Published online 7 June 2011 *Nucleic Acids Research, 2011, Vol. 39, No. 15* ***e105***
doi:10.1093/nar/gkr447

Computational identification of insertional mutagenesis targets for cancer gene discovery

Johann de Jong[1,2], Jeroen de Ridder[1,3], Louise van der Weyden[4], Ning Sun[1], Miranda van Uitert[5], Anton Berns[5,6], Maarten van Lohuizen[2,6], Jos Jonkers[6], David J. Adams[4] and Lodewyk F. A. Wessels[1,2,3,*]

[1]Bioinformatics and Statistics, The Netherlands Cancer Institute, Plesmanlaan 121, 1066CX Amsterdam, [2]Netherlands Consortium for Systems Biology, Science Park 904, 1098XH Amsterdam, [3]Delft Bioinformatics Lab, Faculty of EEMCS, TU Delft, Mekelweg 4, 2628 CD Delft, The Netherlands, [4]Experimental Cancer Genetics Laboratory, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom, [5]Academic Medical Center, University of Amsterdam, Meibergdreef 9, 1105 AZ Amsterdam and [6]Division of Molecular Biology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066CX Amsterdam, The Netherlands

Received January 3, 2011; Revised May 11, 2011; Accepted May 13, 2011

ABSTRACT

Insertional mutagenesis is a potent forward genetic screening technique used to identify candidate cancer genes in mouse model systems. An important, yet unresolved issue in the analysis of these screens, is the identification of the genes affected by the insertions. To address this, we developed Kernel Convolved Rule Based Mapping (KC-RBM). KC-RBM exploits distance, orientation and insertion density across tumors to automatically map integration sites to target genes. We perform the first genome-wide evaluation of the association of insertion occurrences with aberrant gene expression of the predicted targets in both retroviral and transposon data sets. We demonstrate the efficiency of KC-RBM by showing its superior performance over existing approaches in recovering true positives from a list of independently, manually curated cancer genes. The results of this work will significantly enhance the accuracy and speed of cancer gene discovery in forward genetic screens. KC-RBM is available as R-package.

INTRODUCTION

Large-scale insertional mutagenesis screens using retroviruses and transposons are of great importance in cancer research. By integration into the host DNA, retroviruses and transposons can mutate the genome. This process is referred to as insertional mutagenesis. Insertional mutagenesis can disrupt cellular processes, alter gene expression and thereby cause cancer. For this reason, large-scale insertional mutagenesis screens have been successfully employed to identify new putative cancer genes, see e.g. (1–5); J. Kool (personal communication). In addition, retroviral vectors have been shown to be useful in gene therapy, and transposon-based systems also show great potential for this same purpose. However, it is currently still very difficult to predict which surrounding genes will be affected by insertions.

To identify potential cancer genes from an insertional mutagenesis screen, the initial step typically involves the definition of common insertion sites (CISs), see e.g. (1,3,6–9). Insertions are clustered based on inter-insertion-distance and clusters that are unlikely to occur by chance are declared CISs. The CISs are then manually mapped to putative target genes. This manual mapping could potentially introduce biases. For example, known cancer genes may be preferred, thus potentially and unintentionally excluding novel cancer genes. An additional drawback of this approach is that in defining CISs, properties of individual insertions, such as distances to genes and orientation relative to genes are disregarded.

In contrast, nearest-gene mapping (NGM), maps each insertion to the nearest gene [e.g. (10)]. While this approach does operate on individual insertions, and takes the distance of insertions to genes into account, it still disregards the relative orientation of insertions, and does not aggregate insertion data across tumors.

*To whom correspondence should be addressed. Tel: +31 20 5127987; Fax: +31 20 6691383; Email: l.wessels@nki.nl
Present address:
Ning Sun, Pluton IT, Rotterdamseweg 183c, 2629HD Delft.

The orientation of an insertion occurring in the immediate upstream promotor region of a gene is a highly important modulator of the effect of that insertion on the gene. More specifically, if the viral promoter has the same orientation as the host promoter it can take over its function (4). For larger upstream and downstream distances from genes, relative orientation also plays a role: enhancing insertions are predominantly oriented away from target genes (4). It is therefore clear that the orientation of an insertion should be taken into account when determining putative target genes. Furthermore, since the nearest gene is not necessarily the only or best target gene, it is important to allow the assignment of multiple target genes to a single insertion.

To address the issues described above, we developed Kernel Convolved Rule Based Mapping (KC-RBM). KC-RBM integrates GKC (7), a method for identifying statistically significant CISs, with rule-based mapping of individual insertions to genes. Without user intervention, KC-RBM maps insertions to genes based on orientation-dependent windows defined around transcripts, and exploits the information contained in the repetitive occurrence of insertions at a given locus across tumors, i.e. CIS information. We perform extensive analyses of associations between insertion occurrence and same-sample gene expression to evaluate the parameter choices for KC-RBM. We demonstrate the benefits of KC-RBM in cancer gene discovery through the more accurate identification of target genes from two insertional mutagenesis screens, a Murine Leukemia Virus (MuLV) screen and a Sleeping Beauty (SB) transposon screen. KC-RBM represents the first ever approach for mapping SB transposon insertions to target genes.

MATERIALS AND METHODS

Data sets

Murine Leukemia Virus (MuLV) data. In total 20312 MuLV insertions were extracted from 8 insertional mutagenesis screens. These screens produced 1020 tumors in total, and were produced in mice with various genetic backgrounds (3); J. Kool (personal communication). For a subset of 1986 insertions in 97 samples (p19 ko, p53 ko, and wild-type), Illumina MouseWG-6 v2.0 expression data was available and used.

Sleeping Beauty (SB) transposon data. 58266 SB transposon insertion loci were extracted from 255 lymphomas collected from T2/Onc2 and Rosa26 SBase mice. For a subset of 26955 insertions in 135 samples, Illumina MouseWG-6 v2.0 expression data was available and used.

List of mappings

KC-RBM. Maps insertions to multiple putative target genes, using four window sizes, one for upstream-sense insertions one for upstream-antisense insertions, one for downstream-sense insertions and one for downstream-antisense insertions (with respect to transcription start site). Per transcript, these window size parameters are flexibly applied using two additional parameters, a GKC scale parameter (7) and a orientation homogeneity parameter. A gene is a target gene of an insertion if at least one of its transcripts is targeted. As an additional step in selecting a single target gene for each insertion, a prioritization can be made among the target genes identified by KC-RBM according to the number of times they were targeted by all insertions taken together. Then select the gene with the highest count to be the single target gene for that insertion.

Nearest-gene mapping (NGM). For each insertion, find the nearest gene start site, and select this gene to be the single target gene of that insertion. This method is compared to KC-RBM.

CIS nearest-gene mapping (CIS-NG). CISs are detected using GKC (7). The peak of each CIS is then mapped to its nearest gene start site. This method is compared to KC-RBM.

Methods

Aligning genes. Figure 1. The set of tumor samples was reduced to the set for which expression data was available ($n = 97$). All gene start sites were aligned with respect to location as well as orientation, and expression values were z-normalized per gene across samples. For all genes, all insertions were identified in a window of 400 kb around these genes. All resulting (relative insertion locus, z-normalized gene expression) pairs were regarded as points in the (x,y) plane, and were then binned along the y-axis, making a distinction between insertions occurring in sense orientation relative to the gene and in antisense orientation relative to the gene start site, and normalizing gene length. The insertion density was computed by binning the insertions, and computing the number of insertions per base pair for each bin. These values were then normalized to a scale from 0 to 1.

