BIOINFORMATICS

Vol. 23 ISMB/ECCB 2007, pages i133–i141
doi:10.1093/bioinformatics/btm202

Co-occurrence analysis of insertional mutagenesis data reveals cooperating oncogenes

Jeroen de Ridder[1,3], Jaap Kool[2], Anthony Uren[2], Jan Bot[1], Lodewyk Wessels[1,3] and Marcel Reinders[1,*]

[1]Information and Communication Theory Group, Faculty of EEMCS, Delft University of Technology, Delft, The Netherlands, [2]Division of Molecular Genetics and [3]Division of Molecular Biology, The Netherlands Cancer Institute, Amsterdam, The Netherlands

ABSTRACT

Motivation: Cancers are caused by an accumulation of multiple independent mutations that collectively deregulate cellular pathways, e.g. such as those regulating cell division and cell-death. The publicly available Retroviral Tagged Cancer Gene Database (RTCGD) contains the data of many insertional mutagenesis screens, in which the virally induced mutations result in tumor formation in mice. The insertion loci therefore indicate the location of putative cancer genes. Additionally, the presence of multiple independent insertions within one tumor hints towards a cooperation between the insertionally mutated genes. In this study we focus on the detection of statistically significant co-mutations.

Results: We propose a two-dimensional Gaussian Kernel Convolution method (2DGKC), a computational technique that identifies the cooperating mutations in insertional mutagenesis data. We define the Common Co-occurrence of Insertions (CCI), signifying the co-mutations that are statistically significant across all different screens in the RTCGD. Significance estimates are made on multiple scales, and the results visualized in a scale space, thereby providing valuable extra information on the putative cooperation.

The multidimensional analysis of the insertion data results in the discovery of 86 statistically significant co-mutations, indicating the presence of cooperating oncogenes that play a role in tumor development. Since oncogenes may cooperate with several members of a parallel pathway, we combined the co-occurrence data with gene family information to find significant cooperations between oncogenes and families of genes. We show, for instance, the interchangeable cooperation of *Myc* insertions with insertions in the *Pim* family.

Availability: A list of the resulting CCIs is available at: http://ict.ewi.tudelft.nl/~jeroen/CCI/CCI_list.txt

Contact: m.j.t.reinders@tudelft.nl

1 INTRODUCTION

Cancers arise when the regulatory pathways that govern healthy cell proliferation (cell division) are disrupted. Moreover, one of the hallmarks of cancer is that multiple oncogenic events, disrupting multiple pathways, are required before the state of uncontrolled proliferation is reached (Hanahan and Weinberg, 2000). For instance, (mutational) activation of the *Myc* protooncogene together with the loss of the *p53* tumor-suppressor gene in mice, is a commonly observed co-occurrence of mutations that can cause cancer. In this respect, these two genes can be considered to 'cooperate' in the development of the tumor.

In retroviral insertional mutagenesis experiments, genes involved in the development of cancer are identified by determining the loci of viral insertions from tumors induced by retroviruses in cancer-predisposed mice (reviewed in Mikkers and Berns, 2003; Uren *et al.*, 2005). In van Lohuizen *et al.* (1991), for example, the cancer-predisposition is acquired by inserting an Eμ*Myc* transgene in the mouse DNA. After infecting a host cell, the retrovirus inserts its own DNA into the host cell's genome, mutating the host cell's DNA in the process. The mutation may cause alteration in expression of genes in the vicinity of the insertion or, when inserted within a gene, alteration of the gene product. When the affected gene is a cancer gene, activation of a proto-oncogene or inactivation of a tumor-suppressor gene can, in cooperation with the cancer predisposition, cause uncontrolled proliferation of cells. Eventually this may give rise to tumors. Throughout this text these cancer-causing insertions are referred to as oncogenic insertions.

The tumor tissue contains many copies of the cell bearing the oncogenic insertions, but only a few copies of cells carrying non-oncogenic (random, background) insertions. Consequently, cloning the flanking sequences of the inserted virus to determine the insertion loci, will result in a data set of insertion loci (the oncogenic insertions) that are indicative for the presence of nearby cancer genes contaminated with noise (the non-oncogenic insertions). This is schematically depicted in Figures 1A and B. The challenge is to find the regions in the genome that carry insertions in multiple independent tumors significantly more frequently than expected by chance. Such a region is called a Common Integration Site (CIS), and its location is highly correlated with the location of genes involved in tumor development. An important factor to consider is that viral insertions can disrupt gene functioning from various distances around or within the gene. It is therefore essential that significance estimates are made for a range of different CIS widths in order not to miss interesting loci. The discovery of CISs in insertion data will be referred to as a 1D analysis, for which recently a kernel convolution method has been developed (de Ridder *et al.*, 2006).

*To whom correspondence should be addressed.

