaneously localizes the $Z \in \mathcal{Z}$ (chromosome, start and end coordinates) providing the maximum evidence for the presence of an hotspot and gives a measure of its goodness of fit with respect to a constant rate null hypothesis. From a computational perspective, to proceed with the calculation of Eq. 1, we need to define the total amount of success and trials available on G, respectively X and N. In addition, conditioning on a specific zone Z, n_Z and x_Z are the count of trials and success observed within Z. Finally, to identify S is necessary to maximize the likelihood:

$$L(Z, p_Z, q_Z) \propto p_Z^{x_Z}(1-p_Z)^{n_Z-x_Z} q_Z^{X-x_Z}(1-q_Z)^{(N-n_Z)-(X-x_Z)}.$$

for all $Z \in \mathcal{Z}$ by means of the following functions:

$$L(Z) = L\left(p_Z, q_Z | Z = Z_j\right) = \left(\frac{x_Z}{n_Z}\right)^{x_Z} \left(1 - \frac{x_Z}{n_Z}\right)^{n_Z - x_Z} \times \left(\frac{X - x_Z}{N - n_Z}\right)^{X-x_Z} \left(1 - \frac{X - x_Z}{N - n_Z}\right)^{(N-n_Z)-(X-x_Z)}$$

if $\frac{x_Z}{n_Z} > \frac{X-x_Z}{N-n_Z}$, and

$$L(Z) = \left(\frac{X}{N}\right)^X \left(\frac{N-X}{N}\right)^{N-X}.$$

otherwise. Under the null hypothesis, corresponding to a constant probability of success over G, the likelihood is given by:

$$L_0 = \left(\frac{X}{N}\right)^X \left(\frac{N-X}{N}\right)^{N-X}$$

for all $Z \in \mathcal{Z}$.

Multivariate extension to novel relative scan statistics for Bernoulli model

Let now introduce a novel multivariate extension of the described method for identifying the most highly significant *relative cluster*. The method is described as referred to a bivariate case, in order to ensure clarity of the underlying idea, but can be easily extended for the comparison of more than two processes.

We define a *relative cluster* as an area $Z \in \mathcal{Z}$ where two Bernoulli processes show different behaviour, in terms of success probability variation with respect to $Z^C = G \setminus Z$. Conditioning on a particular area $Z \in \mathcal{Z}$ let define p_{Z1} and p_{Z2} as the probability of being an event within Z respectively for $Process_1$ and $Process_2$ and q_{Z1} and q_{Z2} be referred to Z^C. Bernoulli trials location, assumed as known over G, can differ between the two processes. All the analyses are conditioned on the total count of observed events X_1 and X_2. The aim is here to highlight regions where the difference between probability of success in the two series is maximum and statistically significant, accounting for possible different data sets size and non-constant but shared underlying probability variations.

To measure and compare within each process the behaviour observed within/outside Z, we propose the success probability ratio $\frac{p_{Zi}}{q_{Zi}}$. The ratio takes values in R^+ and more specifically $0 \leq \frac{p_{Zi}}{q_{Zi}} < 1$ if the probability of success is lower within Z than outside and $1 < \frac{p_{Zi}}{q_{Zi}} < \infty$ otherwise.

Let now define as relative cluster for $Process_i$ with respect to $Process_j$ the region $Z \in \mathcal{Z}$ where the probability ratio $\frac{p_{Zi}}{q_{Zi}}$ is greater than corresponding ratio $\frac{p_{Zj}}{q_{Zj}}$. Conditioning on $Z \in \mathcal{Z}$ it is possible to define hypothesis system as:

$$\begin{cases} H_{0Z} : \frac{p_{Z1}}{q_{Z1}} = \frac{p_{Z2}}{q_{Z2}} \\ H_{1Z} : \frac{p_{Z1}}{q_{Z1}} \neq \frac{p_{Z2}}{q_{Z2}} \end{cases}$$

or alternatively as:

$$\begin{cases} H_{0Z} : \{p_{Z1} = k_Z q_{Z1}\} \cap \{p_{Z2} = k_Z q_{Z2}\} \\ H_{1Z} : \{p_{Z1} \neq k_Z q_{Z1}\} \cup \{p_{Z2} \neq k_Z q_{Z2}\} \end{cases} \quad (2)$$

Under the null hypothesis, the probability of success may vary over G but it must be shared among processes and characterized by the same value of k_Z. To estimate the scan statistics S, we first need to define the likelihood ratio conditioned on Z. Let now:

- N_1 and N_2 be the total count of Bernoulli trials for each process.
- X_1 and X_2 be the total count of success
- n_{1Z} and n_{2Z} be the size, in terms of trials, of the Z with respect of each series
- x_{1Z} and x_{2Z} be the success amount within Z with respect of each series

According to biological motivations related to virus integration mechanisms, supported and derived from several studies on IS data analysis, it is reasonable to assume that within each treated cell's genome, only one integration event can occur [20]. In addition, there are no

biologically meaningful reasons to suppose that any interaction between IS events occurs in distinct cells. From a modelling perspective, this is equivalent to assume independence among observations. Even more so, the two series can be assumed to be independent and the likelihood function associated to the joint model corresponds to the product of the likelihoods of each process.

$$\begin{aligned} L\left(Z, p_{Z1}, q_{Z1}, p_{Z2}, q_{Z2}\right) \propto & {p_{Z1}}^{x_{1Z}} \left(1-p_{Z1}\right)^{n_{1Z}-x_{1Z}} \\ & {q_{Z1}}^{X_1-x_{1Z}} \left(1-q_{Z1}\right)^{(N_1-n_{1Z})-(X_1-x_{1Z})} \\ & \times {p_{Z2}}^{x_{2Z}} \left(1-p_{Z2}\right)^{n_{2Z}-x_{2Z}} {q_{Z2}}^{X_2-x_{2Z}} \\ & \left(1-q_{Z2}\right)^{(N_2-n_{2Z})-(X_2-x_{2Z})} \end{aligned}$$