The influence of window size. Figure 3. The set of tumor samples was reduced to the set for which expression data was available. For each window size value and each gene, the following approach was taken. Tumor samples were divided in two groups. The first group contained the samples for which at least one insertion was mapped to that gene. The second group contained the samples for which no insertion was mapped to that gene. Between these two groups, a Wilcoxon-score was computed for elevated expression in the first group. Having computed this Wilcoxon-score for all genes, a significance threshold was determined per mapping by permuting ($n = 10\,000$) gene-wise expression profiles across samples with respect to gene-wise insertion profiles across samples, and setting a 5% significance threshold. Per window, each gene with at least one insertion was classified as significant or not significant exactly once. Per gene, each insertion is counted only once. Note that when computing the statistics for one of four window sizes, the other window sizes were set to zero. Furthermore, for all insertions within transcripts, association of insertion occurrence with increased expression levels was computed while

disregarding the insertions outside transcripts. Permutation thresholds (5%) were calculated per window size.

The influence of the GKC scale. Figure 4. For each KC-RBM scale, insertions were mapped to genes, and numbers and fractions of significant genes were computed as described above. KC-RBM was performed using window sizes (20 kb, 120 kb, 40 kb, 5 kb) (MuLV) and (20 kb, 10 kb, 25 kb, 5 kb) (SB transposon) for (upstream-sense, upstream-antisense, downstreamsense, downstream-antisense) insertions, and an orientation homogeneity fraction of 0.75. For each scale, all insertions (for MuLV all insertions from screen 1: p19 ko, p53 ko and wild-type) were mapped to genes, but insertion-expression association was necessarily only computed for the samples for which expression data was available.

Comparing KC-RBM, RBM, CIS-NG and CIS-manual mapping. Figure 5. All insertions were mapped using KC-RBM, setting the window sizes to (20 kb, 120 kb, 40 kb, 5 kb) (MuLV) and (20 kb, 10 kb, 25 kb, 5 kb) (SB transposon) for (upstream-sense, upstream-antisense, downstream-sense, downstream-antisense) insertions. The orientation homogeneity fraction was set to 0.75, and the scale was set to 10 kb (MuLV) and 2 kb (SB transposon).

For KC-RBM, lists of top 20 CTGs were obtained by counting for each gene the number of times it was targeted, and then sorting this list, based on the number of times a gene was identified as a target. Specifically for SB transposon insertions, the CTGs were corrected for the fact that SB transposons only integrate at TA-sites: in determining CTGs, SB transposon insertions were each weighted by 1 divided by the local TA-density determined using the same kernel width as was used for the mapping of insertions (2 kb). The total SB transposon CTG score across all genes was normalized to be equal to the total number of insertions. For NGM, also the top 20 CTGs were determined. CISs were detected using GKC (7), with a scale of 30 kb. The 20 CISs with the highest peaks were then mapped to their nearest gene start site.

For both the MuLV and the SB transposon data set, the top 20 results as well as the overall results were compared to a reference list. For MuLV a manually curated list based on the same data set exists (4). The complete lists of genes identified by the three methods KC-RBM, NGM and CIS-NG were compared to this manually curated list (CIS-M) with respect to presence and rank in this list. Regarding the presence of genes in either of the three methods in the CIS-M list, the three lists were made the same size by taking the top N of each list (where N is the length of the shortest list), to allow for a fair comparison. For each resulting list, the number of genes in that list also present in the CIS-M list was counted. For the comparison between the top two methods, KC-RBM and NGM, significance of the difference in numbers present in the CIS-M list was determined by permutation ($n = 100\,000$). Regarding rank, the following steps were taken. First, all lists were restricted to genes also occurring in CIS-M. Then, the three lists were made the same size by selecting only the top N from each list (where N is the length of the shortest list). This is necessary since the highest ranking CISs and CTGs are the easiest to retrieve, which may negatively affect the average rank of longer lists. Then, for each of the three lists, the average rank in that list of the genes also present in the manually curated list was calculated. For the comparison between the top two methods, KC-RBM and NGM, significance of the difference in average rank was determined by permutation ($n = 100\,000$).

For the SB transposon insertions a similar approach was taken, using as a reference the Cancer Gene Census (11), a list of human cancer-related genes. Mouse homologs were identified by mapping the human EntrezGene identifiers to mouse EntrezGene and Ensembl identifiers using the Bioconductor biomaRt 2.2.0 package (12).

RESULTS

Insertion occurrence and gene expression

Since the orientation of an insertion relative to a target gene and the distance of an insertion to a target gene determine how an insertion may activate that gene, one may expect association between orientation, distance and gene expression. Figure 1 depicts an alignment of all genes. A point represents the average normalized deviation of the gene expression from the mean as a function of the distance between a gene and an insertion. A technical explanation of this figure can be found in the 'Materials and Methods' section. For both the MuLV insertions and the SB transposon insertions, it can be seen that association of insertion occurrence with gene expression of nearby genes is dependent on the distance and orientation of the insertion to the gene.

For MuLV, Figure 1a shows higher average expression deviation for insertions inside and near genes (demarcated by the two vertical black lines). There is a clear peak in expression levels for antisense insertions (red) just upstream of the gene start site. For sense insertions (green), a slightly less pronounced peak can be seen just downstream of the gene start site. These observations are consistent with mechanisms described in the literature by which retroviral insertions affect their target genes (4,5,13). The insertion density across all aligned genes is plotted in black below the expression values, and demonstrates an explicit preference of MuLV insertions for loci near the gene start site, as previously observed (14).

For the SB transposon, Figure 1b suggests that some association does exist, although much less pronounced when compared to the retroviral case. Especially for the sense insertions inside genes there is some elevation in expression. The insertion density (depicted in black, below the binned z-values) shows that SB transposon insertions are predominantly found inside genes.

KC-RBM

Mechanisms described in the literature (4,5,13), supported by our own observations (Figure 1), suggest that insertions should be mapped to putative target genes using