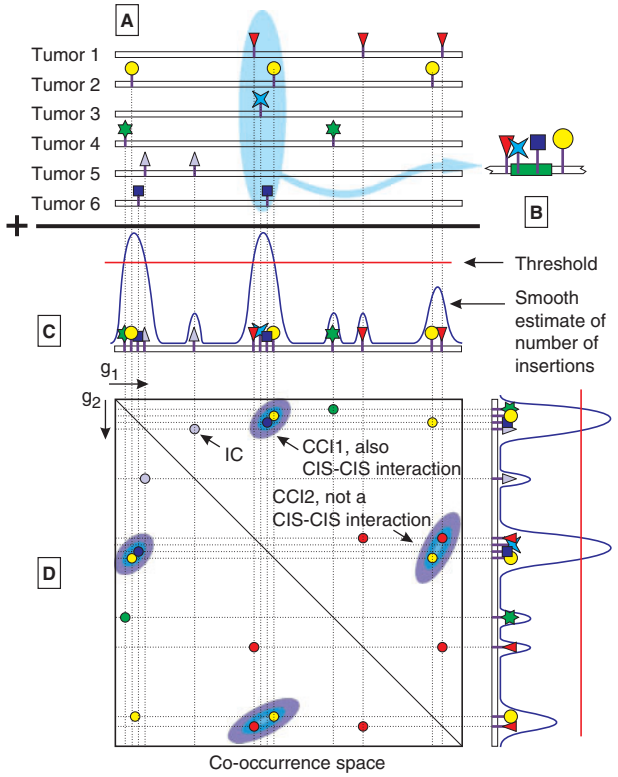

Fig. 1. Schematic depiction of insertion data and mapping to the co-occurrence space. (**A**) Schematic depiction of the data of six tumors. The geometric symbols represent the insertions and are given a different shape for each tumor. The blue region indicates a potential CIS, a region with significantly more insertions than expected by chance. (**B**) An enlargement of the potential CIS. Genes (indicated by the green bar) may be affected from various loci around or within the gene, and there is no unique distance across which viral inserts act on their targets. (**C**) The result of applying a 1D analysis to the aggregate of all the insertions. The blue line represents the 1D estimation of the number of insertions, with peaks indicating high insertion density and therefore putative CISs. The red line is a significance threshold obtained from a permutation analysis. The peaks exceeding this threshold qualify as CISs. (**D**) The mapping of the tumors to the co-occurrence space. Every combination of insertions from one tumor is mapped to a single point in the co-occurrence space, and is referred to as an IC. All co-occurrences are recorded twice, since the co-occurrence space is symmetric in the diagonal. The blue ellipses represent regions with a significantly higher density of co-occurrences, denoted as common co-occurrences of insertions (CCIs). As in the 1D case, significance is determined based on a significance threshold obtained from an empirically generated null-distribution. Note that CCI 1 consists of insertions that also contributed to CISs in both the g_1 and g_2 direction. CCI 2, on the other hand, contains insertions that are only part of a CIS in one direction, the g_2 direction. If a co-occurrence analysis is performed only on insertions that are part of CISs, CCI 1 will be found. For this reason, CCI 1 is a CIS–CIS interaction, since, within one tumor, two distinct CISs are inserted by viruses. However, CCI 2 will not be found, from which it follows that this approach is prone to false negatives. This can be explained by the fact that events in the two-dimensional space are more rare, and hence the threshold for statistical significance can be lower (while still controlling the average number of false positives at the desired α-level), thereby gaining extra power. For this reason the 1D analysis will not be considered any further for the discovery of cooperating genes.

Instead of revealing cooperation of insertionally targeted genes with the cancer-predisposition, this study focuses on revealing the cooperation *between* virally targeted genes (Nakamura *et al.*, 1996; Kim *et al.*, 2003). Ideally, for this purpose the insertions co-occurring in tumors from mice of a uniform genotype should be examined, but a data set that is large enough to acquire statistically significant results is currently absent. Therefore we focus on the co-mutations that are common across a number of different insertional mutagenesis screens from publicly available data. The genes that are targeted by the commonly co-occurring insertions in these tumors are likely to cooperate in the tumor development.

To find the cooperation between virally targeted genes, we propose to analyze the insertion data in the two dimensional co-occurrence space. We define an Insertion Co-occurrence (IC) as a unique combination of insertions within one tumor, and the Common Co-occurrence of Insertions (CCI) as observing the combination of two insertions significantly more frequently than expected by chance across multiple tumors (schematically depicted in Figure 1D). When compared to a 1D analysis, performing a 2D analysis on the insertion data will result in the discovery of new loci that play a role in tumorigenesis. This can be seen by considering a region that is not hit frequently enough to be labeled a CIS in the 1D analysis, but may still be called significant in the 2D analysis, because it co-occurs frequently enough with another inserted region. To ensure all different configurations of insertions around or within genes are taken into account, we evaluate the significance of the CCIs at various scales. Visualizing the CCIs at multiple widths will contribute essential additional information about how insertions disrupt the functioning of their target genes.

Another hallmark of tumorigenesis is the existence of many parallel pathways (Hanahan and Weinberg, 2000), and consequently, the many possibilities of reaching the state of uncontrolled proliferation. This is exemplified by a study using *Pim1* deficient and *Pim2* deficient mice. *Pim1* is frequently hit in screens of Eμ*Myc* transgenic mice. When *Pim1* is knocked out, *Pim2* is frequently hit (van der Lugt *et al.*, 1995), and when *Pim1* and *Pim2* are knocked out, *Pim3* is hit (Mikkers *et al.*, 2002), suggesting all three *Pim* genes promote tumors in cooperation with *Myc*. As a consequence, co-occurring mutations in the RTCGD may not occur frequently enough to be statistically significant, simply because there exist too many parallel possibilities for the cell to become malignant. In this study, we investigate this phenomenon by including gene family information, and assess whether there exists cooperation between genes and a certain gene family.

The data in the RTCGD are publicly available, and the screens in the database have been individually studied and published before. It is therefore likely that the most prominent CCIs will point to cooperations between genes that have been discovered before. However, since we are the first to analyze the *combined* set of screens in the RTCGD for the presence of statistically significant cooperations between virally targeted genes in a systematic fashion, we do expect to discover new interactions. As we expect a subset of our CCIs to be published,

we can partially validate our method by showing that the pairs of genes predicted to cooperate by our method will co-occur in literature abstracts significantly more frequently than expected by chance.

2 METHODS

2.1 The data

Over the last few years an extensive amount of insertional mutagenesis data has been published (see e.g. Hansen *et al.*, 2000; Hwang *et al.*, 2002; Johansson *et al.*, 2004; Joosten *et al.*, 2002; Li *et al.*, 1999; Lund *et al.*, 2002; Mikkers *et al.*, 2002; Suzuki *et al.*, 2002). These data have been compiled in the Retroviral Tagged Cancer Gene Database (RTCGD) (Akagi *et al.*, 2004) (URL: http://RTCGD.ncifcrf.gov, accessed January 4, 2007). Currently, the RTCGD contains 5473 retroviral insertions distributed over 1361 tumors. There are 1031 tumors that contain more than one insertion. The vast majority of the insertions have been acquired in twenty different screens, that used various experimental setups. Therefore, the number of insertions that are found in a tumor varies significantly per screen. Additionally, the mouse models used varied among screens. In this study we analyze the combined data from all the screens in the RTCGD, irrespective of the genetic background or cancer predisposition of the mice used in the screens. Also, we assume that background insertions are distributed uniformly across the genome, and all insertions are independent of each other.

2.2 Insertion Co-occurrence

To exploit the information contained in the joint occurrence of insertions within one tumor, we map the data to the co-occurrence space. In this space a point indicates the location of an IC, that is, two insertions co-occurring in one tumor. Finding the regions in the co-occurrence space that contain ICs more frequently than expected by chance will point to the genes in the genome that cooperate in the development of the tumor.