Conditioned on $Z = Z_j$:

$$\begin{aligned} L_{Z_j} &= \sup_{H_{1Z}} L(p_{Z1}, q_{Z1}, p_{Z2}, q_{Z2}|Z = Z_j) \\ &= \left(\frac{x_{1Z}}{n_{1Z}}\right)^{x_{1Z}} \left(1-\frac{x_{1Z}}{n_{1Z}}\right)^{n_{1Z}-x_{1Z}} \left(\frac{X_1-x_{1Z}}{N_1-n_{1Z}}\right)^{X_1-x_{1Z}} \\ &\quad \left(1-\frac{X_1-x_{1Z}}{N_1-n_{1Z}}\right)^{(N_1-n_{1Z})-(X_1-x_{1Z})} \\ &\quad \times \left(\frac{x_{2Z}}{n_{2Z}}\right)^{x_{2Z}} \left(1-\frac{x_{2Z}}{n_{2Z}}\right)^{n_{2Z}-x_{2Z}} \left(\frac{X_2-x_{2Z}}{N_2-n_{2Z}}\right)^{X_2-x_{2Z}} \\ &\quad \left(1-\frac{X_2-x_{2Z}}{N_2-n_{2Z}}\right)^{(N_2-n_{2Z})-(X_2-x_{2Z})} \end{aligned} \tag{3}$$

The maximum likelihood estimators are given by:

$$\begin{aligned} & \hat{p_{Z1}} = \frac{x_{1Z}}{n_Z};\ \hat{p_{Z2}} = \frac{x_{2Z}}{n_Z};\ \hat{q_{Z1}} = \frac{X_1-x_{1Z}}{N-n_Z}; \\ & \hat{q_{Z2}} = \frac{X_2-x_{2Z}}{N-n_Z}\ . \end{aligned}$$

By introducing the constraint $p_{Z1} = k_Z q_{Z1} \cap p_{Z2} = k_Z q_{Z2}$ as defined in the null hypothesis H_{0Z} in Eq. 2, the likelihood function becomes:

$$\begin{aligned} L_{0Z_j} &= \sup_{H_{0Z}} L\left(k_Z, q_{Z1}, q_{Z2}|Z = Z_j\right) \\ &= \propto q_{Z1} k_Z^{x_{1Z}} \left(1-q_{Z1}k_Z\right)^{n_Z-x_{1Z}} {q_{Z1}}^{X_1-x_{1Z}} \\ &\quad \left(1-q_{Z1}\right)^{(N-n_Z)-(X_1-x_{1Z})} \\ &\quad \times q_{Z2} k_Z^{x_{2Z}} \left(1-q_{Z2}k_Z\right)^{n_Z-x_{2Z}} {q_{Z2}}^{X_2-x_{2Z}} \\ &\quad \left(1-q_{Z2}\right)^{(N-n_Z)-(X_2-x_{2Z})}\ . \end{aligned} \tag{4}$$

Since a closed analytical formula for $\hat{q_{Z1}}, \hat{q_{Z2}}, \hat{k_Z}$ is computationally difficult to derive, we search for a numerical solution to calculate likelihood value L_{0Z_j} and parameters estimates. We remark that differently from the univariate case, the likelihood under the null depend on Z and is not constant over the whole study area G.

To evaluate hypothesis Eq. 2 we exploit Wilks' theorem [21] regarding procedure to test nested hypothesis.

$$\lambda_{Z_j} = 2\left(\ell_{Z_j} - \ell_{0Z_j}\right) \tag{5}$$

which is distributed under the null hypothesis according to:

$$\lambda_{Z_j} \xrightarrow{d} \chi_1^2\ . \tag{6}$$

The relative scan statistics S is defined as:

$$S = \lambda_{\hat{Z}}$$

where:

$$\hat{Z} = \{Z : \lambda_{\hat{Z}} \geq \lambda_{Z_j}\}.$$

Once $\hat{Z}$ has been identified, for potential downstream analysis it could be of interest to characterize zones by *Process*1 and *Process*2 events rate increment. This could be done by comparing the ratios $\frac{\hat{p_{Z1}}}{\hat{q_{Z1}}}$ and $\frac{\hat{p_{Z2}}}{\hat{q_{Z2}}}$ and by classifying $\hat{Z}$ as *Relative Cluster* for *Process*1 when $\frac{\hat{p_{Z1}}}{\hat{q_{Z1}}} > \frac{\hat{p_{Z2}}}{\hat{q_{Z2}}}$ and as *Relative Cluster* for *Process*2 otherwise.

We next describe a particular property of our procedure, graphically represented in Fig. 1, that might overcome the problem of dimensionality occurring in genomic applications where the total amount Z areas can quickly approach infinity. For fixed number of successes over Z, namely x_{1Z} and x_{2Z}, the number of failures - n_{1Z} and n_{2Z} - increases. This causes a progressive decrease of λ_Z, until a new event occurs within the window. Since we are interested in finding $\hat{Z}$, that corresponds to the maximum λ_Z, it is sufficient to focus on zones delimited by events (or in general success outcome).

Thus, the upper bound for the total amount of element in $\mathcal{Z}$ is $[(X_1 + X_2) * (X_1 + X_2 - 1)/2]$. Whenever is possible to define a minimum/maximum length threshold for the relative cluster, a further reduction of complexity and computational efforts holds.

The interpretation of p-value associated to relative scan statistic S must take into account the dimension of set $\mathcal{Z}$, corresponding to the total amount of performed tests. Since dependence between tests varies in strength and can be both positive or negative (it depends on the respective location of the zones associated to tests considered), we adopted the Holm-Bonferroni [22] method for *family wise error rate* (FWER) control. If S results significant, it is possible to scan the study area to identify eventual secondary significant relative cluster $\hat{Z}^*$ disjoint with $\hat{Z}$. For this purpose, we implement a sequential approach, thus ensuring I type error rate control and higher power [23]. The method consists in removing from G zone(s) previously detected as significant, redefining a new the set $\mathcal{Z}^*$ and values for N_1^*, N_2^*, X_1^*, X_2^*, n_{1Z}^*, n_{2Z}^*, x_{1Z}^* and x_{2Z}^* and sequentially performing maximization-FWER control steps.