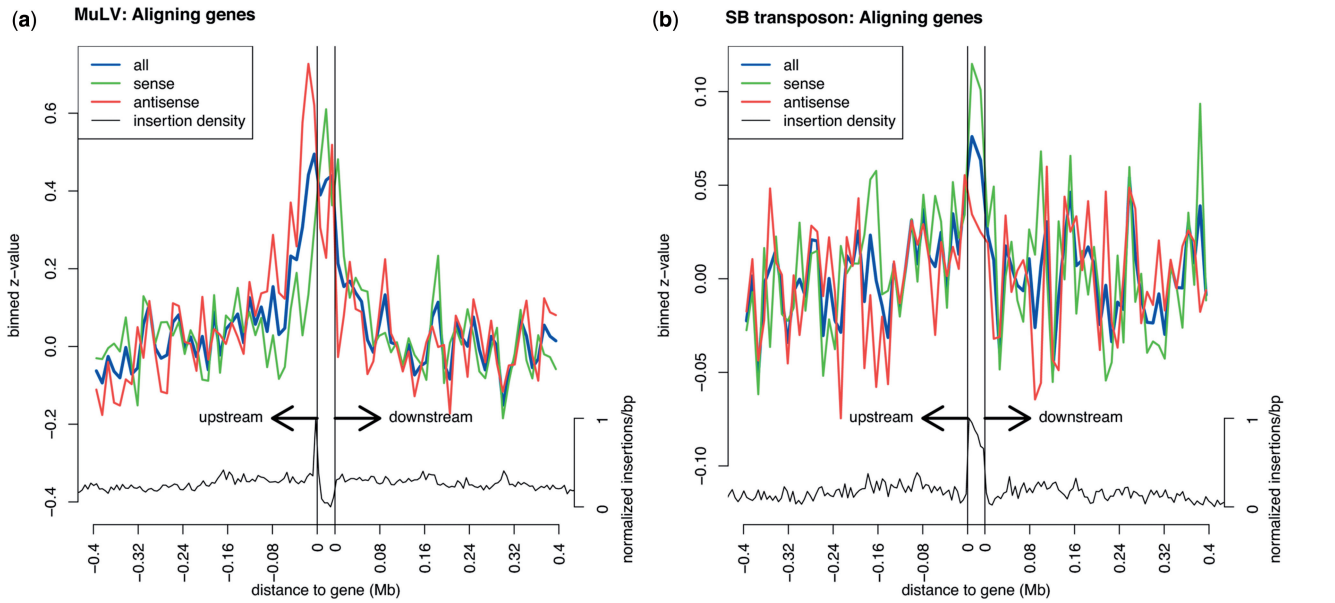

Figure 1. Normalized deviation of gene expression from the mean as a function of insertion distance, for (**a**) MuLV and (**b**) SB transposon. For all genes, all insertions are identified in a window of 500 kb around these genes, from the gene start site and from the gene termination site. All genes are then aligned with respect to location as well as orientation. *z*-normalized gene expression values are associated with the relative locations of the insertions within the 500 kb window. For all genes and insertions taken together, these expression values are binned, and the distinction is made between insertions occurring in sense orientation relative to the gene (green) and in antisense orientation relative to the gene (red). The blue line represents the all insertions taken together. The (aligned) insertion density is plotted in black below the binned *z*-values. The gene boundaries are represented by two vertical black lines.

orientation-dependent windows defined on either side of transcripts. Depending on the orientation and location of an insertion, the insertion will fall within or outside the relevant mapping window. When the insertion falls within a given window, it will be mapped to the associated gene. This approach, which we will call rule based mapping (RBM) is outlined in Figure 2a. It uses four window size parameters, for upstream-sense, upstream-antisense, downstream-sense and downstream-antisense insertions.

The window sizes used by RBM provide strict boundaries outside of which insertions are not mapped to a gene. However, as it is presented in Figure 2a, RBM does not directly exploit the fact that information from across tumor samples is available. After all, cancer genes harbor mutations across many independent tumors. Furthermore, it might be that, in an insertion cluster, a minority of insertions occur that contradict the window sizes set for RBM. As an example, consider a cluster of insertions, a CIS. Suppose that a number of these insertions lie outside the mapping window relative to a certain gene, and the other insertions lie within the mapping window. RBM will not map the insertions outside the mapping window to the gene. However, since the insertions constitute a cluster, it is not unreasonable to assume that all these insertions will all target the same gene. As another example, consider again a cluster of insertions. Let us suppose that a large majority of the insertions have a sense orientation relative to a target gene, and just a few insertions are oriented antisense. Here it will again make sense to map the cluster as a whole to the same target gene, thereby disregarding the antisense orientation of a small minority of insertions.

The implication of these two examples is that it is sensible to allow exceptions to the strict application of the rules, when this is suggested by information regarding the frequency and orientation of insertions across tumors. We therefore propose a hybrid approach, involving RBM and GKC, to exploit information from across tumor samples to flexibly apply RBM in a data-driven manner. Recall that GKC (7) detects CISs by estimating the insertion density through a Gaussian kernel convolution and identifying insertion hot spots based on a random permutation approach. The hybrid approach will be referred to as KC-RBM, and is illustrated in Figure 2b. First, given an insertion profile, a Gaussian kernel convolution is applied to estimate the insertion density, essentially defining clusters of insertions. Second, all insertions are associated with their nearest peak. This results in a number of insertion clusters, one for each peak. Third, if a cluster is orientation-wise homogeneous enough, all individual insertions are merged into a single orientation cluster, otherwise insertions are separated into a sense and an antisense cluster. The positions of the resulting clusters are taken to be the average position of the insertions constituting that cluster. Fourth, all clusters mean loci are mapped using RBM.

In addition to the four window sizes, KC-RBM depends on two parameters: one for determining the level of smoothing of the positions of the insertions, and one for determining the orientation homogeneity of a cluster. The parameter that determines the smoothing is the standard

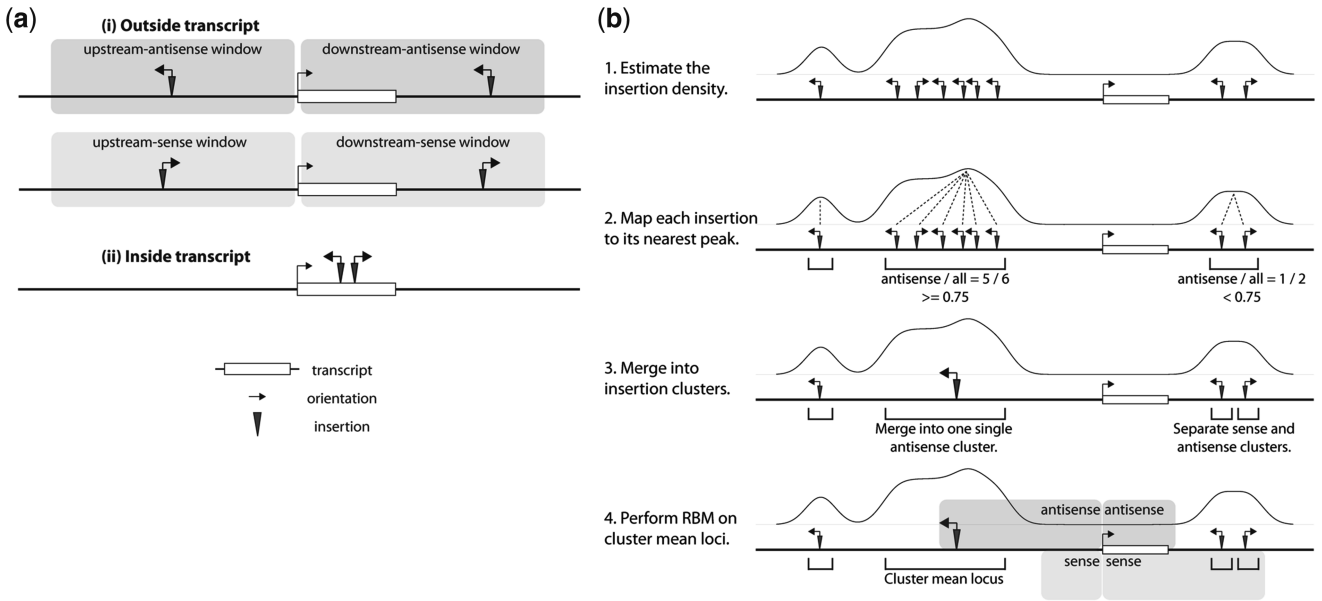

Figure 2. (**a**) RBM for mapping insertions to genes. Distinctions are made based on three properties. Insertions are distinguished by occurrence (i) outside or (ii) inside a transcript, upstream or downstream of a transcript, and in sense or antisense orientation with respect to the orientation of the transcript. (**b**) KC-RBM for mapping insertions to transcripts. First, given an insertion profile, a Gaussian kernel convolution is applied to estimate the insertion density. Second, all insertions are associated with their nearest peak. This results in a number of insertion clusters, one for each peak. Third, if the cluster is orientation-wise homogeneous enough, all individual insertions are merged into a single-orientation cluster, otherwise insertions are separated into a sense and an antisense cluster. Fourth, all clusters mean loci are mapped using RBM. Finally, a gene is considered a target of an insertion if at least one of its transcripts is a target.

deviation of the Gaussian kernel, and is called the scale parameter. The parameter that controls the orientation homogeneity of a cluster is defined as the minimal fraction of the insertions constituting a cluster that need to have the same orientation. The kernel width reflects the degree of strictness with which one wishes to enforce the mapping window: the smaller the scale, the less flexibility is allowed in the chosen sizes of the mapping windows. The orientation homogeneity parameter controls the level of noise tolerated in the insertion orientation: the higher the orientation homogeneity fraction, the higher the stringency on the orientation of insertions.