We propose to apply a 2D Gaussian Kernel Convolution (2DGKC) to determine the statistical significance of the regions with multiple ICs. The 2DGKC, which is very similar to Parzen density estimation, results in a smooth estimate for the number of ICs, $\hat{x}(\mathbf{g})$, at a position $\mathbf{g} \in \{0, G\}$ in the co-occurrence space:

$$\hat{x}(\mathbf{g}) = \sum_{n=0}^{N} K([\mathbf{g} - \mathbf{d}_n]_1) K([\mathbf{g} - \mathbf{d}_n]_2) \quad \text{with } \{0 < g_1 < G, 0 < g_2 < G\}, \tag{1}$$

where G is the total genome length, $K(\cdot)$ is a univariate kernel function, $\mathbf{d}_n$ is the position of the n-th IC, and $[\cdot]_i$ denotes the selection of the i-th element from the vector between brackets. By using the product of two univariate kernel functions local independence is assumed, but by summing multiple kernel functions complex correlation structures can still be discovered. In this study a Gaussian kernel function is used, given by: $K(z) = e^{-2z^2/h^2}$, where h is the kernel width. Note that the kernel function used in our study is not normalized, as is done in traditional density estimates (Parzen, 1962). As a result, the modified density estimate can be interpreted as a continuous estimation of the number of co-occurrences at a given position. The local maxima in $\hat{x}$ (the peaks) will now indicate the location of putative CCIs. Since we are only interested in the local maxima, we reduce the number of evaluations of Equation (1) (required to find the maxima), by applying a standard non-linear optimization algorithm (`fminunc`, MATLAB Optimization toolbox) started from every IC in the data.

2.3 Significance estimates

Significance of the putative CCIs is evaluated by testing against the following null-hypothesis:

$$H_0^{\text{2DGKC}} : \mu_0 = \mu_{\text{observed}}(\mathbf{g})$$

where μ_0 is the mean height of the peaks under the null-hypothesis and $\mu_{\text{observed}}(\mathbf{g}) = \hat{x}(\mathbf{g})$ is the observed height of the peak at position $\mathbf{g}$. The null-hypothesis is rejected if the observed height of the peak significantly *exceeds* the mean height of the peaks under the null-hypothesis.

The null-distribution is acquired by a permutation approach, schematically depicted in Figure 2. The kernel convolution is applied to the ICs that result from a random permutation of the insertions (Fig. 2A and B). This results in random peaks in the co-occurrence space. This is repeated K times, to obtain a set of random realizations (Fig. 2C). From this set, the height of all the peaks is collected, and the null-distribution is computed (Fig. 2D). Using the null-distribution we can convert the α-level to a threshold for the real data. This threshold can now be applied to the smoothed estimate of the number of ICs, that was obtained by applying the 2DGKC to the real co-occurrence data (Fig. 2E). We correct for multiple testing using the Bonferroni multiple testing correction, by dividing the α-level by the number of tests. Since we only evaluate the height of the peaks, we take the number of tests to be equal to the number of peaks in the co-occurrence density.

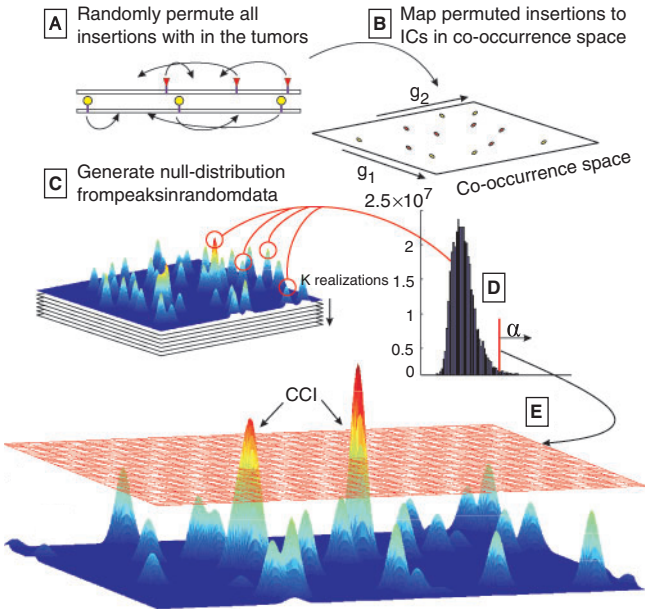

Fig. 2. Schematic depiction of the significance analysis of the smoothed estimate of number of co-occurrences in the insertion data. (**A**) Within each tumor, the position of the insertions are permuted. (**B**) The permuted set of insertions is mapped to the co-occurrence space and the 2D Gaussian Kernel Convolution (2DGKC) is applied. This is repeated to obtain a set of K realization of the density estimate on random data. (**C**) From these realizations the peak heights are collected, and a null-distribution is computed. (**D**) Using a predefined α-level the significance threshold on real data is computed. (**E**) Applying this threshold to the estimated number of Insertion Co-occurrences (ICs) in the real data results in the Common Co-occurrences of Insertions (CCIs), statistically significant co-occurrences of insertions.

2.4 Scale space

The kernel width h can be considered as a scale parameter, thereby providing an excellent way of controlling at which scale the significance of the ICs are evaluated. By increasing h, the kernel functions cover a larger region, and, since potentially more kernel functions will contribute to the smoothed estimate of the number of ICs, this results in higher peaks in this estimate. This mechanism will ensure that the CCIs for which the ICs are confined to one or more very specific regions (narrow CCIs), will only become significant for small values of h (small scales), and conversely, the broad CCIs will only be present at larger scales. This motivates the definition of a cross scale CCI (csCCI), defined as the detection of a CCI at one or more scales.

Visualizing these phenomena will aid the biologist in determining the targeted genes. For this purpose we construct three-dimensional scale space diagrams (see e.g. Figs 5 and 6). In these diagrams the contour, defined by the intersection of the threshold with the smoothed estimate of the number of ICs (Fig. 2E), is plotted in the (g_1/g_2)-plane, as a function of the scale parameter (z-axis). The scale parameter is chosen to cover a range of biologically relevant scales ($10\text{k} \leq h \leq 500\text{k}$). Since for every scale the - computationally intensive - permutation procedure has to be performed, the threshold value is computed only for eight log-uniformly spaced scales. For the 100 intermediate scales, that are used to build the scale space diagrams, the necessary threshold values are computed using a piecewise linear interpolation of the threshold values that were computed using the actual permutation procedure.