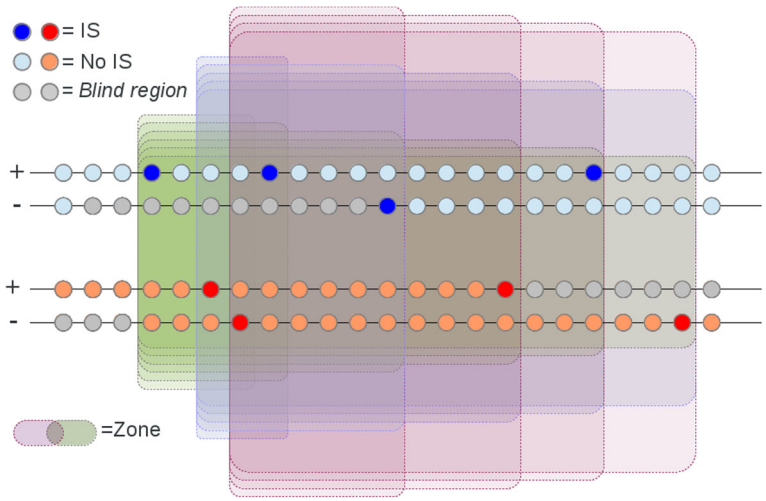

Fig. 1 Schema of the relative scan statistics. Two data sets of Bernoulli trials are represented on an hypothetical small portion of a chromosome. *Dark blue* and *red circle*: genomic coordinate in which events (IS) was observed respectively for $DataSet_1$ and $DataSet_2$. *Light blue* and *orange circle*: genomic coordinates technically investigable but no-event (no integrations retrieved). *Grey circle*: *blind region* of the genome. Transparent area: example of moving windows of variable size regarding first three IS on the left

Algorithm

We next describe the procedure for identifying relative clusters. We designed the script for genomic binary data (e.g. viral integration data). When referring in particular to gene therapy settings, the input information needed are data sets (one data sets in univariate analysis and two data sets for multivariate comparison) relative to IS coordinates (chromosome, position and strand), *blind regions* locations if available, maximum length for candidate interesting regions, L_{max}, and a minimum event counts, EC_{min}. These two input parameters play a crucial role in the definition of the final output and have a strong impact on the computational effort. Their setting must be chosen carefully, according to the data sets size and computational resources available. We suggest, to avoid to exceed half of the support G for L_{max} (clusters greater than this threshold are not very informative) and to set EC_{min} to a small value ($EC_{min} \geq 3$) in order to preserve the capability to detect possible smaller interesting regions.

A description of the algorithm in the multivariate case follows:

1. Using IS data sets and *blind regions* annotation file, calculate effective genome size X_1, X_2 and N
2. Chromosome based definition of the full set of zones, $\mathcal{Z}$.
3. Filter zones with $length(Z) \geq L_{max}$ and $EventCount(Z) \leq EC_{min}$.
4. Using IS data sets and *blind regions* annotation file, calculate effective zones size x_{1Z}, x_{2Z} and n_Z.
5. For each zone Z, calculate L_{0Z} (Eq. 3) and L_Z (Eq. 4) and corresponding λ_Z (Eq. 5).
6. Using χ^2_1 distribution, assign to each λ_Z a p-value (Eq. 6).
7. Apply multiple testing procedure.
8. If adjusted p-value associated to $\lambda_{\hat{Z}}$ is significant, define $G^* = G \setminus \hat{Z}$.
9. Calculate new X_1^*, X_2^* and N^* and restart from step 2.

The algorithm is implemented with a R script available upon request to the corresponding author.

Results and discussion

Datasets

Our application considers data sets that are comparable, for size and type of data, to those used in the literature [10] where alternative methods have been implemented to analyze and compare the profile of MLV and HIV integrations in human hematopoietic stem cells CD34+ in order to study their behaviour within the same cell type. To reduce possible technical bias the same laboratory protocol and sequencing platform was adopted.

For a detailed description of the biotechnological protocols adopted in the laboratories and subsequent bioinformatics processing steps performed, we refer to [10] and its supplementary materials. The final ISs data sets size were respectively 32631 for MLV (X_1) and 28,382 for HIV (X_2).

Due to various reasons related to sequencing technique (e.g. restriction enzymes) and mappability issue of the human genome (e.g. repeated sequences), the whole genome is not technically investigable. *Blind regions* are defined in the literature [17] as unobserved genomic portions which are strictly dependent on different laboratory settings and their distribution, position and total

amount may change a lot across studies. However, using sophisticated and computationally intensive algorithm, it is possible to calculate and predict them quite precisely.

Regarding the univariate setting, taking into account for mappability condition allow to reduce possible systematic/technical bias and to compare clustering behaviour among experiments performed under different setting. Incorporating *blind regions* information in the multivariate scan statistics makes our approach more straightforward as compared to density estimations procedure, and their asymmetry with respect to strand does not necessary require to split analysis into two strand specific tasks. In this paper we adopt results in the literature [17] for selecting predicted *blind regions* thus reducing the genome representation to a set of $N = 4398094578$ (about 2.20×10^9 each strand) independent Bernoulli random variable.

A filtering procedure was applied to $\mathcal{Z}$ generated, consisting in eliminating zones longer than 2.5×10^7 bps (considering simple difference between ISs position) and containing less then 3 ISs. This is performed in order to reduce maximization space and to focus on more biologically meaningful regions without loss of arbitrariness. The size of each zone n_Z is determined subtracting to the theoretical size (2 x ISs distance) the total amount (considering both strand separately) of *blind regions* contained.

Univariate analysis results

We run single IS series analysis with scan statistics approach and we compare the results with hotspots reported in the literature [10], obtained using DBSCAN algorithm [13] (see Supplementary Material and Method in [10] for DBSCAN setting used). Some preliminary results for this analysis has been previously published in [18], without taking into account blind regions bias and focusing only on most significant findings. In HIV data set, DBSCAN identify 2446 clusters, containing 50.6 % (14,369 IS) of the total amount of IS. Clusters' length is on average 19220 bps, but varies from a minimum of 100 to a maximum of 200500 bps. The majority (90 %) of HIV clusters are composed by 3–10 ISs.