Varying the window sizes

This section explores the influence of varying the four window size parameters on insertion–expression association, while setting the scale parameter to 0 and the orientation homogeneity parameter to 1.0, i.e. strictly applying the mapping window widths and without smoothing the insertion orientation. When compared to the analysis represented in Figure 1, this analysis is more refined in that for a particular value of a window size parameter, a Wilcoxon test is performed to determine whether the median difference between the expression of samples with and without a given insertion is significantly different from zero (for more detail, please refer to the 'Materials and Methods' section).

For MuLV, Figure 3a shows the influence of varying the window sizes on insertion–expression association, as measured by the fraction of significant associations (true positive rate) and the number of detected significant associations (number of true positives). In Figure 3a(i), cases with at least one insertion per gene across samples were taken into account. The insertions oriented away from genes, upstream-antisense and downstream-sense, show the largest association (large fraction of significant genes). This is in concordance with the literature, where these cases are often denoted as enhancer insertions, and can activate genes across large distances (4,5,13). In contrast, the association of upstream-sense insertions is very local. This is also in concordance with the literature, where these insertions are often denoted as promoter insertions (4,5,13). Downstream-antisense insertions show the least association. However, there is a clear association for small window sizes (<10 kb).

In contrast to Figure 3a(i), in Figure 3a(ii) we only included genes with at least two insertions per gene across all samples. This shows that, in general, insertion occurrence associates even better with elevated expression levels, although for small downstream window sizes (<20 kb) the data are very sparse. In both these figures the association of downstream antisense insertions is less pronounced than that of downstream sense insertions.

For the SB transposon, Figure 3b shows the influence of different window sizes on insertion–expression association. Also in this figure it is evident that the association is far less pronounced than for the retroviral case, although it can be seen that SB transposon insertions are predominantly found inside genes. Requiring at least one insertion per gene across samples gives a noisy result and only shows a slight association for sense insertions. Requiring at least two insertions (Figure 3b(ii)) gives a

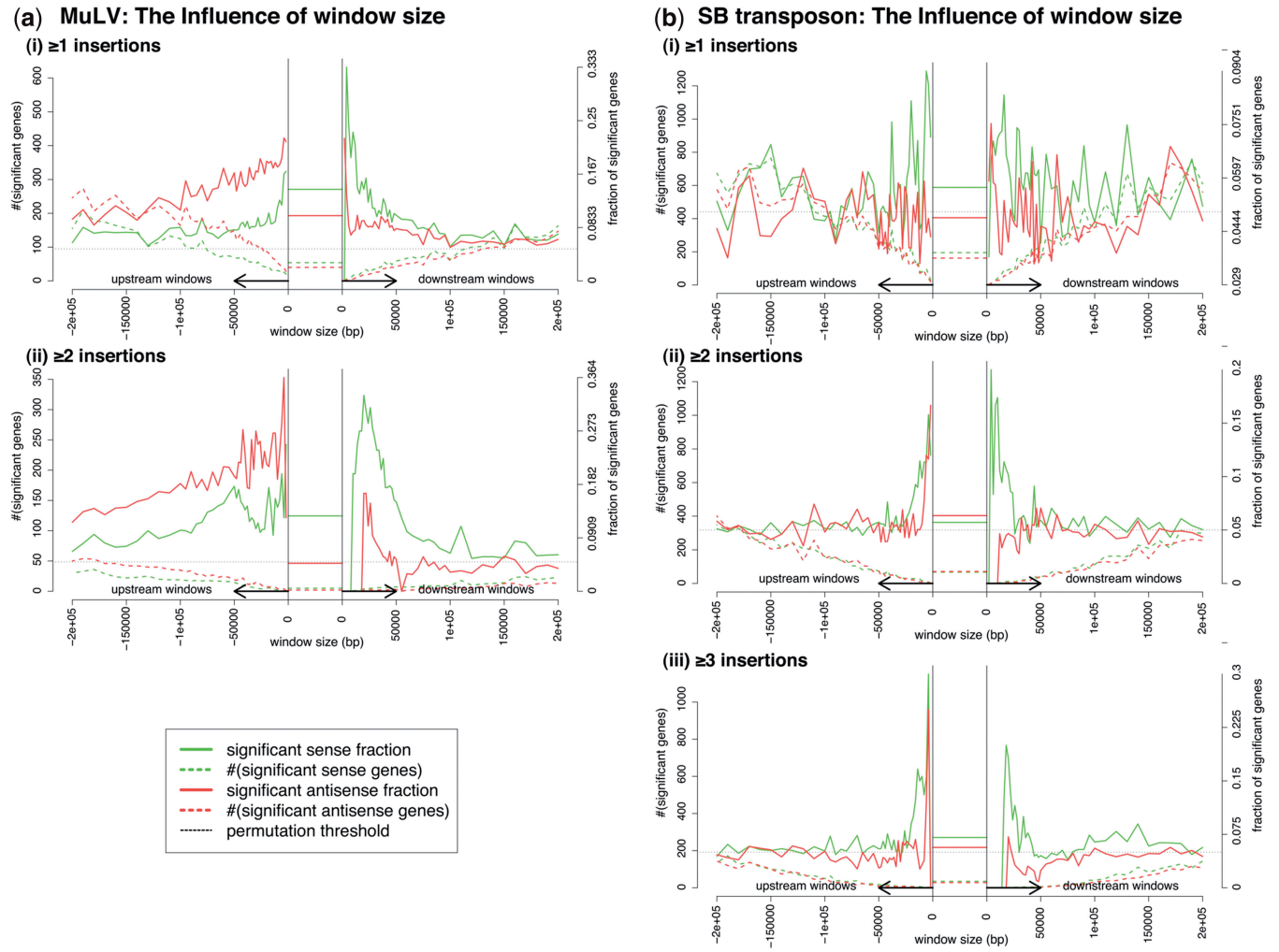

Figure 3. The influence of the four window sizes on mapping quality for (**a**) MuLV and (**b**) the SB transposon, for requiring (**i**) at least one insertion per gene across samples, (**ii**) at least two insertions per gene across samples or (**iii**) at least three insertions per gene across samples. A distinction is made between upstream-sense, upstream-antisense, downstream-sense, and downstream-antisense windows. For an explanation of the computation of the fractions and numbers of significant genes (true positive rate and number of true positives, respectively) refer to the 'Materials and Methods' section. First, computation is done for multiple window sizes. Second, when computing the statistics for one of four window sizes, the other window sizes are set to zero. Third, for all insertions within transcripts association of insertion occurrence with increased expression levels was computed while disregarding the insertions outside transcripts. Permutation thresholds (5%, represented by the dotted black line) were calculated per window size.

clearer picture, and results in higher fractions of significant genes. It shows that the insertion–expression association for the SB transposon is far more localized and less pronounced when compared to retroviral insertions. Again, the sense insertions associate better with increased expression. For both downstream-antisense insertions and downstream-sense insertions, the data are too sparse to draw meaningful conclusions.