2.5 χ^2-ranking

In addition to ranking the csCCIs on their average peak height across the scales, it is also interesting to rank the csCCIs according to a one-tailed χ^2-test, which corrects for the frequency with which the individual co-occurring loci are hit. Using the P-value from the χ^2-test, it is possible to filter the csCCIs at a user-defined α-level, which is an often employed pruning technique in the context of association rule mining (Liu *et al.*, 2001). Note that, by filtering the results, statistically significant interactions (based on peak height) are lost, and should therefore only be employed in case too many interactions were discovered.

Per CCI and per scale a P-value is computed for the χ^2-test performed on the following table:

	A_{g_1}	$\neg A_{g_1}$	
A_{g_2}	N_{g_1,g_2}		N_{g_2}
$\neg A_{g_2}$			
	N_{g_1}		N

In this table, A_{g_1} denotes an area in the co-occurrence space: $A_{g_1} = \{CCI_{g_1} - h < g_1 < CCI_{g_1} + h; 0 < g_2 < G\}$, that is, an area of width $2h$ around CCI_{g_1}, the g_1 position of the CCI under investigation, and the height spanning the complete g_2 axis. A_{g_2} is defined in an analogous fashion. Now, N_{g_1,g_2} can be defined as the number of ICs in the intersection of the areas A_{g_1} and A_{g_2}. Likewise, N_{g_1}, N_{g_2} and N are defined as the total number of ICs in the areas A_{g_1}, A_{g_2} and the complete co-occurrence space, respectively. The csCCIs can now be ranked according to their average P-value across the scales in which the CCI was found to be significant.

2.6 Family mapping

The presence of parallel pathways may prevent co-occurring insertions from reaching the significance threshold. A clear example is the previously mentioned cooperation of the *Myc* proto-oncogene and the *Pim1* and *Pim2* proto-oncogenes. Since more than one possibility exists to cooperate with *Myc*, the spatial correlation in the g_2 direction of the ICs in the *Myc* locus will be diminished, that is, the ICs will be divided into two separate clusters: one near the *Pim1*/*Myc* locus on Chromosome 17/Chromosome 15 and one near the *Pim2*/*Myc* locus on Chromosome X/Chromosome 15. This results in lower peaks at these positions, and, because the data is far from saturated, possibly even causes one or both of these peaks to fail the significance test.

This problem is circumvented by increasing spatial correlation of the regions surrounding the genes that can substitute for each other. There is, however, no data source available that contains information on functional substitution. For this reason, we revert to Ensembl gene family information, which is based on sequence similarity (Hubbard *et al.*, 2005), and is an indirect indication that the genes in such a family can act as functional substitutes. To increase the level of confidence that genes from one family can indeed substitute for each other, only families with up to ten family members are considered. The spatial correlation is increased by mapping the regions surrounding genes within the same family on top of each other, by aligning them with respect to a common reference (schematically depicted in Fig. 3). In this alignment the transcriptional direction of the genes is taken into account. The common reference, referred to as the pivot, is chosen to be the 5' end of the genes. A major advantage is that ICs that were previously separated now may be close enough to reach the significance threshold. Before the mapping is performed, a few conditions need to be satisfied: (1) ICs from the same tumor are not mapped, since common cooperations can only be called significant when encountered in more than one tumor. (2) Genes within one family that are close together are excluded, since the ICs in their neighborhood will already be spatially correlated. (3) ICs with a distance to the pivot exceeding five times the scale parameter are not mapped. These ICs will not contribute to the peak height, but may introduce false positives.

After the family mapping is performed, the 2DGKC method is applied to the ICs in the family mapped space. A Family Mapped CCI

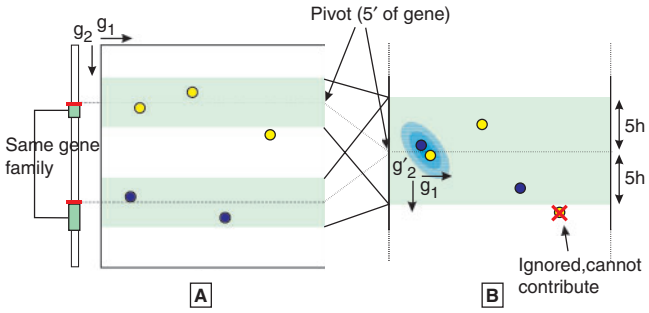

Fig. 3. Schematic depiction of the mapping of the ICs to the families. (**A**) The IC space with five ICs. Two genes have been depicted (green bars) that are members of the same family. The red bars denote the 5' ends of the genes. (**B**) The region around the genes are mapped onto each other, taking into account the direction of transcription of the gene, and using the pivot (5' end of the gene) as common reference. Only a region of five times the scale parameter is considered, since only ICs within this range will have an additive effect on the smoothed estimate of the number of ICs belonging to the family under investigation. ICs outside the region are therefore ignored. From the schematic it can be seen that, before the mapping, ICs that did not result in a peak exceeding the significance threshold, after the mapping may become close enough to have an additive effect on the smoothed estimate of the number of ICs, resulting in the discovery of Family Mapped CCI (indicated by the blue ellipse). Note that mapping changes only the g_2 dimension (denoted by g_2'), the g_1 dimension remains the same.

(FM-CCI) is defined as a peak that exceeds the significance threshold. The FM-CCIs indicate the cooperation of a region in the g_1 direction with one or more members of a certain gene family in the g_2 direction. Note that the mapping and 2DGKC is applied per family.

By mapping the regions around the genes from a family onto each other, the peak height that is expected by chance will increase. As a consequence, the null-distribution, against which the resulting peaks are compared, should incorporate this effect. This is achieved by including the family mapping before the permutation procedure depicted in Figure 2. The number of regions that are mapped onto each other changes as a function of the family size, and therefore a null-distribution is computed per family size. The multiple testing correction factor is equal to the total number of peaks evaluated in the family mapped space, which is approximately equal to the one used in the detection of CCIs.