By running univariate scan statistics methods, with a significance threshold fixed at $\alpha = 0.01$ and using Holm [22] procedure for adjusting p-values, 282 clusters are identified (see Table 1 and Additional file 1), corresponding to 45.5 % (12,935 IS) of the HIV data set. Hotposts length is between 4053 bps and 8,264,000 bps, on average 742,000, and ISs content vary from 4 to 651.

For MLV, DBSCAN identifies 3497 clusters, corresponding to 65.3 % (21,307 IS) of MLV data set. Clusters are on average 8385 bps long, the observed minimum and maximum length are respectively 19 bps and 78,530 bps. Using univariate scan statistics, 803 clusters has been identified (see Table 2 and Additional file 1), grouping 18,388 ISs equivalent to 56.3 % of MLV data set. Length mean value results equal to 270,400 bps, with a minimum of 1932 bps and a maximum of 5,449,000 bps.

In general, the two methods provide consistent results and highlight different clustering behaviour proper of the two viral vectors, in particular in terms of clusters length and events density. Both methods confirm HIV preference for active transcriptional units, such as coding regions, typically wider than regulatory regions preferentially targeted by MLV viral vectors. This characteristic is well captured in particular by the success probability ratio, $\frac{p_{\hat{HIV}_Z}}{q_{\hat{HIV}_Z}}$ for HIV candidate hotspots, generally lower with respect to MLV counterpart, $\frac{p_{\hat{MLV}_Z}}{q_{\hat{MLV}_Z}}$ (see Tables 1, 2 and Additional files 1 and 2). The count distributions of ISs belonging to the same cluster are similar across virus type but not across methods. Taking into account summary data and graph in Fig. 2, is clear that DBSCAN lead to a bigger selection of over targeted regions than scan statistics, characterized by both smaller length and size. We remark that both methods suggest a clear difference in terms of length between vectors type and a homogeneity for size distributions, reinforcing the findings known about virus preferences.

Table 1 List of first 10 clusters identified in HIV data by scan statistics

S	Chr	Start	End	IS count	$\frac{p_{\hat{HIV}_Z}}{q_{\hat{HIV}_Z}}$	Raw *p*-value	Adj *p*-value
2463.2	chr11	63175583	68111375	651	17.2	<2e-16	<2e-16
1795.1	chr16	95090	3640598	444	19.6	<2e-16	<2e-16
1390.0	chr17	70634094	73732441	386	15.5	<2e-16	<2e-16
1189.8	chr17	75720251	78604915	323	16.2	<2e-16	<2e-16
1063.8	chr3	46999507	52978572	424	8.5	<2e-16	<2e-16
1046.8	chr6	30563526	33532447	325	12.6	<2e-16	<2e-16
1041.8	chr9	138245676	139772487	224	26.9	<2e-16	<2e-16
732.0	chr8	144469820	146194757	188	18.1	<2e-16	<2e-16
721.1	chr19	572963	3118599	209	14.3	<2e-16	<2e-16
629.1	chr17	1483915	4578114	238	9.2	<2e-16	<2e-16

Table 2 List of first 10 clusters identified in MLV data by scan statistics

S	Chr	Start	End	IS count	$\frac{p_{\hat{MLV_Z}}}{q_{\hat{MLV_Z}}}$	Raw *p*-value	Adj *p*-value
386.5	chr20	51646845	51991770	89	22.8	<2E-16	<2E-16
326.4	chr20	10362242	10450134	55	51.8	<2E-16	<2E-16
318.4	chr17	26646082	26672265	41	131.1	<2E-16	<2E-16
302.6	chr17	76325116	76460372	56	39.5	<2E-16	<2E-16
285.6	chr19	59566413	59591310	37	127.9	<2E-16	<2E-16
284.6	chr21	38671040	39311896	90	12.2	<2E-16	<2E-16
279.2	chr17	51718847	53782415	142	6.2	<2E-16	<2E-16
278.7	chr1	25046795	28847012	183	4.7	<2E-16	<2E-16
267.7	chr18	72291047	72971441	87	11.6	<2E-16	<2E-16
264.4	chr12	6084417	10441567	197	4.2	<2E-16	<2E-16

We next investigate how methods agree in identifying locations of most significant regions. DBSCAN clusters are sorted in terms of size, i.e. the amount of IS falling within cluster limits, to allow for possible the comparison with scan statistics results.

The list of the first 10 Most Significant Clusters (MSCs) coordinates discovered by Scan Statistics in HIV data set are showed in Table 1, together with some related measures. The complete list is available in Additional file 1. The most significant cluster is located at chromosome 11, interval 63,175,583;68,111,375 and within the same region DBSCAN identifies 40 out of 2446 distinct clusters, including the top 2 for ISs content (interval 65,586,752;65,736,062, 110 ISs and interval 66651503-66776194, 96 ISs). The second most significant cluster, named MSC_2 is located on chromosome 16, interval 71,294,851;77,821,445 and is composed by 610 IS. Within this genomic region, DBSCAN reported 38 clusters, including the third in terms of ISs.