One remark should be made on the visual presentation in Figure 3. To allow for a 2D presentation, the four window types and the within-transcript case are treated separately. i.e. when computing the statistic for one specific window size, the other three are set to zero. A more comprehensive view is offered in a more complex 4D visualization in the Supplementary Data (Supplementary Figures S7, S8, S11 and S12), but does not lead to different observations.

Varying the smoothing

This section explores the influence of varying the Gaussian Kernel Convolution scale parameter on insertion–expression association, while keeping the orientation homogeneity parameter and the four window sizes constant. Figure 3 showed there are substantial differences in strength of insertion–expression association for the four window size parameters. Therefore, while varying the smoothing, these window sizes are fixed to values reflecting this relative strength of insertion–expression association. This implies that the upstream-antisense window (ua) is the largest window, followed by the downstream-sense (ds) window, the upstream-sense (us) window and the downstream-antisense (da) window, respectively. Furthermore, window sizes are chosen such that the fraction of significant genes never falls below the permutation threshold of 5%, and a particular window size is

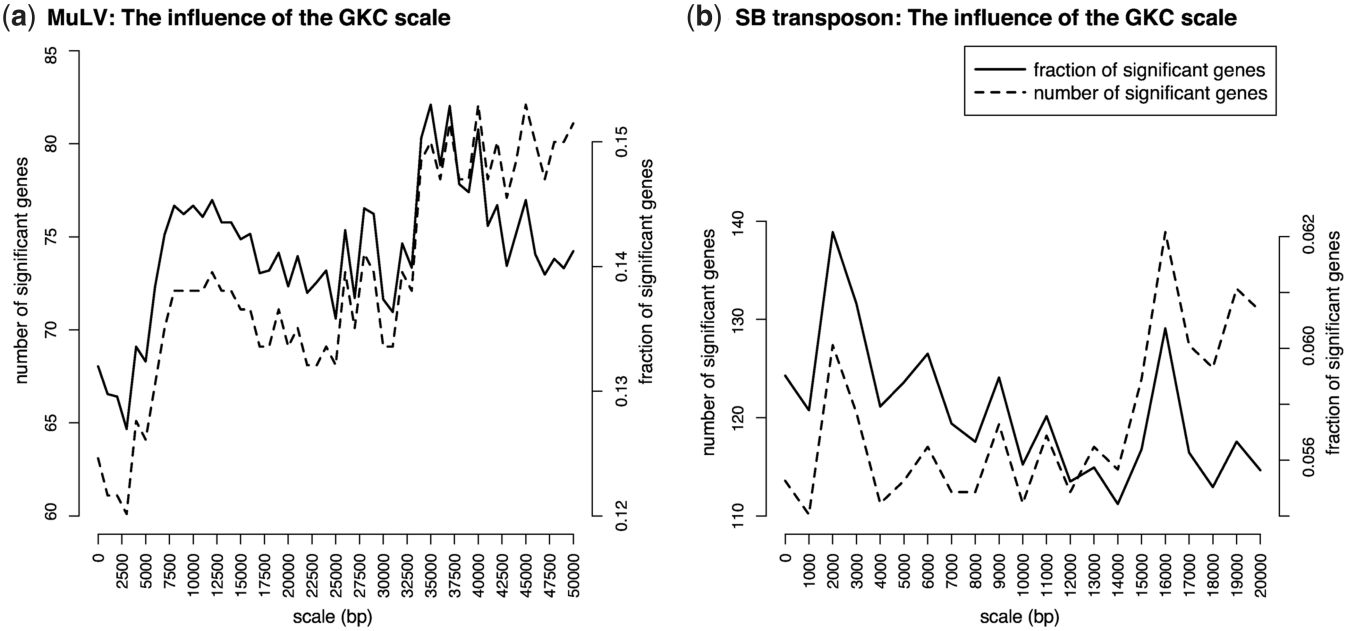

Figure 4. The influence of the GKC scale parameter on the number of true positives and the true positive rate, for performing RBM on (**a**) the MuLV data and (**b**) the SB transposon data using window sizes (20 kb, 120 kb, 40 kb, 5 kb) (MuLV) and (20 kb, 10 kb, 25 kb, 5 kb) (SB transposon) for (upstream-sense, upstream-antisense, downstream-sense, downstream-antisense) insertions. The orientation homogeneity parameter was set to 0.75.

never too small to retrieve at least one significant gene in that window. This resulted in the window sizes (us, ua, ds, da) = (20 kb, 120 kb, 40 kb, 5 kb). Furthermore the orientation homogeneity minimal fraction was set to 0.75.

For these parameter values, the influence of the scale parameter on the number of significant genes and the fraction of significant genes is visualized in Figure 4. For the MuLV data set, KC-RBM achieves stronger associations if smoothing is applied, for all scales larger than 5 kb, and especially for scales around 10 kb and 35 kb (Figure 4a). For scales larger than 35 kb the performance deteriorates, with the number of significant genes increasing and the fraction of significant genes decreasing.

For the SB transposon, Figure 4b again shows that the association of SB transposon insertions with increased gene expression is less pronounced than in the case of MuLV. However, the strongest association is attained for small scales around 2 kb. Furthermore, similar to the retroviral case, the number and fraction of significant genes diverges for larger scales. Hence, we fixed the the kernel width for MuLV at 10 kb and for the SB transposon at 2 kb.

Using KC-RBM for cancer gene identification

In previous sections, we have investigated how the association between insertion presence and gene expression is modulated by the parameters of KC-RBM (window sizes and kernel width for a fixed homogeneity parameter), and fixed these parameters to appropriate values. Now, we will demonstrate how KC-RBM is employed for the identification of cancer genes based on the insertion data only.

Recall that for each insertion, KC-RBM identifies a list of putative target genes. However, for a given insertion, not all identified targets may be of equal importance. As a first step in extracting interesting genes from a KC-RBM mapping, we will select at most a single target for each insertion. More specifically, among the targets identified per insertion, we rank the genes according to the number of times they were targeted across insertions, and select the top ranking gene as the single target for that insertion. This selects for genes frequently targeted across insertions. The set of all targeted genes can subsequently be ranked by counting for each gene the number of times that gene was targeted by an insertion, resulting in a list of commonly targeted genes (CTGs).

Figure 5a shows a top 20 CTG list for the MuLV data set, obtained by applying the procedure described above. Figure 5d shows the results for the SB transposon data set.

Evaluating KC-RBM

In this section, we will evaluate the effectiveness of KC-RBM in identifying cancer genes by following the steps described in the previous section, and comparing the results obtained with two other methods for computationally identifying cancer genes from insertional mutagenesis screens. The first method performs a GKC-based CIS analysis (7), and then maps each CIS peak to the nearest gene. We will refer to this method as CIS-nearest-gene mapping (CIS-NG). The second method consists of mapping each insertion to the nearest gene, and then determining the CTGs. We will refer to this method as NGM. The results obtained by these three methods will be evaluated by comparing them to manually curated lists of cancer genes.