2.7 Validation from literature

In order to validate the most prominent csCCIs that resulted from our analysis, we evaluated how often the two genes, close to a csCCI, co-occurred in the same MEDLINE abstract according to the online database PubGene (http://www.pubgene.org) (Jenssen *et al.*, 2001). This required a non-trivial mapping of the csCCI to their target genes. Although it has been shown that viral insertions most frequently target their closest neighboring gene (Erkeland *et al.*, 2006), it is likely that this simple heuristic will introduce some false negatives, thereby diluting the number of discovered co-occurring gene pairs in the PubGene database. To overcome this problem we evaluate all nine combinations of the three nearest genes surrounding the region marked by a csCCI in the g_1 direction against their three counterparts in the g_2 direction, and use only the combination that resulted in the maximum number hits in PubGene. We compare the results obtained by this procedure against the result obtained by repeating the same procedure with 2500 random combinations with the genes in our list.

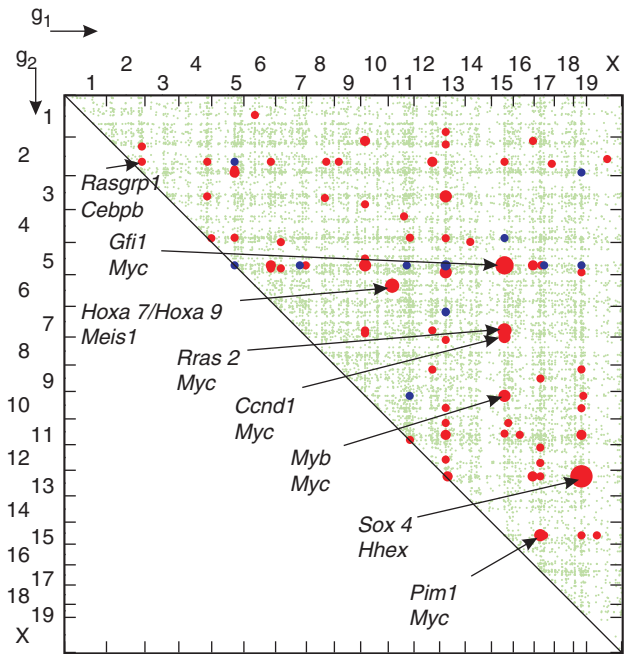

Fig. 4. (**A**) Co-occurrence plot for all the data in the RTCGD, where the axis markings denote the chromosomes and the green dots indicate the ICs. The red and blue dots mark the locations of the csCCIs, where the blue ones indicates the csCCIs for which non of the scales reached the additional 5% threshold according to the χ^2-test, described in the Methods section. The radius of the csCCI marker is proportional to the score obtained by normalizing the peak heights of the csCCIs per scale, and averaging this normalized peak height across the scales at which the csCCI was found to be significant. The arrows indicate the gene pairs discussed in the Results section.

3 RESULTS

3.1 Common co-occurrence of insertions

We have applied the proposed 2DGKC method to the combined data from the screens in the RTCGD. We evaluated the data at the following eight log-uniformly spaced scales: [10000, 17487, 30579, 53472, 93506, 163512, 285930, 500000] at a significance level of $\alpha = 0.05$. This resulted in the discovery of 86 csCCIs, that is, we find 86 pairs of loci that cooperate with each other in the development of the tumor. An overview of the results are given in Figure 4 and the top ten csCCIs are listed in Table 1 (a complete list is available online).

A number of interactions identified in retroviral mutagenesis screens have previously been characterized. *Myc* collaborates with *Pim1* (Verbeek *et al.*, 1991), *Myb* (Davies *et al.*, 1999), *Gfi1* (Schmidt *et al.*, 1998), and *Cyclin D1* (Lovec *et al.*, 1994) and *Hoxa9*/*Hoxa7* collaborate with *Meis1* (Kroon *et al.*, 1998). The majority of co-occurences however, have not been studied in mouse models of lymphoma, but in some cases the literature provides supporting evidence for their cooperation. For instance, the csCCI near *Rasgrp1*/*Cebpb* ranked 43rd in the list. *Rasgrp1* is a guanine nucleotide exchange factor that activates *Ras* signalling. *Cebpb* (CCAAT/enhancer-binding protein beta) is a transcription factor that mediates interleukin-6 (*IL-6*) signalling. *Cebpb* is also an important mediator of *Ras* induced oncogenesis (Zhu *et al.*, 2002).

Interestingly, when ranking the csCCIs according to the χ^2-test, a rather different top 10 is found (Table 2). These interactions are of special interest, since the individual loci are inserted in relatively few tumors, which makes it more likely that the combination of the two mutations is causal for development of the tumor. Figure 2 shows the result after applying an additional 0.05 threshold to the *P*-value resulting from the χ^2-test. Indeed, it can be seen that 12 csCCIs (colored blue in Fig. 4) do not reach this additional threshold, and may therefore be of less interest. Notably, they mainly represent interactions with either *Sox4* or *Gfi1*, which, by themselves, are both frequently targeted in insertional mutagenesis screens.

3.2 Validation from literature

Table 1 lists the candidate target gene pairs, as indicated by the top ten of the 86 csCCIs. By searching the PubGene database we found six of these ten gene pairs to co-occur in the literature abstracts. This is statistically significant ($P < 6.3 \times 10^{-4}$), when compared to the 322 hits that resulted from querying 2500 random, and therefore mostly unrelated, combinations in our set. Also when considering the complete list of 86 gene pairs indicated by the csCCIs, we find a statistically significant

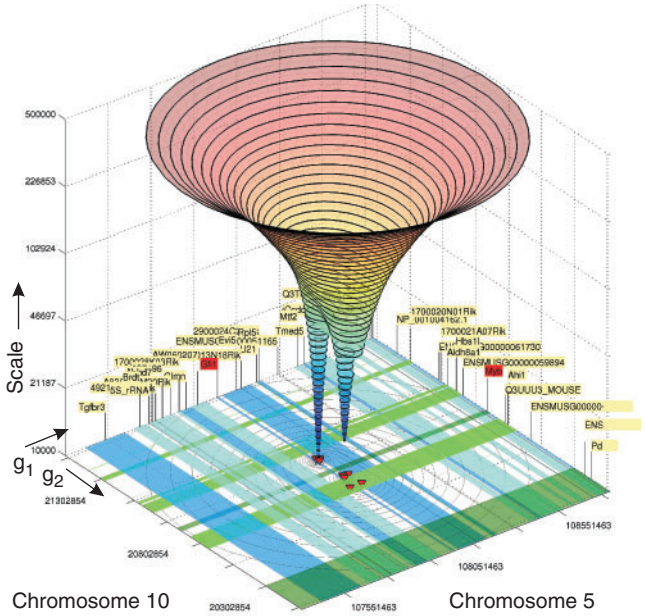

Fig. 5. Scale space diagram of a csCCI located on the Chromosome 10/Chromosome 5 intersection, near the *Gfi1* and *Myb* genes. The dark blue and light blue areas indicate the genes on the top and bottom strand in the g_1 direction, respectively. The dark green and light green areas indicate the genes on the top and bottom strand in the g_2 direction, respectively. The red triangles mark the location of the ICs. From the scale space diagram it becomes clear that there are in fact two distinct loci of integration on either side of the *Myb* gene.