Univariate analysis results for MLV data set are tabulated in Table 2 and Additional file 2. Region on chromosome 20, interval 51,646,845;51,991,770 contain 89 ISs and is suggested to be the most evident hotspot region for MLV vector. Within the same interval, DBSCAN identify 8 distinct clusters, but not among the top in ranking. The second, MSC_2,is on chromosome 20, interval 10,362,242;10,450,134 and is composed by 55 ISs. It overlaps with the 50-th hotspost retrieved using DBSCAN. A perfect correspondence is observed between MSC_3 and the 4-th cluster derived from DBSCAN, both located

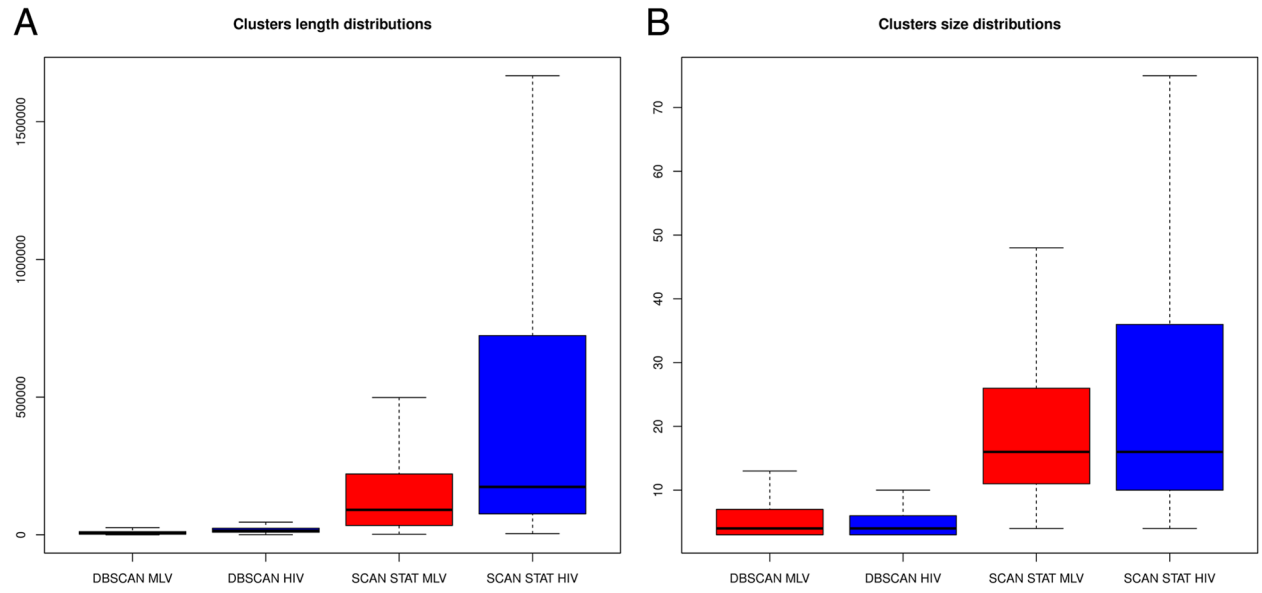

Fig. 2 a Length distributions of clusters identified by DBSCAN and scan statistics algorithm in MLV and HIV data sets. **b** Size distributions of clusters identified by DBSCAN and scan statistics algorithm in MLV and HIV data sets

on chromosome 17, interval 26,659,383;26,672,265. Conversely, the first cluster calculated using DBSCAN is on chromosome 22 27,525,356;27,545,150, its size is 42 ISs and corresponds to 85-th MLV scan statistics derived cluster.

In simple terms we reveal that the most important part of the difference in identifying the total amount of clusters can be attributed to a *fragmentation* of scan statistics cluster in more DBSCAN clusters. Despite that, an overall clear correspondence in terms of localization was observed, while agreement in ranking is more dependent on clustering behaviour.

Multivariate analysis results

The Relative Scan Statistics identified 292 genomic intervals showing a difference in targeting propensity by the two viral vectors. Totally, 174 of them could be classify as relative clusters for MLV. Conversely 119 of them are labeled as HIV relative clusters. Chromosome 17 is the one with the highest amount of detected interesting regions (Fig. 3).

We remark that the a big advantage of the proposed methods is the ability to detect both long and short regions. Long relative cluster can be usually easily visualized by using density estimate superposition. Short relative clusters or closed *opposite* relative cluster are much more difficult to detect, due to the smoothness of kernel estimator. This is in our opinion a crucial feature of our proposal, and it may be of particular utility for data analysis and for vector safety assessment. We now compare our list with the suggested 100 regions (51 for MLV and 49 for HIV) proposed in the literature [17].

Although the total amount of interesting regions might vary considerably, it is not clear which one performs better since true differently targeted regions are not known. In our opinion, since the underlying biological mechanism and target site selection process are deeply different (MLV belongs to the gammaretroviral genus and HIV to the lentiviral), a longer list of candidate regions can be considered more realistic.

This idea seems to be supported by visual comparison of chromosome based kernel density estimations. The length and the size of regions identified using the two different approach are similar (Fig. 4), nevertheless [17] method discriminates between MLV and HIV regions, since the latter are longer and include more events. By comparing intervals localization and their overlapping, we can highlight that all previously identified regions are associated to a Relative Scan Statistics derived clusters.

In Table 3 first 20 relative clusters are reported (complete list available as Additional file 3). For both methods, the most significant regions are labeled as cluster for HIV vector, suggesting that it is easier to detect wider regions characterized by moderate increase of targeting