For MuLV, a manually curated list exists, based on a subset of the same MuLV insertion data (3). In Figure 5b

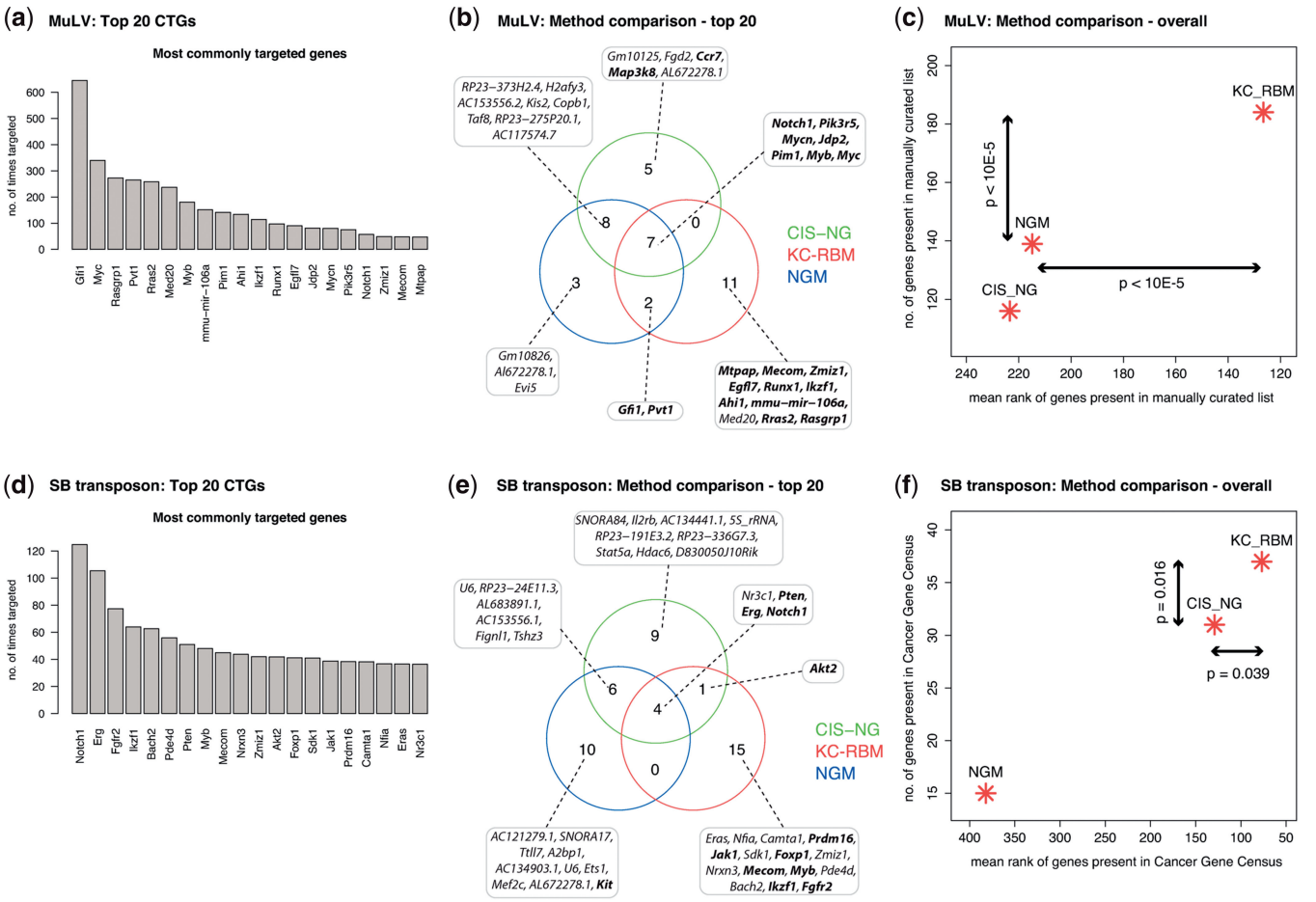

Figure 5. Comparison between genes identified by CTG mapping and by CIS mapping for MuLV and SB transposon. The top 20 CTGs identified by KC-RBM for (**a**) MuLV and (**d**) SB transposon; on the *x*-axis the gene symbol, on the *y*-axis the number of times a certain gene was identified as a target. KC-RBM was performed using window sizes (20 kb, 120 kb, 40 kb, 5 kb) for the MuLV data and (20 kb, 10 kb, 25 kb, 5 kb) for the SB transposon data, for upstream-sense, upstream-antisense, downstream-sense and downstream-antisense windows, respectively. The scale parameters were set to 10 kb (MuLV) and 2 kb (SB transposon). The orientation homogeneity parameter in both cases was set to 0.75. Venn diagrams for both (**b**) MuLV and (**e**) SB transposon depicting the overlap between the top 20 CTGs identified by KC-RBM-CTG mapping, and the top 20 CIS-nearest-genes. Gene names in bold face refer to genes that were also identified in the manually curated list of 346 CISs (3) (MuLV) or in the Cancer Gene Census (SB transposon). For (**c**) MuLV and (**f**) SB transposon, the complete lists of CTGs are also more extensively compared to the manually curated set (MuLV) and the Cancer Gene Census list (B transposon), again taking the manually curated list as a reference. This shows that KC-RBM performs better than both CIS-NG and NGM.

this manually curated list is taken as a reference and compared to the top 20 lists for KC-RBM, NGM and CIS-NG mapping. The seven genes identified in all three mappings were also present in the manually curated list. With the exception of *Med20*, all genes in the KC-RBM lists were identified in the manually curated list as well. The presence of *Med20* is explained by its proximity to *Ccnd3*. In the manually curated list, the insertions in the neighborhood of *Med20* were mapped to *Ccnd3*. CIS-NG finds two targets neither of which were present in the other two mappings nor in the manually curated list. These two genes, *Gm10125* and *Fgd2*, lie near *Zhfx1a* (*Zeb1*) and *Pim1*, respectively, which are present in the manually curated list. NGM finds three targets neither identified in the other two mapping nor in the manually curated list, *Gm10826*, *Al672278.1* and *Evi5*. These genes lie near *Myb*, *Runx1* and *Gfi1*, respectively, which were identified in the manually curated list as the target genes of corresponding CISs.

For MuLV, the lists of CTGs are also more extensively compared in Figure 5c, not restricting the comparison to only the top 20 CTGs, and again taking the manually curated list as a reference. This shows that, of all three methods, KC-RBM identifies the largest number of genes present in the manually curated list. The difference in the number of genes identified by KC-RBM and the second-best method, NGM, is highly significant ($P < 10^{-5}$, based on a permutation approach). Furthermore, the genes identified by KC-RBM that were also identified in the manually curated list, rank higher (lower average rank in the KC-RBM list), when compared to the other methods. The difference in rank between KC-RBM and the second-best method, NGM, is also highly significant ($P < 10^{-5}$, based on a permutation approach)

No manually curated list exists for the SB transposon insertions. Therefore, the Cancer Gene Census (11) was taken as a reference for evaluating the targets identified by the three approaches based on this data set. The results of the comparison of KC-RBM, NGM and CIS-NG mapping are depicted in Figure 5. Four genes are present in all three mappings, *Erg*, *Pten*, *Notch1* and *Nr3c1*. All these genes are known cancer-related genes, see e.g. (15–18), and the first three are also present in the Cancer Gene Census. NGM is the only approach that identifies *Kit* as an additional Cancer Gene Census gene. Both CIS-NG and KC-RBM identify *Akt2* in addition to the Cancer Gene Census jointly detected by all three approaches. However, KC-RBM finds eight additional Cancer Gene Census genes. Similar to the retroviral case, somewhat more obscure proximal targets can score very high in CIS-NG mapping and NGM. Examples are *AC153556.1*, *RP23-336G7.3* and *RP23-24E11.3*, which are located in the vicinity of *Myb*, *Jak1* and *Bach2*, respectively. *Myb*, *Jak1* and *Bach2* are identified exclusively by KC-RBM and have been shown to be involved in cancer, see e.g. (19–21).