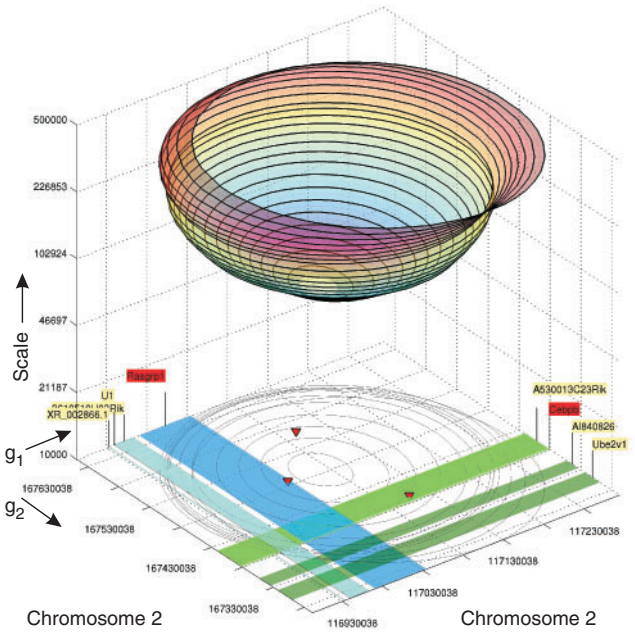

Fig. 6. Scale space diagram of a csCCI located on the Chromosome 2/Chromosome 2 intersection, near *Rasgrp1* and *Cebpb*. Nomenclature is equivalent to Figure 5. Note that this csCCI only is significant at higher scales, and can therefore be missed if the wrong (subset of) scale(s) is evaluated.

Table 1. Top ten of the csCCIs, ranked according to their average peak height across the scales, and their candidate targets and hits in PubGene. The candidate targets are defined as the gene pairs with most hits. When no PubGene hits were scored, the RTCGD consensus genes are listed

csCCI rank	Gene(s) 1	Gene(s) 2	Number of PubGene hits
1	*Sox4*	*Hhex*	0
2	*Gfi1*	*Myc*	17
3	*Rras2*	*Myc*	0
4	*Hoxa9/Hoxa7*	*Meis1*	25
5	*Myb*	*Myc*	151
6	*Gfi1*	*Myb*	1
7	*Myc*	*Pim1*	56
8	*AA881470/Iqce*	*Sox4*	0
9	*Notch2*	*Sox4*	0
10	*Ccnd1*	*Myc*	141

overrepresentation in the literature abstracts ($P < 5.6 \times 10^{-4}$), since 23 of these co-occurred in the PubGene database. For the ten gene pairs listed in Table 2, no significant overrepresentation in literature abstracts was established. This is not surprising, since these genes are hit relatively infrequently, and are therefore less likely to be well-characterized in literature.

3.3 Scale space diagrams

The list in Table 2 contains some interesting putative cooperations between genes, but by plotting the csCCIs in the scale space, valuable extra information about the cooperation can be gained. From Figure 5 it is clear that, at the largest scales, insertions near *Myb* clearly co-occur with *Gfi1* insertions. *Gfi1* and *Myb* are transcription factors with roles in hematopoiesis (Mucenski *et al.*, 1991: Zeng *et al.*, 2004). At the smaller scales however, inserts surrounding *Myb* can be divided into two separate clusters, and independently associate with the *Gfi1* locus. This suggests that inserts from both clusters are functionally equivalent, thereby strengthening the case for grouping them into a single CCI at larger scales, but possibly also indicates a different mechanism by which they disrupt functioning of *Myb*. This diagram can thus give valuable insight in the mechanisms that disrupts gene functioning. Other examples exist where csCCIs are only significant at a certain range of scales, for instance the previously mentioned csCCI near *Rasgrp1* and *Cebpb* (Fig. 6). Clearly, when evaluating this csCCIs at a single scale or subset of scales, one runs the risk to miss this significant cooperation if the scale at which it is evaluated does not match the scale of the CCI.

3.4 Family mapping

Figure 7A shows the previously mentioned example of the possible substitution of insertions near *Pim2* for *Pim1* mutations. The figure exemplifies that, by performing the family mapping, indeed meaningful extra interactions are found.

The IC near *Pim2* and *Myc* would have gone undetected in the normal co-occurrence analysis, the family mapping proves capable of exploiting the additional information contained in this IC.

Similarly interesting is the discovered FM-CCI indicating cooperation between *Sox4* and the *Cyclin dependent kinases* family. Seven from the nine genes in this family are hit in eight independent tumors. Figure 7B shows the scale space diagram for this interaction. Apparently, *Sox4* insertions cooperate interchangeably with one of the members of the *Cyclin dependent kinases* family. Figure 8 shows how the ICs targeting the *Sox4*/*Cyclin dependent kinases* family are distributed over the tumors. Notably, none of the genes in the *Cyclin dependent kinases* family is hit frequently enough to reach significance on its own account (the two ICs near *Sox4*/*Cdk6* are too far from each other to reach significance). It is only by applying the family mapping that cooperation between *Sox4* and the *Cyclin dependent kinases* family can be discovered.