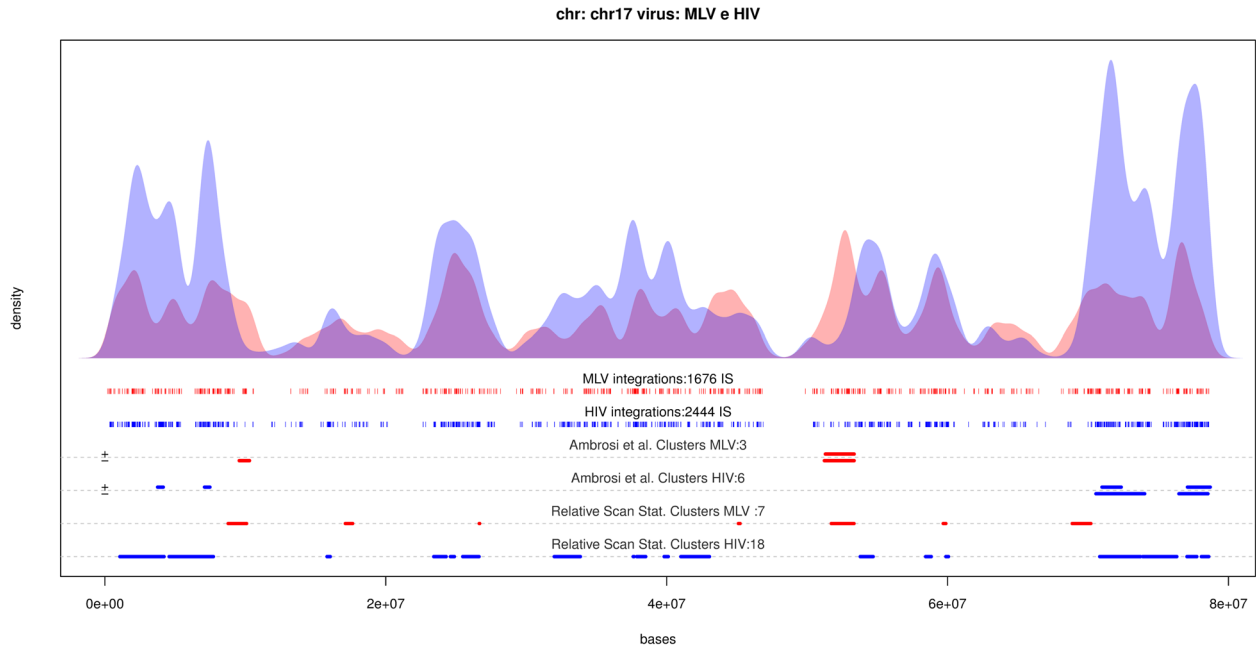

Fig. 3 HIV and MLV IS distributions on chr 17. HIV and MLV IS distributions on chromosome 17 estimated by means of Gaussian kernel with unbiased cross validation bandwidth selection (*blue curve* and *red curve* respectively). Comparative hotspots reported in [17] correspondent to segments indicated on third line in red (MLV comparative hotspot) and fourth line in blue (HIV comparative hotspots) taking into account for strand annotation. Fifth and sixth lines are dedicated to relative scan statistics. First two significant cluster identified using relative scan statistics with no correspondent comparative hotspots are highlighted (*black box*)

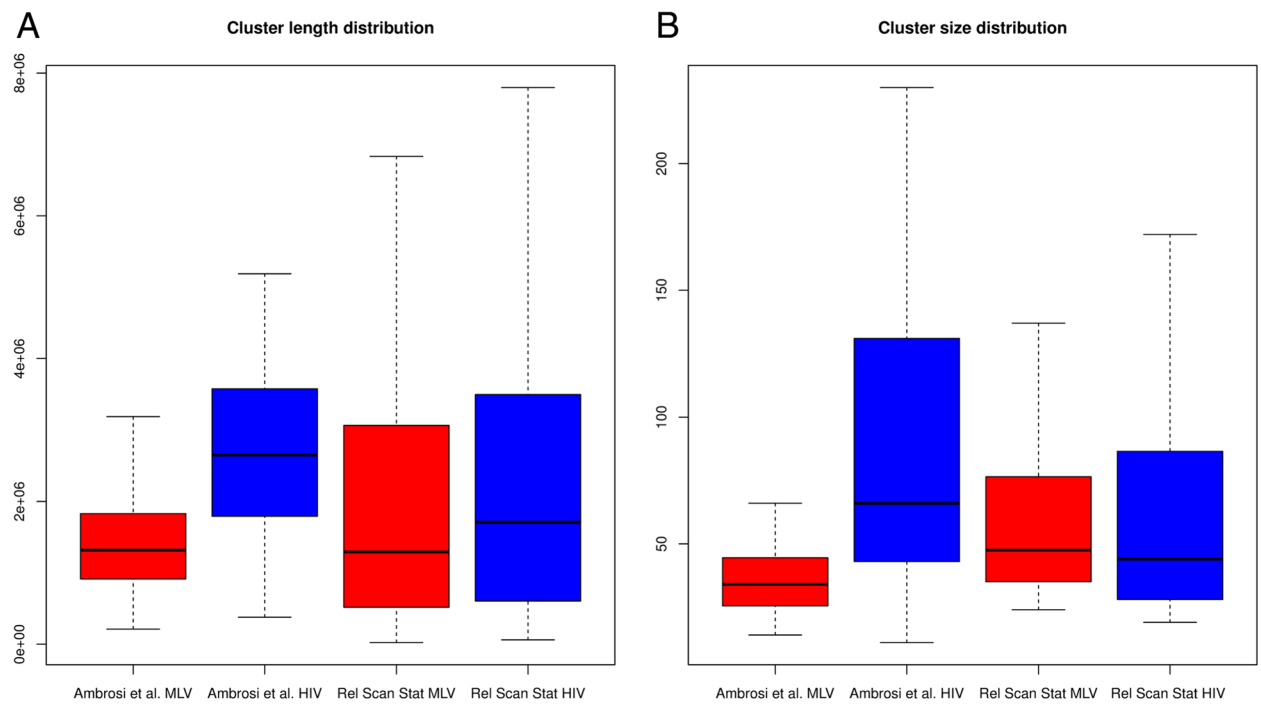

Fig. 4 a Length distributions of clusters identified by Ambrosi et al. methods and relative scan statistics algorithm in MLV and HIV data sets. **b** Size distributions of clusters identified by Ambrosi et al. methods and relative scan statistics algorithm in MLV and HIV data sets