The results in Figure 5e are further substantiated by the more extensive comparison in Figure 5c, comparing larger lists of genes (refer to the 'Materials and Methods' section), and again taking the Cancer Gene Census as a reference. KC-RBM again performs best in terms of the average rank of the Cancer Gene Census genes it identifies as well as the number of Cancer Gene Census genes identified. Furthermore, the differences in presence and rank between KC-RBM and the second-best method (CIS-NG in this case), are again significant ($P = 0.016$ and $P = 0.039$, respectively). Note that the Cancer Gene Census is less relevant for the SB insertion data than the manually curated list is for the MuLV data. It is not based on a SB transposon insertional mutagenesis screen. Furthermore, it is a list of human genes, which additionally have to be mapped to mouse homologs. Consequently, the overlap between the genes identified by the three methods and the Cancer Gene Census genes is much smaller compared to the MuLV case, resulting in higher *P*-values. Nevertheless, the *P*-values are significant, and demonstrate clearly that KC-RBM retrieves more cancer-related genes than the other methods.

DISCUSSION

In this article, we presented KC-RBM, a method for automatically mapping individual retroviral and SB transposon insertions to putative target genes. KC-RBM represents the first ever approach for mapping SB transposon insertions to target genes. In addition, the analyses presented here constitute the first genome-wide analysis of insertion and same-sample gene expression for retroviruses and transposons. Such a comprehensive data set provides significant power in determining the factors that govern the associations between insertions and neighboring genes.

It is important to emphasize that, while mapping individual insertions to target genes, KC-RBM also exploits cross-sample information in multiple ways. The KC-RBM scale parameter allows for smoothing the positions of insertions based on cross-sample information. The insertion orientations are smoothed based on cross-sample information by setting the orientation homogeneity fraction parameter in KC-RBM. Furthermore, determining CTGs by aggregating insertions also integrates information from across tumor samples.

Although the presence of insertions is associated with increased gene expression (Figure 1) the analysis presented in Figure 1 does not provide sufficient evidence to distinguish cause from effect. In other words, we cannot conclude from this associative analysis that the presence of an insertion causes higher or lower gene expression. It is certainly possible that insertions are more likely to occur in actively transcribed regions, i.e. near genes with elevated expression levels. This and other factors inducing insertion biases have, in fact, been demonstrated for retroviruses as well as for transposons, e.g. (14,22–28). However, regardless of the determinants of insertion bias, it is a fact that insertions cause tumors since there is a statistically significant increase in the tumor incidence in animals infected with MuLV or in animals where SB transposons are activated (3,29). It is therefore very likely that these insertions caused, among other, changes in gene expression that resulted in oncogenesis. To further explore the causal relationship between insertions and aberrant gene expression, we performed additional analyses, the results of which are presented in the Supplementary Data.

The first analysis involves the distribution of insertion orientation in CTGs. If insertions were simply occurring in regions that were already transcriptionally active, i.e. elevated expression being the cause of insertions, one would expect insertions in these regions to show no preference for orientation. On the other hand, if the insertions were the cause of elevated gene expression, one would expect a strong preference for an orientation consistent with activation of the target gene(s). To test this hypothesis, we performed two tests. First, we randomized the insertion orientation in both the MuLV and SB data sets and regenerated the expression–position–orientation plots as presented in Figure 1. As can be seen in Supplementary Figures S2 and S10, the orientation-dependent effect on gene expression disappears. Second, we analyzed the orientation distribution of insertions assigned to the top 214 MuLV CTGs (only including CTGs with 10 or more insertions). The results show a highly significant preference for orientation, consistent with the known mechanisms employed by retroviruses to activate target genes (see Supplementary Figures S3 and S4). Strong preferences for activating orientations were also established for the SB transposon (Supplementary Figure S13). To further explore the causal relationship between insertions and gene expression, we performed a second analysis where we compared expression data from normal and tumor tissue. If insertions would simply target genes that are already active in normal tissue, there will be little difference between the activity of genes carrying insertions in tumor tissue and the activity of these same genes in normal tissue. To test this hypothesis, we compared gene

expression of MuLV CTGs in cancer tissue to gene expression of these same genes in normal tissue. For the cancer tissue, only tumors with insertions in a specific CTG were included in the analysis. These analyses showed that in the cases where the average expression was higher in the cancer tissue compared to normal tissue, the increase was significant in 84.2% of the cases. This implies that in the vast majority of cases a gene carrying an insertion in the tumor tissue was significantly differentially expressed with respect to that same gene in normal tissue. For the cases where the average expression was lower in the tumor tissue with respect to normal tissue, the decrease was significant in 72.7% of the cases (see Supplementary Figure S5). This strongly suggests that frequently targeted genes showing aberrant expression in tumor tissue, are not already active in normal tissue, hence challenging the aforementioned hypothesis that insertions simply tend to insert in genes that are already active in normal tissue.

Taken together, the strong orientation preference and the strong association of insertion presence with a change in gene expression suggest that a large fraction of insertions play a causal role in aberrant gene expression in tumor samples. Lastly and perhaps most importantly, it should be emphasized that while we employed the association of insertions and gene expression to obtain settings of the four window sizes in KC-RBM, this is not a requirement for the approach. In fact, depending on a researcher's interests, different window sizes may be appropriate.

As the definitive test for the performance of KC-RBM, and regardless of the direction of causality, we evaluated its ability to identify known cancer-associated genes from a manually curated list (3), and from the Cancer Gene Census (11). KC-RBM gives superior results when compared to automated CIS-NG mapping and NGM. Without the need for human intervention, it avoids more obscure proximal targets and finds a clean list of well-known cancer-related genes, as demonstrated by the comparison with a manually curated list (3), and the Cancer Gene Census (11). This is important, since human interference could cause a bias, for example toward known or expected cancer genes, thus actually preventing the discovery of new or unknown cancer genes. This emphasizes the added value that a reliable automated insertion mapping procedure such as KC-RBM can have for analyzing insertional mutagenesis data and discovering novel oncogenes and tumor suppressor genes. As such, we believe that KC-RBM will significantly increase the efficiency of cancer gene discovery from insertional mutagenesis screens.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

This work was financed by the Netherlands Consortium for Systems Biology (NCSB) which is part of the Netherlands Genomics Initiative/Netherlands Organisation for Scientific Research. David J. Adams is supported by Cancer Research UK and the Wellcome Trust. Funding for open access charge: The Netherlands Consortium for Systems.

Conflict of interest statement. None declared.