Table 2. Top 10 of the ranked csCCIs, according to the χ^2-ranking procedure. RTCGD consensus genes are listed

csCCI rank	Gene(s) 1	Gene(s) 2
1	*Hoxa9*/*Hoxa7*	*Meis1*
2	*Meis1*	*Dnalc4*
3	*Lmo2*	*Il2rg*
4	*Ramp1*	*Hoxa9*/*Hoxa7*
5	*Gabpb1*	*Eml4*
6	*Ccr7*	*Hexim1*
7	*Pptc7*	*Pou2f2*
8	*Sox4*	*Hhex*
9	*Zdhhc18*/*Arid1a*	*Map3k14*/*Fmnl1*
10	*Rap1a*/*6530418L21Rik*	*Nfix*/*Lyl1*

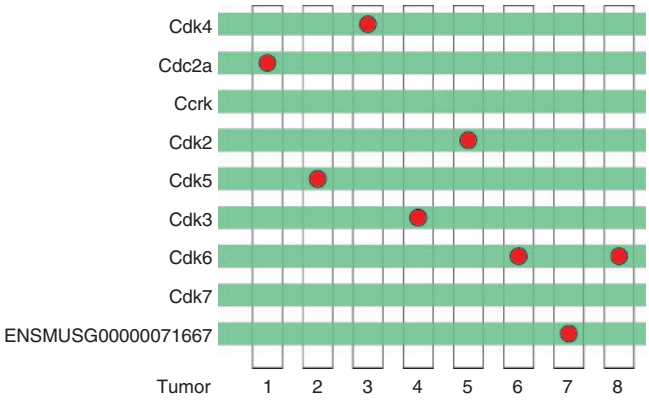

Fig. 8. Schematic depiction of the distribution of ICs that were encountered near *Sox4* (within a 1 Mbp square window), over the nine members from the *Cyclin dependent kinases* family. Only *Cdk6* is hit twice, but the ICs were too far from each other to reach significance by themselves. The figure shows that this interaction, among others, can only be found by applying a family mapping.

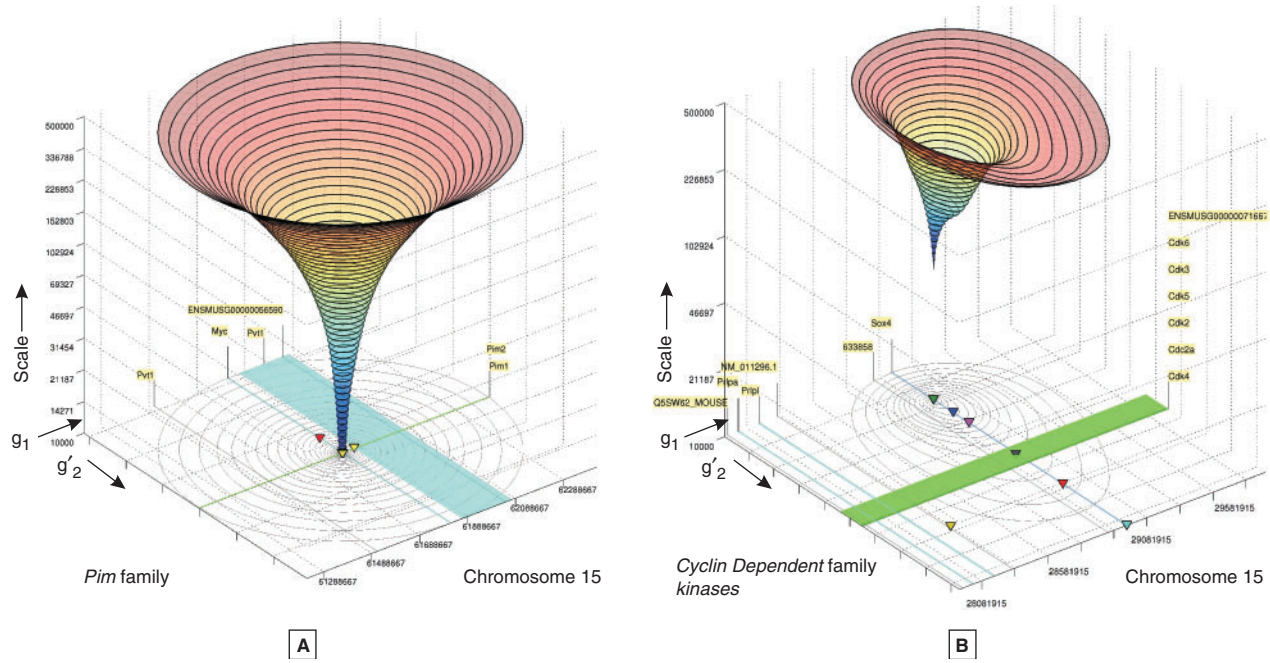

Fig. 7. Scale space diagrams of FM-CCIs. Nomenclature is equivalent to Figure 5, with the exception of the green area, which indicate the genes of the gene family under investigation, in the g'_2 direction. (**A**) The interaction between *Myc* and the *Pim* family (ENSF00000001108: SERINE/THREONINE KINASE PIM) in the scale space. The red triangles mark ICs near *Pim2*, and yellow triangles mark ICs near *Pim1*. (**B**) The interaction between *Sox4* and the *Cyclin dependent kinases* Family (ENSF00000000186: CELL DIVISION). The coloring of the ICs indicate near which seperate family member it occurred. Notably, seven of the nine genes in this family are hit.

4 CONCLUSIONS AND DISCUSSION

Until now, the main focus of analysis on insertional mutagenesis data has been one-dimensional, that is, discovering regions in the genome that are causal for tumor development, the CISs. In this article we analyzed the data from publicly available retroviral insertional mutagenesis screens in the 2D co-occurrence space. By evaluating the significance of co-occurring insertions we found 86 statistically significant csCCIs, that indicate cooperation between insertionally targeted genes. By analyzing the data in a scale space we are able to detect csCCIs that are only significant at a limited subset of the scales, for instance the putative cooperation between *Rasgrp1* and *Cebpb*. In addition, the scale space provides essential information about mechanisms that underlie the viral disruption of gene functioning. This was exemplified by the putative cooperation between *Myb* and *Gfi1*, where the scale space showed two sub-CCIs at low scales, indicating two confined regions of integration.

To assess whether also known cooperation between genes are found, we showed that the set of candidate gene pairs, resulting from our study, is significantly overrepresented in the PubGene database, a literature network containing gene-to-gene co-citations. In addition to known cooperations, our study also revealed previously unknown putative cooperations, that are interesting targets for possible follow-up studies. We have presented two rankings of the resulting csCCIs, one based on average peak height and one based on the average *P*-value resulting from a χ^2-test. The latter ranking takes into account the possibility that a csCCI is caused by frequent insertion of one or both of the individual loci. We can conclude that, by analyzing the data in the co-occurrence space, and at multiple scales, we can find new statistically significant regions in the genome that play a role in tumor development.