Table 3 List of relative clusters identified by relative scan statistics

S	Chr	Start	End	HIV IS	MLV IS	$\log\left(\frac{\frac{\hat{p_{HIV_Z}}}{q_{HIV_Z}}}{\frac{\hat{p_{MLV_Z}}}{q_{MLV_Z}}}\right)$	Type	Adj *p*-value
474.1	chr11	63153734	68347426	659	129	1.91	hiv	<2E-16
450.9	chr6	30095760	33488528	332	7	4.49	hiv	<2E-16
434.2	chr16	95090	3561021	430	41	2.74	hiv	<2E-16
260.9	chr17	70835415	73732441	372	75	1.86	hiv	<2E-16
227.0	chr3	47041751	52978572	422	119	1.47	hiv	<2E-16
219.4	chr9	134493480	139818935	307	60	1.89	hiv	<2E-16
213.5	chr17	77047796	77746204	172	7	3.70	hiv	<2E-16
191.9	chr8	144548769	146194757	182	15	2.89	hiv	<2E-16
122.0	chr19	1027304	6006371	292	104	1.20	hiv	<2E-16
115.4	chr22	48983597	49573459	115	11	2.71	hiv	<2E-16
105.6	chr21	37559632	39311896	9	126	-3.02	mlv	<2E-16
102.1	chr19	54074745	55048471	122	18	2.21	hiv	<2E-16
99.3	chr17	1069411	4213267	229	79	1.23	hiv	<2E-16
96.4	chr1	153550587	154168170	90	7	2.94	hiv	<2E-16
91.8	chr18	70832211	73059134	6	103	-3.26	mlv	<2E-16
91.5	chr17	4573721	7723628	194	62	1.32	hiv	<2E-16
86.5	chr20	49745347	52129713	7	102	-3.07	mlv	<2E-16
86.0	chr12	11729500	14430150	8	105	-2.95	mlv	<2E-16
83.3	chr20	60901158	62379063	109	19	2.02	hiv	<2E-16
81.3	chr6	6536008	13289623	22	141	-2.13	mlv	<2E-16

rate, typical of HIV vector, than shorter genomic portions with high increase of targeting probability as observed for MLV derived vector. To compare also the ranking of regions, we sorted the results obtained in [17] using p-value associated to Fisher exact test calculated for assess regions significance. Due to strand specificity, top 6 results in reported in [17] map to the top 3 regions in Table 3. However the known method missed the firsts 2 regions both located on chromosome 19 on p-arm, Fig. 5 which is a gene dense portions of the genome. Gene density is known to be a particular feature in the genome able to attract particularly HIV derived vectors and this support our result.

Graphs for remaining chromosomes are available in Additional file 4.

Conclusions

In this paper we present two methods for clustering identification of genomic events based on scan statistics approach. Results retrieved from both methods are consistent with the biological literature and findings thus revealing deep biological differences between integration process and target sites selection characterizing different viral vectors. Speculating on cluster dimensions and length, our analysis confirms the well known preferences of MLV in integrating more likely in regulatory elements or in general over small genomic interval, whereas HIV integrates over wider regions corresponding to active coding elements. Independently from the total amount of identified interesting regions, a substantial spatial overlap between results was observed in HIV data set, as regarding both localization and significance. For MLV data set, a good agreement is showed in terms of localization but for significance ranking. The intrinsic behaviour of HIV probably helps this results correspondence, since aggregation is less strong than MLV but affects wider regions, leading to cluster formed by many IS rewarded by DBSCAN ranking scheme based on dimension. For MLV instead, generally the aggregation tendency is characterized by higher event density but limited to narrow genomic intervals and less ISs.

Relative Scan Statistics seems to be able to identify regions characterized by unshared variation of events rate, potentially allowing for focusing downstream analysis only on differently targeted regions. This may help clinicians/researcher in improve viral vectors safety. The results obtained agree with previous published literature and avoid the necessity to split analysis according to strands.

In conclusion, starting from a probabilistic approach based on estimation and comparison of probability of success, we recommended scan statistics as a fundamental

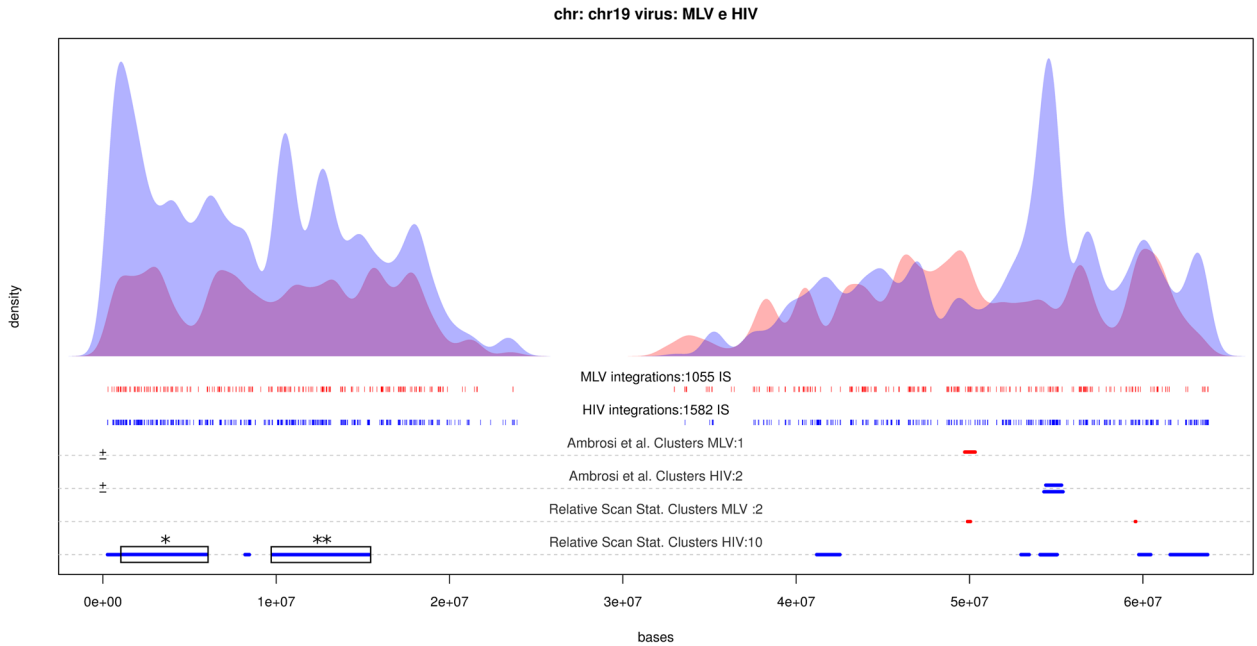

Fig. 5 HIV and MLV IS distributions on chr 19. HIV and MLV IS distributions on chromosome 19 estimated by means of Gaussian kernel with unbiased cross validation bandwidth selection (*blue curve* and *red curve* respectively). Comparative hotspots reported in [17] correspondent to segments indicated on third line in red (MLV comparative hotspot) and fourth line in blue (HIV comparative hotspots) taking into account for strand annotation. Fifth and sixth lines are dedicated to relative scan statistics. First two significant cluster identified using relative scan statistics with no correspondent comparative hotspots are highlighted (*black box*)

inferential tool able to exploit an hypothesis testing procedure to sort candidate regions in terms of significance instead of size or additional testing procedure.