REFERENCES

1. Mikkers,H., Allen,J., Knipscheer,P., Romeijn,L., Hart,A., Vink,E., Berns,A. and Romeyn,L. (2002) High-throughput retroviral tagging to identify components of specific signaling pathways in cancer. *Nat. Genet.*, **32**, 153–159.
2. Lund,A., Turner,G., Trubetskoy,A., Verhoeven,E., Wientjens,E., Hulsman,D., Russell,R., DePinho,R., Lenz,J. and van Lohuizen,M. (2002) Genome-wide retroviral insertional tagging of genes involved in cancer in Cdkn2a-deficient mice. *Nat. Genet.*, **32**, 160–165.
3. Uren,A., Kool,J., Matentzoglu,K., deRidder,J., Mattison,J., van Uitert,M., Lagcher,W., Sie,D., Tanger,E., Cox,T. *et al.* (2008) Large-scale mutagenesis in p19(ARF)- and p53-deficient mice identifies cancer genes and their collaborative networks. *Cell*, **133**, 727–741.
4. Uren,A., Kool,J., Berns,A. and van Lohuizen,M. (2005) Retroviral insertional mutagenesis: past, present and future. *Oncogene*, **24**, 7656–7672.
5. Kool,J. and Berns,A. (2009) High-throughput insertional mutagenesis screens in mice to identify oncogenic networks. *Nat. Rev. Cancer.*, **9**, 389–399.
6. Suzuki,T., Shen,H., Akagi,K., Morse,H.C., Malley,J., Naiman,D., Jenkins,N. and Copeland,N. (2002) New genes involved in cancer identified by retroviral tagging. *Nat. Genet.*, **32**, 166–174.
7. deRidder,J., Uren,A., Kool,J., Reinders,M. and Wessels,L. (2006) Detecting statistically significant common insertion sites in retroviral insertional mutagenesis screens. *PLoS Comput. Biol.*, **2**, 1530–1542.
8. Sauvageau,M., Miller,M., Lemieux,S., Lessard,J., Hébert,J. and Sauvageau,G. (2008) Quantitative expression profiling guided by common retroviral insertion sites reveals novel and cell type specific cancer genes in leukemia. *Blood*, **111**, 790–799.
9. Mattison,J., Kool,J., Uren,A., deRidder,J., Wessels,L., Jonkers,J., Bignell,G., Butler,A., Rust,A., Brosch,M. *et al.* (2010) Novel candidate cancer genes identified by a large-scale cross-species comparative oncogenomics approach. *Cancer Res.*, **70**, 883–895.
10. Erkeland,S., Verhaak,R., Valk,P., Delwel,R., Lowenberg,B. and Touw,I. (2006) Significance of murine retroviral mutagenesis for identification of disease genes in human acute myeloid leukemia. *Cancer Res.*, **66**, 622–626.
11. Futreal,P., Coin,L., Marshall,M., Down,T., Hubbard,T., Wooster,R., Rahman,N. and Stratton,M. (2004) A census of human cancer genes. *Nat. Rev. Cancer*, **4**, 177–183.
12. Durinck,S., Moreau,Y., Kasprzyk,A., Davis,S., De Moor,B., Brazma,A. and Huber,W. (2005) BioMart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, **21**, 3439–3440.
13. Jonkers,J. and Berns,A. (1996) Retroviral insertional mutagenesis as a strategy to identify cancer genes. *Biochim. Biophys. Acta.*, **1287**, 29–57.
14. Wu,X., Li,Y., Crise,B. and Burgess,S. (2003) Transcription start regions in the human genome are favored targets for MLV integration. *Science*, **300**, 1749–1751.
15. Sashida,G., Bazzoli,E., Menendez,S., Liu,Y. and Nimer,S. (2010) The oncogenic role of the ETS transcription factors MEF and ERG. *Cell Cycle*, **9**, 3457–3459.
16. Salmena,L., Carracedo,A. and Pandolfi,P. (2008) Tenets of PTEN tumor suppression. *Cell*, **133**, 403–414.
17. Allenspach,E., Maillard,I., Aster,J. and Pear,W. (2002) Notch signaling in cancer. *Cancer Biol. Ther.*, **1**, 466–476.
18. Lind,G., Kleivi,K., Meling,G., Teixeira,M., Thiis-Evensen,E., Rognum,T. and Lothe,R. (2006) ADAMTS1, CRABP1, and NR3C1 identified as epigenetically deregulated genes in colorectal tumorigenesis. *Cell Oncol.*, **28**, 259–272.

19. Ramsay,R. and Gonda,T. (2008) MYB function in normal and cancer cells. *Nat. Rev. Cancer*, **8**, 523–534.
20. Ono,A., Kono,K., Ikebe,D., Muto,A., Sun,J., Kobayashi,M., Ueda,K., Melo,J., Igarashi,K. and Tashiro,S. (2007) Nuclear positioning of the BACH2 gene in BCR-ABL positive leukemic cells. *Gene. Chromosome. Canc.*, **46**, 67–74.
21. Verma,A., Kambhampati,S., Parmar,S. and Platanias,L. (2003) Jak family of kinases in cancer. *Cancer Metast. Rev.*, **22**, 423–434.
22. Berry,C., Hannenhalli,S., Leipzig,J. and Bushman,F. (2006) Selection of target sites for mobile DNA integration in the human genome. *PLoS Comput. Biol.*, **2**, e157.
23. Hematti,P., Hong,B., Ferguson,C., Adler,R., Hanawa,H., Sellers,S., Holt,I., Eckfeldt,C., Sharma,Y., Schmidt,M. *et al.* (2004) Distinct genomic integration of MLV and SIV vectors in primate hematopoietic stem and progenitor cells. *PLoS Biol.*, **2**, 2183–2190.
24. Mitchell,R., Beitzel,B., Schroder,A., Shinn,P., Chen,H., Berry,C., Ecker,J. and Bushman,F. (2004) Retroviral DNA integration: ASLV, HIV and MLV show distinct target site preferences. *PLoS Biol.*, **2**, e234.
25. Bushman,F., Lewinski,M., Ciuffi,A., Barr,S., Leipzig,J., Hannenhalli,S. and Hoffmann,C. (2005) Genome-wide analysis of retroviral DNA integration. *Nat. Rev. Microbiol.*, **3**, 848–58.
26. Lewinski,M., Yamashita,M., Emerman,M., Ciuffi,A., Marshall,H., Crawford,G., Collins,F., Shinn,P., Leipzig,J., Hannenhalli,S. *et al.* (2006) Retroviral DNA integration: viral and cellular determinants of target-site selection. *PLoS Pathog.*, **2**.
27. Ambrosi,A., Cattoglio,C. and Di Serio,C. (2008) Retroviral integration process in the human genome: Is it really non-random? A new statistical approach. *PLoS Comput. Biol.*, **4**, e1000144.
28. Plachy,J., Kotáb,J., Divina,P., Reinisova,M., Senigl,F. and Hejnar,J. (2010) Proviruses selected for high and stable expression of transduced genes accumulate in broadly transcribed genome areas. *J. Virol.*, **84**, 4204–4211.
29. Starr,T., Allaei,R., Silverstein,K., Staggs,R., Sarver,A., Bergemann,T., Gupta,M., Gerard O'Sullivan,M., Matise,I., Dupuy,A. *et al.* (2009) A Transposon-Based genetic screen in mice identifies genes altered in colorectal cancer. *Science*, **323**, 1747–1750.