To deal with the possibility that cells choose alternative pathways to become malignant, we have incorporated information about gene families in the analysis. By remapping the data according to putative substitutions derived from gene family membership, we were able to discover significant cooperations between genes and genes from a gene family. Examples of the known substitution of *Pim2* insertions for insertions near *Pim1* in tumors with virally activated *Myc*, as well as the putative cooperation between *Sox4* and the *Cyclin dependent kinases* family were given. These examples show that much is to be gained by integrating insertional mutagenesis data with other data sources, such as gene family information, since the insertion data in itself is far from saturated.

The methods presented are especially beneficial for data from high throughput screens with many insertional mutations per tumor. Therefore, the methods may be applied to other types of genome wide mutagenesis data as well, for example data from transposon screens (Collier and Largaespada, 2005). As the amount of data increases, extensions to a multi-occurrence analysis become interesting. For the proposed 2DGKC method, these extensions are fairly straightforward.

ACKNOWLEDGEMENTS

This work was part of the BioRange programme of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI).

Conflict of interest: none declared.

REFERENCES

Akagi,K. *et al.* (2004) RTCGD: retroviral tagged cancer gene database. *Nucleic Acids Res.*, **32**(Database issue), D523–D527.

Collier,L.S. and Largaespada, D.A. (2005) Hopping around the tumor genome: transposons for cancer gene discovery. *Cancer Res.*, **65**, 9607–9610.

Davies,J. *et al.* (1999) Cooperation of myb and myc proteins in t cell lymphomagenesis. *Oncogene*, **18**, 3643–3647.

de Ridder,J. *et al.* (2006) Detecting statistically significant common insertion sites in retroviral insertional mutagenesis screens. *PLoS Comput. Biol.*, **2**, e166.

Erkeland,S.J., *et al.* (2006) Significance of murine retroviral mutagenesis for identification of disease genes in human acute myeloid leukemia. *Cancer Res.*, **66**, 622–626.

Hanahan,D. and Weinberg, R.A. (2000) The hallmarks of cancer. *Cell*, **100**, 57–70.

Hansen,G.M. *et al.* (2000) Genetic profile of insertion mutations in mouse leukemias and lymphomas. *Genome Res.*, **10**, 237–243.

Hubbard,T. *et al.* 2005. *Nucleic Acids Res*, **33**, D447–D453.

Hwang,H.C. *et al.* (2002) Identification of oncogenes collaborating with p27Kip1 loss by insertional mutagenesis and high-throughput insertion site analysis. *Proc. Natl Acad. Sci. USA*, **99**, 11293–11298.

Jenssen,T.K. *et al.* (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.*, **28**, 21–28.

Johansson,F.K. *et al.* (2004) Identification of candidate cancer-causing genes in mouse brain tumors by retroviral tagging. *Proc. Natl Acad. Sci. USA*, **101**, 11334–11337.

Joosten,M. *et al.* (2002) Large-scale identification of novel potential disease loci in mouse leukemia applying an improved strategy for cloning common virus integration sites. *Oncogene*, **21**, 7247–7255.

Kim,R. *et al.* (2003) Genome-based identification of cancer genes by proviral tagging in mouse retrovirus-induced T-cell lymphomas. *J Virol*, **77**, 2056–2062.

Kroon,E. *et al.* (1998) Hoxa9 transforms primary bone marrow cells through specific collaboration with meis1a but not pbx1b. *EMBO J.*, **17**, 3714–3725.

Li,J. *et al.* (1999) Leukaemia disease genes: large-scale cloning and pathway predictions. *Nat. Genet.*, **23**, 348–353.

Liu,B. *et al.* (2001) Identifying non-actionable association rules. In *KDD '01: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM Press: New York, NY, USA, pp 329–334..

Lovec,H. *et al.* (1994) Cyclin d1/bcl-1 cooperates with myc genes in the generation of b-cell lymphoma in transgenic mice. *EMBO J.*, **13**, 3487–3495.

Lund,A.H. *et al.* (2002) Genome-wide retroviral insertional tagging of genes involved in cancer in Cdkn2a-deficient mice. *Nat. Genet.*, **32**, 160–165.

Mikkers,H. and Berns, A. (2003) Retroviral insertional mutagenesis: tagging cancer pathways. *Adv. Cancer Res.*, **88**, 53–99.

Mikkers,H. *et al.* (2002) High-throughput retroviral tagging to identify components of specific signaling pathways in cancer. *Nat. Genet.*, **32**, 153–159.

Mucenski,M.L. *et al.* (1991) A functional c-myb gene is required for normal murine fetal hepatic hematopoiesis. *Cell*, **65**, 677–689.

Nakamura,T. *et al.* (1996) Cooperative activation of Hoxa and Pbx1-related genes in murine myeloid leukaemias. *Nat. Genet.*, **12**, 149–153.

Parzen,E. (1962) On estimation of a probability density function and mode. *The Ann. Math. Stat.*, **33**, 1065–1076.

Schmidt,T. *et al.* (1998) Zinc finger protein gfi-1 has low oncogenic potential but cooperates strongly with pim and myc genes in t-cell lymphomagenesis. *Oncogene*, **17**, 2661–2667.

Suzuki,T. *et al.* (2002) New genes involved in cancer identified by retroviral tagging. *Nat Genet*, **32**, 166–174.

Uren,A.G. *et al.* (2005) Retroviral insertional mutagenesis: past, present and future. *Oncogene*, **24**, 7656–7672.

vander Lugt,N.M. *et al.* (1995) Proviral tagging in e mu-myc transgenic mice lacking the pim-1 proto-oncogene leads to compensatory activation of pim-2. *EMBO J.*, **14**, 2536–2544.

van Lohuizen,M. *et al*. (1991) Identification of cooperating oncogenes in E mu-myc transgenic mice by provirus tagging. *Cell*, **65**, 737–752.

Verbeek,S. *et al*. (1991) Mice bearing the e mu-myc and e mu-pim-1 transgenes develop pre-b-cell leukemia prenatally. *Mol. Cell. Biol.*, **11**, 1176–1179.

Zeng,H. *et al*. (2004) Transcription factor gfi1 regulates self-renewal and engraftment of hematopoietic stem cells. *EMBO J.*, **23**, 4116–4125.

Zhu,S. *et al*. (2002) Ccaat/enhancer binding protein-beta is a mediator of keratinocyte survival and skin tumorigenesis involving oncogenic ras signaling. *Proc. Natl Acad. Sci. USA*, **99**, 207–212.