Additional files

Additional file 1: Table S1. Full list of HIV clusters. (PDF 49 kb)

Additional file 2: Table S2. Full list of MLV clusters. (PDF 80 kb)

Additional file 3: Table S3. Full list of relative clusters. (PDF 51 kb)

Additional file 4: Figure S1. Relative scan statistics results for remaining chromosomes. (PDF 5089 kb)

Acknowledgements
The authors thank all members of the CUSSB for helpful suggestions.

Declaration
Publication charges for this article were funded by the grant FFC 27/2014 of the Fondazione per la Ricerca sulla Fibrosi Cistica.
This article has been published as part of BMC Bioinformatics Volume 17 Supplement 11, 2016. Selected articles from the 11th International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics (CIBB 2014). The full contents of the supplement are available online https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-17-supplement-11.

Availability of data and materials
The raw sequence reads for HIV and MLV IS are available the GenBank Short Read Archive under the accession number SRA024251.1.

Authors' contributions
Contributed statistical methodology: DP CDS. Wrote the paper: DP CDS. Implemented the methods: DP. Both authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Published: 22 September 2016

References

1. Naus J. The distribution of the size of maximum cluster of points on the line. J Am Stat Assoc. 1965;60:532–8.
2. Naus J. Clustering of random points in two dimensions. Biometrika. 1965;52:263–7.
3. Loader CR. Large-deviation approximation to the distribution of scan statistics. Ann Appl Probab. 1991;23:751–71.
4. Chen J, Glaz J. Two dimensional discrete scan statistics. Stat Probab Lett. 1996;31:59–68.
5. Kulldorff M. A spatial scan statistic. Commun Stat:Theory Methods. 1997;26:1481–1496.
6. Kulldorff M, Athas W, Feuer E, Miller B, Key C. Evaluating cluster alarms: A space-time scan statistics and brain cancer in Los Alamos. Am J Public Health. 1998;88:1377–1380.
7. Kelsall JE, Diggle PJ. Kernel estimation of relative risk. Bernoulli. 1995;1:3–16.
8. Kelsall JE, Diggle PJ. Non parametric estimation of spatial variation in relative risk. Stat Med. 1995;14:2335–343.
9. Bushman F, Lewinski M, Ciuffi A, Barr S, Leipzig J, Hannenhalli S, Hoffmann C. Genome-wide analysis of retroviral DNA integration. Nat Microbiol. 2005;3:848–58.
10. Cattoglio C, Pellin D, Rizzi E, Maruggi G, Corti G, Miselli F, Sartori D, Guffanti A, Di Serio C, Ambrosi A, De Bellis G, Mavilio F. High-definition mapping of retroviral integration sites identifies active regulatory elements in human multipotent hematopoietic progenitors. Blood. 2010;116:5507–517.
11. Grubbs FE. Sample criteria for testing outlying observations. Ann Math Stat. 1950;21:27–58.
12. Biffi A, Bartolomae CC, Cesana D, Cartier N, Aubourg P, Ranzani M, Cesani M, Benedicenti F, Plati T, Rubagotti E, et al. Lentiviral vector common integration sites in preclinical models and a clinical trial reflect a benign integration bias and not oncogenic selection. Blood. 2011;117(20):5332–5339.
13. Ester M, Kriegel H-P, Sander J, Xu X, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd. 1996;96(34):226–231.
14. Hacein-Bey-Abina S, Garrigue A, Wang GP, Soulier A, Lim J, et al. Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. J Clin Investig. 2008;118:3132–142.
15. Cattoglio C, Facchini G, Sartori D, A A, Antonelli A, Miccio A, Cassani B, Schmidt M, von Kalle C, Howe S, Thrasher AJ, Aiuti A, Ferrari G, Recchia A, Mavilio F. Hot spots of retroviral integration in human CD34+ hematopoietic cells. Blood. 2007;110:1770–1778.
16. Ambrosi A, Di Serio C. Vectors and integration in gene therapy: Statistical considerations. J Comput Sci Syst Biol. 2009;2:117–23.
17. Ambrosi A, Glad I, Pellin D, Cattoglio C, Mavilio F, Di Serio C, Frigessi A. Estimated comparative integration hotspots identify different behaviors of retroviral gene transfer vectors. PLoS Comput Biol. 2011;7:12.
18. Pellin D, Di Serio C. Clusters identification in binary genomic data: The alternative offered by scan statistics approach. Comput Intell Methods for Bioinforma Biostat. 2014;1:149–58.
19. Kulldorff M, Mostashari F, Duczmal L, Yih K, Kleinman K, Platt R. Multivariate spatial scan statistics for disease surveillance. Stat Med. 2007;26:1824–1833.
20. Aiuti A, Biasco L, Scaramuzza S, Ferrua F, Cicalese MP, Baricordi C, Dionisio F, Calabria A, Giannelli S, Castiello MC, et al. Lentiviral hematopoietic stem cell gene therapy in patients with wiskott-aldrich syndrome. Science. 2013;341(6148):1233151.
21. Wilks SS. The large-sample distribution of the likelihood ratio for testing composite hypotheses. The Annals of Mathematical Statistics. 1938;9(1):60–62.
22. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat. 1979;6:65–70.
23. Zhang Z, Assunção R, Kulldorff M. Spatial scan statistics adjusted for multiple clusters. Journal of Probability and Statistics. 2010;2010.